Joint Searching and Grounding: Multi-Granularity Video Content Retrieval

Zhiguo Chen*
Center for Future Media & School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu, China

Xun Jiang*
Center for Future Media & School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu, China

Xing Xu†
Center for Future Media & School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu, China

Zuo Cao
Meituan
Shanghai, China

Yijun Mo
Hubei Specialized Institute of Intelligent Edge Computing
Huazhong University of Science and Technology
Wuhan, China

Heng Tao Shen
Center for Future Media & School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu, China
Peng Cheng Laboratory
Shenzhen, China

ABSTRACT

Text-based video retrieval is a well-studied task aimed at retrieving relevant videos from a large collection in response to a given text query. Most existing TVR works assume that videos are already trimmed and fully relevant to the query thus ignoring that most videos in real-world scenarios are untrimmed and contain massive irrelevant video content. Moreover, as users' queries are only relevant to video events rather than complete videos, it is also more practical to provide specific video events rather than an untrimmed video list. In this paper, we introduce a challenging but more realistic task called *Multi-Granularity Video Content Retrieval (MGVCR)*, which involves retrieving both video files and specific video content with their temporal locations. This task presents significant challenges since it requires identifying and ranking the partial relevance between long videos and text queries under the lack of temporal alignment supervision between the query and relevant moments. To this end, we propose a novel unified framework, termed, *Joint Searching and Grounding (JSG)*. It consists of two branches: (1) a glance branch that coarsely aligns the query and moment proposals using inter-video contrastive learning, and (2) a gaze branch that finely aligns two modalities using both inter- and intra-video contrastive learning. Based on the glance-to-gaze design, our JSG method learns two separate joint embedding spaces for moments and text queries using a hybrid synergistic contrastive learning strategy. Extensive experiments on three public benchmarks, *i.e.*, Charades-STA, DiDeMo, and ActivityNet-Captions demonstrate the superior performance of our JSG method on both video-level retrieval and event-level retrieval subtasks. Our open-source implementation code is available at https://github.com/CFM-MSG/Code_JSG.

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**; • **Information systems** → *Multimedia information systems.*

KEYWORDS

Multi-Granularity Video Content Retrieval; Contrastive Learning; Multimodal Learning; Video Understanding; Multimedia Applications;

ACM Reference Format:
Zhiguo Chen, Xun Jiang, Xing Xu, Zuo Cao, Yijun Mo, and Heng Tao Shen. 2023. Joint Searching and Grounding: Multi-Granularity Video Content Retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3612349

1 INTRODUCTION

Text-based Video Retrieval (TVR) [4, 7, 22, 24] aims at retrieving videos from a large video collection with natural language queries. In the last decades, the TVR task has been fully explored and achieved impressive progress. However, as illustrated in Fig. 1(a), most existing TVR methods assume that the target videos are already trimmed and fully relevant to the query, thus limiting their performance on the untrimmed long video collections. Recently, Dong *et al.* [5] extended the conventional TVR task to a Partially Relevant Video Retrieval task, which aims to retrieve untrimmed videos with partially relevant text queries. However, they still focus

*Equal Contributions.
†Corresponding Author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612349

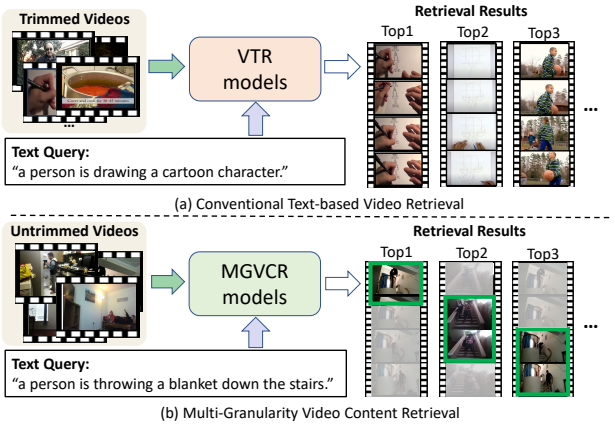

Figure 1: An illustrative example of conventional Text-based Video Retrieval and our proposed Multi-Granularity Video Content Retrieval: (a) The TVR aims at retrieving video files from a trimmed video collection. (b) The MGVCR not only retrieves video files but also specific video content.

on retrieving video files only and ignore the specific video content that is highly related to users' queries, thus can not be applied to tackle the event-specific retrieval requirements. Due to these two defects, these existing methods show more limitations in most real-world scenarios.

To this end, we are dedicated to studying a novel and more practical text-based video retrieval paradigm, which could retrieve video files and specific relevant video events synchronously and has not been fully explored yet. As is depicted in Fig. 1(b), given a short text query, it requires models to retrieve the most relevant video files, as well as the specific video events with their starting and ending timestamps. Such a paradigm is more practical as it could tackle the coarse-grained and fine-grained retrieval requirements in an end-to-end manner. Nevertheless, it is also more challenging compared with the conventional TVR task: (1) As the queries are only partially relevant to target videos, it is difficult to learn a joint embedding space for metric learning with previous methods. (2) As a cross-modal retrieval task, the temporal annotations of video events are hard to get since labeling boundary-wise annotations for a large untrimmed video collection is often impractical or expensive to obtain in real-world scenarios. Due to these challenges, the early attempts [9, 22, 30, 36] exist various limitations attribute to annotation cost, performance, or efficiency.

In this paper, we formulate an advanced text-based video content retrieval paradigm, termed Multi-Granularity Video Content Retrieval (MGVCR), which involves both video-level and event-level retrieval in a unified framework trained with the video-text matching pairs. To overcome the mentioned challenges above, we designed a novel video retrieval method, namely, Joint Searching and Grounding (JSG). Specifically, as illustrated in Fig. 2, our JSG method learns an effective joint embedding space for moments and text queries using a hybrid synergistic contrastive learning strategy. It consists of two key branches: (1) a Glance Branch that coarsely aligns the query and event proposals using inter-video contrastive learning, and (2) a Gaze Branch that finely aligns two modalities using both inter- and intra-video contrastive learning. Moreover, to overcome the challenge of redundant irrelevant video content, we propose two video content proposal generation mechanisms to capture potential video events. We conduct extensive experiments on three public benchmarks, *i.e.*, Charades-STA, DiDeMo, and ActivityNet-Captions, and demonstrate the superior performance of our proposed method on both video-level retrieval and event-level retrieval subtasks.

