Learning Commonsense-aware Moment-Text Alignment for Fast Video Temporal Grounding

ZIYUE WU, Tianjin University of Technology, Tianjin, China
JUNYU GAO, Chinese Academy of Sciences Institute of Automation, Beijing, China
SHUCHENG HUANG, Jiangsu University of Science and Technology, Zhenjiang, China
CHANGSHENG XU, National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China

Grounding temporal video segments described in natural language queries effectively and efficiently is a crucial capability needed in vision-and-language fields. In this paper, we deal with the fast video temporal grounding (FVTG) task, aiming at localizing the target segment with high speed and favorable accuracy. Most existing approaches adopt elaborately designed cross-modal interaction modules to improve the grounding performance, which suffer from the test-time bottleneck. Although several common space-based methods enjoy the high-speed merit during inference, they can hardly capture the comprehensive and explicit relations between visual and textual modalities. In this paper, to tackle the dilemma of speed-accuracy tradeoff, we propose a commonsense-aware cross-modal alignment network (C$^2$AN), which incorporates commonsense-guided visual and text representations into a complementary common space for fast video temporal grounding. Specifically, the commonsense concepts are explored and exploited by extracting the structural semantic information from a language corpus. Then, a commonsense-aware interaction module is designed to obtain bridged visual and text features by utilizing the learned commonsense concepts. Finally, to maintain the original semantic information of textual queries, a cross-modal complementary common space is optimized to obtain matching scores for performing FVTG. Extensive results on two challenging benchmarks show that our C$^2$AN method performs favorably against state-of-the-arts while running at high speed. Our code is available at https://github.com/ZiyueWu59/CCA.

CCS Concepts: · Computing methodologies → Activity recognition and understanding; · Theory of computation → Semantics and reasoning.

Additional Key Words and Phrases: Fast video temporal grounding, Commonsense knowledge, Commonsense-aware interaction, Complementary common space.

1 INTRODUCTION

Video Temporal Grounding (VTG) task aims to localize the temporal segment in a video that is semantically aligned with the given language query. It has various applications such as robotic navigation, video surveillance/entertainment, autonomous driving, etc. Recently, many approaches [15, 19, 41, 42, 48, 49, 56, 58, 59, 61, 64, 69, 77, 79, 81–86] have been proposed for the VTG task. Due to the large cross-modal gap between video and text in the VTG task, existing approaches [16, 41, 48, 73, 80, 83, 88] mainly focus on improving the accuracy of

Authors’ Contact Information: Ziyue Wu, Tianjin University of Technology, Tianjin, China; e-mail: wuziyuewzy@gmail.com; Junyu Gao, Chinese Academy of Sciences Institute of Automation, Beijing, China; e-mail: gaojunyu@ia.ac.cn; Shucheng Huang, Jiangsu University of Science and Technology, Zhenjiang, China; e-mail: shuang@163.com; Changsheng Xu, National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China; e-mail: csxu@nlpr.ia.ac.cn.

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.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM 1551-6865