To sum up, the main contributions of this paper are three-fold: (1) We introduce a novel and practical text-based video retrieval task called Multi-Granularity Video Content Retrieval (MGVCR), which involves retrieving both video files and specific video content with their temporal locations. We also design an effective framework, *i.e.*, Joint Searching and Grounding (JSG), for this challenging task. (2) We design a novel unified framework that learns an effective joint embedding space with a glance-to-gaze learning procedure. It successfully overcomes the difficulty brought by the redundancy of untrimmed videos and the absence of temporal annotations. (3) We propose a hybrid synergistic contrastive learning strategy that collaborates both coarse-grained and fine-grained video content learning. It effectively aligns the local video content and text queries in two separate joint embedding spaces.

2 RELATED WORKS

Temporal Sentence Grounding. This task aims at retrieving specific video content from an untrimmed video with a given text query. The deep learning solution for the VMR task is firstly proposed by [1, 11]. As a fundamental problem in multi-modal video understanding, it has been studied for several years and most existing methods can be categorized into two types: fully-supervised methods [1, 11, 12, 17, 27, 33, 37] and weakly-supervised methods [13, 19, 39, 40]. The former requires fully-supervised training annotations, where each event in videos are labeled with their start and end time in a video. Contrastively, the latter, weakly-supervised methods [13, 39, 40], focuses on reducing the heavy annotation cost in this task by eliminating fine-grained temporal supervision signals. Although a TSG method can be applied to untrimmed videos to align the text to partial video content, it is limited to a single video rather than large-scale video datasets.

Text-based Video Retrieval. Retrieving relevant video content from a large video collection with given text queries is an essential task for a multimedia system. Following the priori assumption that text queries are highly aligned with complete videos, current methods [4, 6, 7, 24] widely explored video-level retrieval with a common embedding space for video and text representations. However, the works above focused on retrieving trimmed videos containing little redundant content and aligning with text queries well, which limits their generalization to the real world. Due to this, several recent works [5, 9, 22, 30, 36] focused on retrieving video content from an untrimmed video collection with partially relevant text queries. The early methods [22, 36] of text-based untrimmed videos tackled this problem by introducing extra supervision signals by manually labeling the corresponding video events. With the extra temporal annotations, these methods achieved promising performance on video-level retrieval and also pioneeringly combined text-based video-level and event-level retrieval. However, as annotating video events in most cross-modal video retrieval is extremely extensive and subjective, it is actually impractical for a multimedia retrieval

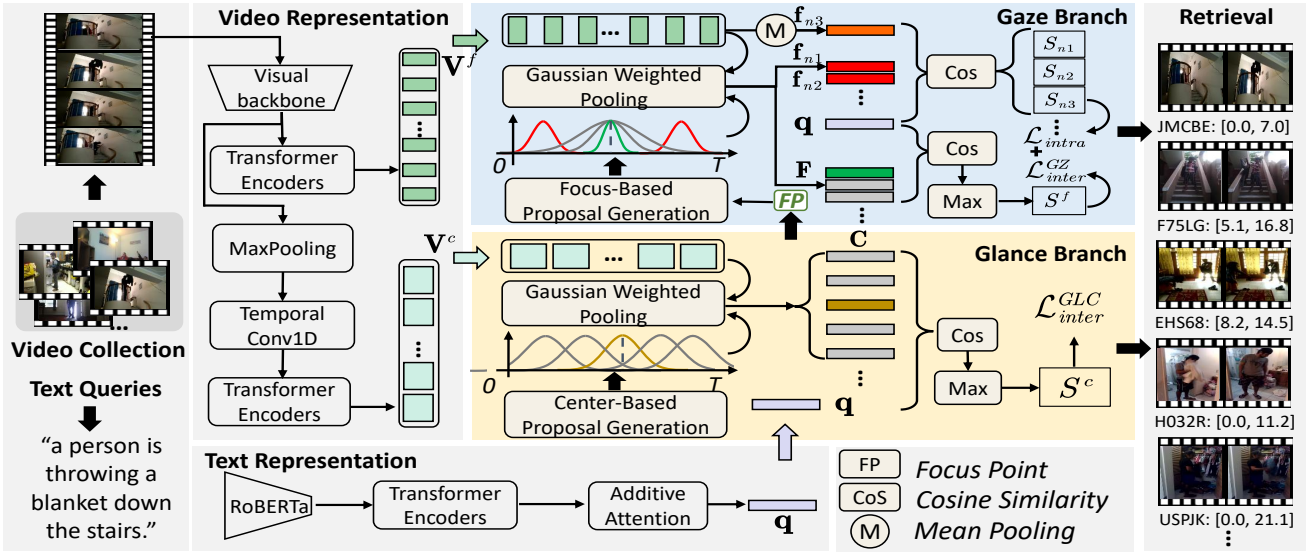

Figure 2: Illustration of our proposed JSG framework for Multi-Granularity Video Content Retrieval task. The key components of our framework are the Glance branch and the Gaze branch. We display the retrieval results on the right side of the figure, where the top 5 results are given with the video IDs and the start and end timestamps.

system. To this end, we are dedicated developing a unified framework for video-level and event-level retrieval, which is only based on conventional video-text matching pairs.

Contrastive Learning. Contrastive learning aims at learning the common features of similar examples and distinguishing the differences between non-similar examples. In recent years, it has attracted great attention in many research fields, such as multimodal representation [14, 28, 29, 35] or multimedia recommendation [34, 41]. Typically, He *et al.* [14] proposed momentum contrastive for unsupervised visual representation learning and achieved superior performance on a group of downstream tasks. Moreover, contrastive learning has also made great success in multimodal video understanding [5, 18, 38, 40]. Taivanbat *et al.* [2] proposed a video highlight detection framework that distinguishes a video from other videos via a contrastive objective. Dong *et al.* [5] employed contrastive learning to discriminate the difference between matched pairs and irrelevant pairs, which inspired us to design a hybrid synergistic contrastive learning strategy in this work.

3 PROPOSAL METHOD

3.1 Problem Statement

Here we first give a specific definition for the Multi-Granularity Video Content Retrieval (MGVCR) task. Consider that we have a set of M untrimmed long videos $\mathcal{V} = \{V_k\}_{k=1}^{M}$, where $V_k = \{f_i\}_{i=1}^{T}$ denotes the k-th video with T frames. Given a text query $Q = \{w_i\}_{i=1}^{n_w}$, where w_i is the i-th word in the query, and n_w is the number of words in the query. The goal of our task is to retrieve: (1) a video V^* that contains the relevant event to the query and (2) the most relevant video moment M^* in the video V^* by providing the start and end time point τ^s and τ^e. Note that the temporal boundary τ^s and τ^e of the video moment M^* is not available in the training phase, and the video moment M^* can be a short clip, long clip, or even the whole video.

3.2 Feature representation

Text Representation. Following [22], we first represent the text queries as word-level features and then aggregate the word-level features into sentence-level features. Specifically, given a sentence query $Q = \{w_i\}_{i=1}^{n_w}$ with n_q words, we employ a pre-trained RoBERTa model to extract the word-level features. We furtherly use a fully connected layer to project the word-level features into a d-dimensional space. To model the contextual information in the sentences, we add the learned positional embedding to the projected word-level features, and encode the features with a Transformer encoder layer[32]. The encoded contextual word-level features can be represented as $\mathbf{Q} = \{\mathbf{q}_i\}_{i=1}^{n_q}$, where $\mathbf{q}_i \in \mathbb{R}^d$ is the i-th word's embedding. Finally, an additional attention mechanism is applied to aggregate the word-level features $\mathbf{Q}$ into a sentence-level feature $\mathbf{q} \in \mathbb{R}^d$:

$$\mathbf{q} = \sum_{i=1}^{n_q} \alpha_i \mathbf{q}_i, \alpha = Softmax(\mathbf{Q}\mathbf{W}^T), \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{1\times d}$ is a trainable vector, and α is the attention weights.

Video Representation. Given an untrimmed video, we employ the pre-trained CNN-based visual backbones to extract the video features, and evently sample a fixed number of vectors to represent the video, which can be denoted as $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^{n_v} \in \mathbb{R}^{nv\times d_v}$, where n_v is fixed for each video. For the fine-grained video representation, we first project the video features $\mathbf{V}$ into a d-dimensional space with an FC layer, and then encode the projected features with a Transformer encoder layer to get contextualized features $\mathbf{V}^f = \{\mathbf{v}_i^f\}_{i=1}^{n_v} \in \mathbb{R}^{n_v\times d}$. Inspired by [17], we employ a learnable downsampling strategy with temporal convolution to obtain the coarse-grained video representation $\mathbf{V}^d \in \mathbb{R}^{n_c\times d}$:

$$\mathbf{V}^d = \text{Conv1D}(\text{MaxPooling}(\mathbf{V}, k1, s), k2, 1), \tag{2}$$

where $k1$ and s are the kernel size and step size of the temporal max-pooling layer, $k2$ is the kernel size of the temporal convolution

layer and the step size is set to 1. n_c is the feature length, where $n_c < n_v$. In this way, each video feature in the coarse-grained video representation has a large temporal receptive field, and the number of feature vectors has reduced significantly, which saves computational costs in the glance branch. To obtain the contextualized coarse-grained video features, we also encode the downsampled features with an FC layer and Transformer encoder layer and denote the features as $\mathbf{V}^c = \{\mathbf{v}_i^c\}_{i=1}^{n_c} \in \mathbb{R}^{n_c \times d}$.

3.3 Glance Branch

Center-Based Proposal Generation (CBPG). The CBPG module is designed to generate local video representations of moment proposals around the whole video. To facilitate the description, we represent a proposal as a center point and a width that indicates the proposal's temporal extent. Specifically, we evently sample n_{pc} centers as the relative temporal position of the moment proposals' centers in the videos:

$$\mathbf{P^c} = \{0, ..., \frac{i}{n_{pc}-1}, ..., 1\}, 0 \leqq i \leqq (n_{pc}-1) \tag{3}$$

For each center, we evenly sample n_{pw} multi-scale moment proposals with the proposals' relative width varying from $w_{lb} \in (0, 1)$ to $w_{ub} \in (w_{lb}, 1]$:

$$\mathbf{P^w} = \{w_{lb}, ..., w_{lb} + \frac{i(w_{ub}-w_{lb})}{n_{pw}-1}, ..., w_{ub}\}, 0 \leqq i \leqq (n_{pw}-1) \tag{4}$$

In this way, the moment proposals can be obtained by combining the centers and widths: $\mathbf{P} = \{\mathbf{p}_i\}_{i=1}^{n_p}$, where $\mathbf{p}_i = \{p_i^c, p_i^w\}$, and $n_p = n_{pc} \times n_{pw}$ is the total number of proposals.

Inspired by the observation that some frames are more representative than others for a given video event, we introduce the Gaussian Weighted Pooling (GWP) mechanism to generate moment representations. Let the representation of the coarse-grained moment proposals be $\mathbf{C} = \{\mathbf{c}_j\}_{j=1}^{n_p}$, $\mathbf{c}_j \in \mathbb{R}^d$ is the j-th proposal's feature aggregated with the GWP mechanism:

$$\mathbf{c}_j = \sum_{i=1}^{n_c} \frac{\mathbf{v}_i^c}{\sqrt{2\pi}(p_j^w/\sigma)} exp\left(-\frac{(i/n_c - p_j^c)^2}{2(p_j^w/\sigma)^2}\right), \tag{5}$$

where σ is a scale factor, which is set to 9 in all experiments.

Coarse-grained Similarity Measure. Given a pair of query and video, we first calculate the cosine similarity between each of the coarse-grained moment representations and the sentence-level query representation, then select the optimal one as the similarity score of the query and video pair. Specifically, the similarity score between a query and a video can be obtained as:

$$S^c(q, v) = max\{S(\mathbf{q}^T, \mathbf{c}_1), S(\mathbf{q}^T, \mathbf{c}_2), ..., S(\mathbf{q}^T, \mathbf{c}_{n_p})\}, \tag{6}$$

where $S(\cdot)$ is the cosine similarity function. The optimal moment representation is denoted as $\mathbf{c}^* = \arg\max_i S(\mathbf{q}^T, \mathbf{c}_i)$, and the corresponding optimal moment proposal can be represented as $\mathbf{p}_* = \{p_*^c, p_*^w\}$. We take the center p_*^c of the optimal proposal p_* as the *focus point* which is a guidance for fine-grained moment proposal generation in our Gaze Branch 3.4.

Alignment with Inter-video Contrastive Learning. Given a text query, it is crucial to distinguish the partially related video from the unrelated ones. Besides, the optimal moment proposal from the partially related video is expected to be more similar to the query in semantics than the proposals from other videos. To align the representations of the moment proposals and the corresponding text queries in a coarse-grained manner, we adopt the inter-video contrastive learning strategy, which encourages the similarity between the positive pairs (*i.e.*, the query and the corresponding optimal proposal from the same video) to be higher than that between the negative pairs (*i.e.*, the query and the proposals from different videos). For the training objectives, we jointly employ the InfoNCE loss [15] and the triplet ranking loss [31] to coarsely align the cross-modal semantics. Given a positive query-video pair, the InfoNCE loss over a mini-batch $\mathcal{B}$ can be defined as:

$$\begin{aligned}\mathcal{L}_{nce}^c = -\frac{1}{n}\sum_{(q,v)\in\mathcal{B}}\Bigg[&\log\left(\frac{S^c(q,v)}{S^c(q,v)+\sum_{v^-\in\mathcal{N}_v}S^c(q,v^-)}\right)\\ &+\log\left(\frac{S^c(q,v)}{S^c(q,v)+\sum_{q^-\in\mathcal{N}_q}S^c(q^-,v)}\right)\Bigg],\end{aligned} \tag{7}$$

where n is the batch size, $\mathcal{N}_v$ and $\mathcal{N}_q$ are sets of all the negative videos and negative queries in the mini-batch that correspond to the given query q and video v, respectively. The triplet ranking loss is defined as:

$$\begin{aligned}\mathcal{L}_{trip}^c = \frac{1}{n}\sum_{(q,v)\in\mathcal{B}}\big[&\max\left(0, \Delta_1 + S^c(q,v^-) - S^c(q,v)\right)\\ &+\max\left(0, \Delta_1 + S^c(q^-,v) - S^c(q,v)\right)\big],\end{aligned} \tag{8}$$

where Δ_1 is the margin, and v^- and q^- are a negative video and a negative query sampled from the mini-batch that corresponds to the given query q and video v, respectively. Moreover, we also follow [5] to sample the negative pairs with an easy-to-hard strategy. The overall inter-video contrastive loss of glance branch is obtained as $\mathcal{L}_{inter}^{GLC} = \mathcal{L}_{trip}^c + \beta_1\mathcal{L}_{nce}^c$, where β_1 is a balance factor.

3.4 Gaze Branch

Focus-Guided Proposal Generation (FGPG). Given the focus $p_* = \{p_*^c, p_*^w\}$ from the Glance Branch 3.3, the FGPG module aims to generate more precise moment proposals. Specifically, for each video-query pair (v,q), we first obtain the optimal proposal $p_* = \{p_*^c, p_*^w\}$ from the glance branch, which has the highest similarity with the query among all the coarse-grained proposals. Then we generate the fine-grained proposals by sampling n_p^f widths around the reference point p_*^c, which can be denoted as:

$$\mathbf{P^{fw}} = \{w_{lb}, ..., w_{lb} + \frac{i(w_{ub}-w_{lb})}{n_p^f - 1}, ..., w_{ub}\}, 0 \leqq i \leqq (n_p^f - 1) \tag{9}$$

where w_{lb} and w_{ub} are hyperparameters of the lower and upper bounds of the width, respectively. In this way, the fine-grained moment proposals can be denoted as $\mathbf{P^f} = \{p_i^f\}_{i=1}^{n_p^f}$, where $p_i^f = \{p_*^c, p_i^{fw}\}$. To aggregate the fine-grained features, we adopt the GWP mechanism similar to the process in the CBPG module. Concretely, let the representation of the fine-grained moment proposals

be $\mathbf{F} = \{\mathbf{f}_j\}_{j=1}^{n_p^f}$, the aggregated feature $\mathbf{f}_j$ is obtained as:

$$\mathbf{f}_j = \sum_{i=1}^{n_v} \frac{\mathbf{v}_i^f}{\sqrt{2\pi}(p_i^{fw}/\sigma)} exp\left(-\frac{(i/n_v - p_*^c)^2}{2(p_i^{fw}/\sigma)^2}\right), \tag{10}$$

where $\mathbf{v}_i^f$ is the i-th fine-grained video feature, n_v is the number of features in V^f. Note that the FGPG module is only used in the training stage. In the inference stage, we directly use the CBPG module to generate the fine-grained moment proposals for the gaze branch, which does not rely on the optimal proposal p_* from the glance branch.

Intra-video Negative Proposal Mining. To further improve the alignment between the fine-grained moment proposals and the query, we adopt an intra-video negative proposal mining strategy to mine the negative proposals from the same video as the positive proposal. Inspired by [40], we generate two kinds of negative proposals: (1) the moments located to the left and right side of the positive proposal, and (2) the whole video that contains the positive proposal, which we called the reference proposal. Formally, given the positive proposal $p_p = \{p_p^c, p_p^w\}$, the first kind of negative proposals can be obtained as:

$$\begin{aligned} p_{n1} &= \{max(p_p^c - p_p^w/2, 0)/2, max(p_p^c - p_p^w/2, 0)\}, \\ p_{n2} &= \{1 - max(1 - p_p^c - p_p^w/2, 0)/2, max(1 - p_p^c - p_p^w/2, 0)\}, \end{aligned} \tag{11}$$

where p_{n1} and p_{n2} are the negative proposals located to the left and right side of the positive proposal, respectively. The feature of p_{n1} and p_{n2} can be obtained with the GWP mechanism similar to the fine-grained proposal representations f_i, and we denote the feature of p_{n1} and p_{n2} as $\mathbf{f}_{n1} \in \mathbb{R}^d$ and $\mathbf{f}_{n2} \in \mathbb{R}^d$, respectively. The second kind of negative proposals is obtained by temporally mean-pooling the features of the fine-grained video representation V^f, and we denote it as $\mathbf{f}_{n3} \in \mathbb{R}^d$.

Fine-grained Similarity Measure. Given a video-query pair, similar to the coarse-grained similarity measure in the glance branch, we calculate the similarity between each of the fine-grained proposals and the queries by using the cosine similarity on their representations, then the similarity between the given query and video is obtained by selecting the maximum similarity among all the fine-grained proposal and query pairs, which can be denoted as:

$$S^f(q, v) = max\{S(\mathbf{q}^T, \mathbf{f}_1), S(\mathbf{q}^T, \mathbf{f}_2), ..., S(\mathbf{q}^T, \mathbf{f}_{n_p})\}, \tag{12}$$

where $S(\cdot)$ is the cosine similarity function. We denote the optimal fine-grained moment representation as $\mathbf{f}^* = \arg\max_i S(\mathbf{q}^T, \mathbf{f}_i)$, which is taken as the positive sample in contrastive learning. For the negative proposals of each video-query pair, we also compute the cosine similarity and denote them as $S_{n1}(q, v) = S(\mathbf{q}^T, \mathbf{f}_{n1})$, $S_{n2}(q, v) = S(\mathbf{q}^T, \mathbf{f}_{n2})$ and $S_{n3}(q, v) = S(\mathbf{q}^T, \mathbf{f}_{n3})$, respectively.

Alignment with Hybrid Contrastive Learning. Given a text query, the MGVCR task requires not only distinguishing the positive video from the negative one but also distinguishing the positive moment proposal from the negative proposals in the positive video. In this case, we propose a hybrid contrastive learning strategy that aligns the representations of the proposals and the queries with both inter- and intra-video contrastive learning. For the inter-video contrastive learning, we follow the process in the glance branch to sample the negative video-query pairs, and jointly use the InfoNCE loss $\mathcal{L}_{nce}^f$ and triplet ranking loss $\mathcal{L}_{trip}^f$, which can be obtained by replacing $S^c(\cdot)$ with $S^f(\cdot)$ in Eq.(7) and Eq.(8), respectively. We denote the inter-video contrastive loss of the gaze branch as $\mathcal{L}_{inter}^{GZ}$, which is the weighted sum of the InfoNCE loss and triplet ranking loss $\mathcal{L}_{inter}^{GZ} = \mathcal{L}_{trip}^f + \beta_2 \mathcal{L}_{nce}^f$, where β_2 is the weight of the InfoNCE loss.

For the intra-video contrastive learning, since there are only three negative proposals for each positive proposal, we use the triplet ranking loss three times. Formally, we compute the intra-video contrastive loss as:

$$\begin{aligned} \mathcal{L}_{intra} = &\mathcal{L}_{trip}(S^f(q, v), S_{n1}(q, v), \Delta_2) \\ &+\mathcal{L}_{trip}(S^f(q, v), S_{n2}(q, v), \Delta_2) \\ &+\mathcal{L}_{trip}(S^f(q, v), S_{n3}(q, v), \Delta_3) \end{aligned} \tag{13}$$

where Δ_2 and Δ_3 are margins, and $\mathcal{L}_{trip}(\cdot)$ is the triplet ranking loss for paired video and query.

3.5 Training and Inference

Training. The overall training objective is the weighted sum of the losses in the glance branch and the gaze branch, which can be formulated as:

$$\mathcal{L} = \mathcal{L}_{inter}^{GLC} + \lambda_1 \mathcal{L}_{inter}^{GZ} + \lambda_2 \mathcal{L}_{intra} \tag{14}$$

where λ_1 and λ_2 are hyperparameters to balance training objectives.

Inference. After the model is trained, the gaze branch can independently align the semantics of the queries and videos, and no longer needs guidance from the glance branch. Hence, we replace the FGPG module with the CGPG module in the gaze branch during the inference phase. In this way, the glance branch and the gaze branch can give retrieval results independently.

Given a text query, we take a ranking-based approach to retrieve the desired video files and event moments synchronously. First, for the video-level retrieval subtask, we jointly use the coarse-grained similarity scores $S^c(q, v)$ and fine-grained similarity scores $S^f(q, v)$ to rank the videos. The final similarity score is given by:

$$S(q, v) = \alpha S^c(q, v) + (1 - \alpha) S^f(q, v) \tag{15}$$

where $\alpha \in [0, 1]$ is a hyperparameter for balancing the contribution of the two similarities.

Second, for the event-level retrieval subtask, we use the fine-grained similarity scores between the representations of the query and the moment proposals from the gaze branch to rank the candidate moments. To reduce the computational cost, we only choose the $top - k$ proposals in each video, where the k is set to 10 by default. Moreover, we perform Non-Maximum Suppression (NMS) [25] on the proposals to remove the redundant ones.

4 EXPERIMENTS

4.1 Implementation Details

Datasets. We conduct extensive experiments on three public datasets, including: **(1) Charades-STA:** [11] is a dataset for temporal activity localization via language query. It contains 6,672 videos of daily indoors activities. We follow the official data split, where

the training set and the testing set contain 12,408 and 3,720 moment-sentence pairs, respectively. **(2) ActivityNet-Captions:** [21] is a dataset for dense video captioning and video moment retrieval. It consists of 20K YouTube videos of various activities. For fairness, we use the common data split adopted in [5, 36] **(3) DiDeMo:** [1] is a dataset for text-based video-moment retrieval. It contains 10.6K videos with maximum lengths of 30 seconds, and splitted into 8,395, 1,065, and 1,004 for training, validation, and testing, respectively. Particularly, as our proposed MGVCR task aims at a practical cross-modal retrieval scenario, we follow the conventional Text-based Video Retrieval settings where the temporal annotations of video events are not available during training but only video-query pairs.

Evaluation Metrics. To evaluate the performance of our method on both video-level retrieval and event-level retrieval subtasks, we adopt the following metrics: **(1) Video-level Retrieval:** we use Recall@K as the evaluation metric following [5, 36]. Recall@K is the fraction of queries that correctly retrieve the desired video in the top K of the ranking list. Higher Recall@K means better performance. **(2) Event-level Retrieval:** we use Recall@K, IoU=m as the evaluation metric following [30, 36]. Recall@K, IoU=m denotes the percentage of test samples that have at least one predicted moment whose intersection over union (IoU) with the ground-truth moment is larger than m in the top-K predictions. We set K=10,100 and m=0.3,0.5,0.7. A prediction is correct if (1) the predicted video matches the ground truth video, and (2) the predicted moment has a high overlap with the ground truth moment, where temporal IoU is used to measure the overlap. Note that the sentence descriptions in the DiDeMo dataset have multiple temporal annotations, which is different from the others. We consider a predicted moment as correct if it overlaps with at least ground truth with a specified IoU.

Experimental Settings. We use the pre-trained I3D [3] backbone as our visual feature extractor for Charade-STA and ActivityNet-Captions datasets. For the DiDeMo dataset, we use the visual feature provided by [23], which is obtained by concatenating the extracted features using pre-trained ResNet-101 [16] and Slowfast[10]. We use pre-trained RoBERTa[26] to extract textual features, with the embedding dimensions of 1024 for the three datasets. During training, the balance factor β_1, β_2, λ_1, and λ_2 are set to 0.04, 0.04, 1, and 1 respectively. The n_{pc} and n_{pw} are set to 16 and 10 for the Charades-STA dataset, 32 and 10 for the ActivityNet-Captions dataset, and 6 and 6 for the DiDeMo dataset, respectively. The Δ_1 and Δ_3 are set to 0.2 for all three datasets. The Δ_2 is set to 0.2, 0.15, and 0.15 for the Charades-STA, ActivityNet-Captions, and DiDeMo datasets, respectively. The event proposals' relative upper bound w_{ub} is set to 1 in our experiments for all datasets, while the lower bound w_{lb} is set to 0.05, 0.05, 0.17 for the Charades-STA, ActivityNet-Captions, and DiDeMo datasets, respectively. We train our model on a single NVIDIA A6000 GPU with the Adam [20] optimizer with a learning rate 1.5×10^{-4} and the batch size is set to 128. We also provide the implementation with Huawei MindSpore toolkit at https://github.com/CFM-MSG/Code_JSG.

4.2 Performance Comparision

Comparison on Event-level Retrieval. As MGVCR is a challenging Text-based Video Retrieval task that is not fully explored yet, only a few VMR works can be directly used to retrieve events from a collection of videos without using temporal annotations for training, including CAL [9] and HMAN [30]. Hence, we select some video corpus moment retrieval methods that do not require early feature fusion processes, including the XML [22] and ReLoCLNet [36]. We re-train them with the temporal boundary information removed to fit our setting. We also compare with the early VMR method MCN [1], which is scaled up by Escorcia *et al.* [9] to retrieve events from a video collection. Besides, we modify the newly proposed untrimmed video retrieval method MS-SL [5] by adding an event retrieval strategy to it. Specifically, we take the top 10 most similar proposals in each video's clip branch proposals as the retrieved events and then evaluate the event-level retrieval performance. The experimental results are shown in Table 1 and 2.

Table 1: Performance comparison on Event-level Retrieval subtask on the ActivityNet-Captions dataset.

Method	IoU=0.3		IoU=0.5		IoU=0.7	
	R@10	R@100	R@10	R@100	R@10	R@100
MCN [1]	-	-	0.18	1.26	0.09	0.70
CAL [9]	-	-	0.21	1.58	0.10	0.90
XML [22]	3.21	12.48	1.69	7.58	0.10	0.90
HMAN [30]	-	-	0.66	4.75	0.32	2.27
ReLoCLNet [36]	4.82	15.8	3.01	11.22	1.47	6.3
MS-SL [5]	10.80	28.31	5.85	15.65	2.46	6.60
JSG (Ours)	**13.27**	**40.61**	**8.76**	**29.98**	**3.83**	**15.78**

From the results we can see that: (1) Our method outperforms the other methods by a huge margin on all of the three datasets. This demonstrates the superiority of our method of retrieving relevant events from different kinds of video collections. (2) We also observe that our method has a significant advantage over the existing methods on the ActivityNet-Captions dataset, which contains longer and more complex videos than the other two datasets. This shows that our method can handle the challenges of retrieving long and complex untrimmed videos better than the existing methods. Moreover, we notice that our method outperforms ReLoCLNet by a large margin, which indicates that our coarse-to-fine contrastive learning process and our unified joint embedding space are more effective than the two-stage retrieval process used by ReLoCLNet. (3) Compared with the latest video-level retrieval method MS-SL, when the IoU threshold increase from 0.3 to 0.7, the recall rate of our model drops slower than the MS-SL, which indicates that our method aligns the queries and videos more accurately.

Comparison on Video-level Retrieval. Retrieving untrimmed videos in the video collection is also a challenging task. We compare our method with the state-of-the-art model MS-SL [5], and some baseline models, and report the results in Table 3 and Table 4 for Charades-STA and ActivityNet-Captions, respectively. From Table 3 and 4, we can observe that: 1) Our performance is superior to counterpart methods on most metrics. Specifically, compared with the latest counterpart method MS-SL, our method achieves a 4.3% improvement on the Charades-STA dataset, which is more significant than the improvement on the ActivityNet-Captions dataset. We speculate the reason is there exists a number of queries that describe the whole video in the ActivityNet-Captions dataset, which leads to a degeneration of our grounding designs. 2) Compared with the conventional Text-based Video Retrieval method DE++ [7] and RIVRL [8], our method reveal remarkable superiority. It proves

Table 2: Performance comparison on Event-level Retrieval subtask on the Charades-STA and DiDeMo datasets. Note that we adopt the performance of MCN [1] reported in [9] for comparisons.

Method	Charades-STA						DiDeMo					
	IoU=0.3		IoU=0.5		IoU=0.7		IoU=0.3		IoU=0.5		IoU=0.7	
	R@10	R@100	R@10	R@100	R@10	R@100	R@10	R@100	R@10	R@100	R@10	R@100
MCN [1]	-	-	0.52	2.96	0.31	1.75	-	-	2.15	12.47	1.55	9.03
CAL [9]	-	-	0.75	4.39	0.42	2.78	-	-	3.90	16.51	2.81	12.79
XML [22]	0.7	2.47	0.32	1.42	0.16	0.78	2.4	5.65	0.99	3.73	0.41	1.63
ReLoCLNet [36]	1.51	3.28	0.94	2.26	0.59	1.21	5.36	16.08	3.15	11.78	1.51	6.3
HMAN [30]	-	-	1.40	7.79	1.05	4.69	-	-	6.25	28.39	4.98	22.51
MS-SL [5]	4.46	17.61	2.55	10.05	0.91	3.76	8.81	29.03	5.45	20.32	2.96	12.41
JSG (Ours)	**7.23**	**28.71**	**5.67**	**22.50**	**3.28**	**12.34**	**10.95**	**39.18**	**9.29**	**35.03**	**7.11**	**27.17**

that our method is more suitable for untrimmed video retrieval, which is more practical in real-world scenarios.

Table 3: Performance comparison on Video-level Retrieval subtask on the Charades-STA dataset.

Method	R@1	R@5	R@10	R@100	SumR
XML [22]	1.6	6.0	10.1	46.9	64.6
DE++ [7]	1.7	5.6	9.6	37.1	54.1
ReLoCLNet [36]	1.2	5.4	10.0	45.6	62.3
RIVRL [8]	1.6	5.6	9.4	37.7	54.3
MS-SL [5]	1.8	7.1	11.8	47.7	68.4
JSG (Ours)	**2.4**	**7.7**	**12.8**	**49.8**	**72.7**

Table 4: Performance comparison on Video-level Retrieval subtask on the ActivityNet-Captions dataset.

Method	R@1	R@5	R@10	R@100	SumR
XML [22]	5.3	19.4	30.6	73.1	128.4
DE++ [7]	5.3	18.4	29.2	68.0	121.0
ReLoCLNet [36]	5.7	18.9	30.0	72.0	126.6
RIVRL [8]	5.2	18.0	28.2	66.4	117.8
MS-SL [5]	**7.1**	22.5	34.7	75.8	140.1
JSG (Ours)	6.8	**22.7**	**34.8**	**76.1**	**140.5**

4.3 Ablation Study

Analysis on Model Structure and Ranking Strategy. We conduct ablation studies on our model structure by removing or changing some of the key components. Moreover, we explore the effect of different ranking strategies for video-level and event-level retrieval. The results are shown in Table 5. Our ablated models contain the following settings: (1) *w/o CBPG*: We remove the CBPG module in the glance branch and use a sliding window to generate proposals, following MS-SL [5]. (2) *w/o GWP*: We replace the GWP mechanism in both the Glance and Gaze branches with the popular used mean-pooling operation. (3) *w/o GD*: We remove the global downsampling strategy used in the video representation module so that the originally used coarse-grained video representation in the glance branch is replaced by the fine-grained ones. (4) *w/o GZ*: We remove the gaze branch, and only train the glance branch. (5) *GLC-sim*: Using the glance branch similarity scores to rank the retrieval results. (6) *GZ-sim*: Using the gaze branch similarity scores to rank the retrieval results. (7) *Sum-sim*: Using the sum of the similarity from two branches to rank the retrieval results.

From Table 5 we have the following observations: (1) The event-level retrieval performance is more sensitive to the model structure modifications, this is because the event-level retrieval is influenced by both the inter-video semantic alignment and the intra-video semantic alignment. (2) All of the explored components in our model contribute to the final performance. Particularly, the Gaze Branch has a significant influence on both the event-level and video-level retrieval performance, which demonstrates the effectiveness of the gaze branch in capturing the fine-grained semantic alignment. (3) Our proposed CBPG strategy is more effective than the sliding window strategy for our framework and the GWP mechanism plays a crucial role in the event-level retrieval among all the other components we explored. The reason is they help the whole model to capture the most representative features for each proposal. As for the ranking strategy, we find that the summed similarity is the best for video-level retrieval, while the gaze branch similarity is the best for event-level retrieval.

Table 5: Ablation on model components on the Charades-STA dataset. We set K = 10 for the Event-level Retrieval.

Method	Event-level			Video-level	
	IoU=0.3	IoU=0.5	IoU=0.7	R@10	R@100
w/o CBPG	5.35	4.11	1.96	11.9	49.5
w/o GWP	3.06	1.80	0.62	12.7	49.0
w/o GD	5.11	3.90	2.18	12.2	49.1
w/o GZ	4.06	2.96	1.75	9.30	44.6
Full Model	**7.23**	**5.67**	**3.28**	**12.8**	**49.8**
GLC-sim	3.41	2.53	1.56	9.8	44.5
GZ-sim	**7.23**	**5.67**	**3.28**	12.5	49.2
Sum-sim	6.29	4.95	2.72	**12.8**	**49.8**

Analysis on Training Objectives To explore the effect of different training objectives, we conduct ablation studies on the Charades-STA dataset. We train the model with different combinations of the proposed training objectives and report the results in Table 6. When only using the glance branch inter-video contrastive loss, the model is trained based on the coarse-grained video features without learning the intra-video alignment. In this case, we only get a basic model that performs poorly on both the event-level retrieval and video-level retrieval subtasks. After adding the gaze branch inter-video contrastive loss, the model is trained on both the coarse-grained video features and the fine-grained ones, which improves the retrieval performance on both of the subtasks significantly. When adding the gaze branch intra-video contrastive loss, the fine-grained video features are learned to be more discriminative for the proposals in the same video. However, it has negative influence on the alignment learning of the inter-video proposals,

Figure 3: Visualization of MGVCR results, where the two cases are selected from the test split of the Charades-STA dataset.

which is crucial for the video-level retrieval subtask. Therefore, the intra-video contrastive loss is only used for event-level retrieval.

Table 6: Ablation on training objectives on the Charades-STA dataset. We set $K = 10$ for the Event-level Retrieval.

L_{inter}^{GLC}	L_{inter}^{GZ}	L_{intra}	Event-level		Video-level	
			IoU=0.3	IoU=0.5	R@10	R@100
✗	✗	✗	0.08	0.03	0.6	7.6
✓	✗	✗	4.49	3.36	9.8	47.3
✓	✓	✗	6.26	4.44	**12.8**	**49.8**
✓	✗	✓	4.44	3.31	11.8	48.4
✓	✓	✓	**7.23**	**5.67**	**12.8**	**49.8**

Further Analysis on Ranking Strategies. We further explore the three ranking strategies, *i.e.*, glance branch similarity, gaze branch similarity and sum similarity score, for event-level retrieval and video-level retrieval on the Charades-STA dataset. For clarity, we abbreviate the three ranking strategies as GLC-sim, GZ-sim, and Sum-sim respectively. We compare the performance for event-level retrieval by giving the R@10 and R@100 results over different IoU thresholds, as shown in Figure 4. According to the results, we can see that: (1) The GZ-sim strategy performs better than the other two strategies over all the IoU thresholds, which demonstrates the consistency of the ranking strategy for event-level retrieval; (2) The Sum-sim strategy lies between the GLC-sim strategy and the GZ-sim strategy, which demonstrates that the gaze branch is enough to capture the subtle difference between the proposals.

Figure 5 plots the video-level retrieval results on more metrics for the three ranking strategies. From the figure we have the following observations: (1) The GZ-sim strategy performs better than the GLC-sim strategy on all the metrics, which demonstrates that our fine-grained hybrid contrastive learning strategy is effective for video-level retrieval; (2) The Sum-sim strategy has the best performance on most of the metrics except for R@5, which indicates that the glance branch and gaze branch are complementary to each other for video-level retrieval.

Qualitative Results We visualize two retrieval results of our method on the Charades-STA dataset in Figure 3. Specifically, we show the top-3 retrieved videos for each query and mark the predicted events with a blue box in each of the videos. From the figure we can see that our method retrieved the video with the correct events, and the predicted events are well aligned with the text query. Moreover, it is interesting to see that, the top2 and top3 retrieved videos also contain events relevant to the query, which demonstrates the effectiveness of our method.

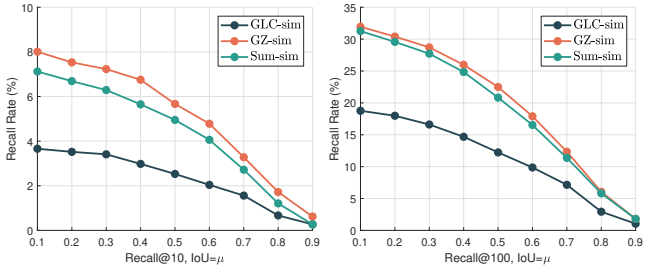

Figure 4: Comparison of the ranking strategies for event-level retrieval.

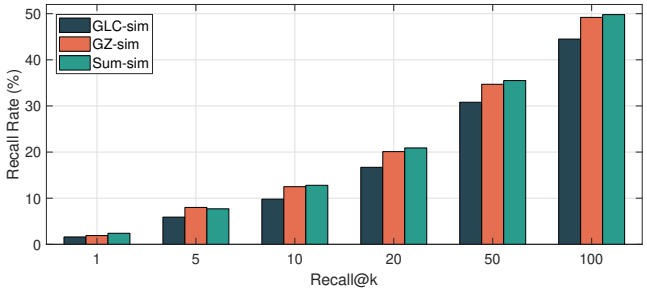

Figure 5: Comparison of the ranking strategies for video-level retrieval.

5 CONCLUSIONS

In this paper, we proposed a practical text-based video retrieval paradigm which is not fully explored yet, *i.e.*, Multi-Granularity Video Content Retrieval (MGVCR). It aims at synchronously retrieving videos and specific video content from a large untrimmed video collection with given text queries. To address this challenging task, we designed a novel unified video retrieval framework, termed Joint Searching and Grounding (JSG), which effectively overcame the redundancy of video content and the absence of temporal supervision through a glance-to-gaze process and synergistic contrastive learning strategy. Extensive experiments on three public benchmarks demonstrated the superiority of our proposed JSG method. In the future, we will furtherly explore video content understanding and retrieval and facilitate the development of multimedia applications.

6 ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (62222203, 61976049) and CAAI-Huawei MindSpore Open Fund.

REFERENCES

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision*. 5803–5812.

[2] Taivanbat Badamdorj, Mrigank Rochan, Yang Wang, and Li Cheng. 2022. Contrastive Learning for Unsupervised Video Highlight Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14042–14052.

[3] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[4] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. 2020. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10638–10647.

[5] Jianfeng Dong, Xianke Chen, Minsong Zhang, Xun Yang, Shujie Chen, Xirong Li, and Xun Wang. 2022. Partially Relevant Video Retrieval. In *Proceedings of the ACM International Conference on Multimedia*. 246–257.

[6] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. 2019. Dual encoding for zero-example video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9346–9355.

[7] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. 2021. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[8] Jianfeng Dong, Yabing Wang, Xianke Chen, Xiaoye Qu, Xirong Li, Yuan He, and Xun Wang. 2022. Reading-strategy inspired visual representation learning for text-to-video retrieval. *IEEE Transactions on Circuits and Systems for Video Technology* (2022).

[9] Victor Escorcia, Mattia Soldan, Josef Sivic, Bernard Ghanem, and Bryan Russell. 2019. Temporal localization of moments in video collections with natural language. (2019).

[10] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE International Conference on Computer Vision*. 6202–6211.

[11] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision*. 5267–5275.

[12] Junyu Gao and Changsheng Xu. 2021. Fast video moment retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*. 1523–1532.

[13] Junyu Gao and Changsheng Xu. 2021. Learning video moment retrieval without a single annotated video. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 3 (2021), 1646–1657.

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9726–9735.

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9726–9735.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 770–778.

[17] Xun Jiang, Xing Xu, Jingran Zhang, Fumin Shen, Zuo Cao, and Heng Tao Shen. 2022. SDN: Semantic Decoupling Network for Temporal Language Grounding. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[18] Xun Jiang, Xing Xu, Jingran Zhang, Fumin Shen, Zuo Cao, and Heng Tao Shen. 2022. Semi-Supervised Video Paragraph Grounding With Contrastive Encoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2466–2475.

[19] Xun Jiang, Zailei Zhou, Xing Xu, Yang Yang, Guoqing Wang, and Heng Tao Shen. 2023. Faster Video Moment Retrieval with Point-Level Supervision. *arXiv preprint arXiv:2305.14017* (2023).

[20] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.

[21] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-Captioning Events in Videos. In *Proceedings of the IEEE International Conference on Computer Vision*. 706–715.

[22] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *European Conference on Computer Vision*. Springer, 447–463.

[23] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. HERO: Hierarchical Encoder for Video+Language Omni-representation Pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 2046–2065.

[24] Xirong Li, Chaoxi Xu, Gang Yang, Zhineng Chen, and Jianfeng Dong. 2019. W2vv++ fully deep learning for ad-hoc video search. In *Proceedings of the ACM International Conference on Multimedia*. 1786–1794.

[25] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. 2018. Bsn: Boundary sensitive network for temporal action proposal generation. In *European Conference on Computer Vision*. 3–19.

[26] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[27] Jonghwan Mun, Minsu Cho, and Bohyung Han. 2020. Local-Global Video-Text Interactions for Temporal Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10807–10816.

[28] Ryumei Nakada, Halil Ibrahim Gulluk, Zhun Deng, Wenlong Ji, James Zou, and Linjun Zhang. 2023. Understanding multimodal contrastive learning and incorporating unpaired data. In *International Conference on Artificial Intelligence and Statistics*. 4348–4380.

[29] Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. 2022. Probabilistic representations for video contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14711–14721.

[30] Sudipta Paul, Niluthpol Chowdhury Mithun, and Amit K Roy-Chowdhury. 2021. Text-based localization of moments in a video corpus. *IEEE Transactions on Image Processing* 30 (2021), 8886–8899.

[31] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 815–823.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[33] Gongmian Wang, Xing Xu, Fumin Shen, Huimin Lu, Yanli Ji, and Heng Tao Shen. 2022. Cross-modal dynamic networks for video moment retrieval with text query. *IEEE Transactions on Multimedia* 24 (2022), 1221–1232.

[34] Yinwei Wei, Xiang Wang, Qi Li, Liqiang Nie, Yan Li, Xuanping Li, and Tat-Seng Chua. 2021. Contrastive Learning for Cold-Start Recommendation. In *Proceedings of the ACM International Conference on Multimedia*. 5382–5390.

[35] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. 2022. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15671–15680.

[36] Hao Zhang, Aixin Sun, Wei Jing, Guoshun Nan, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. 2021. Video corpus moment retrieval with contrastive learning. In *Proceedings of the ACM International Conference on Research & Development in Information Retrieval*. 685–695.

[37] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020. Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language. In *AAAI Conference on Artificial Intelligence*. 12870–12877.

[38] Zhu Zhang, Zhou Zhao, Zhijie Lin, Jieming Zhu, and Xiuqiang He. 2020. Counterfactual Contrastive Learning for Weakly-Supervised Vision-Language Grounding. In *Advances in Neural Information Processing Systems*. 18123–18134.

[39] Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu. 2022. Weakly Supervised Video Moment Localization with Contrastive Negative Sample Mining. In *AAAI Conference on Artificial Intelligence*. 3517–3525.

[40] Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu. 2022. Weakly Supervised Temporal Sentence Grounding With Gaussian-Based Contrastive Proposal Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15555–15564.

[41] Ding Zou, Wei Wei, Xian-Ling Mao, Ziyang Wang, Minghui Qiu, Feida Zhu, and Xin Cao. 2022. Multi-level Cross-view Contrastive Learning for Knowledge-aware Recommender System. In *Proceedings of the ACM International Conference on Research & Development in Information Retrieval*. 1358–1368.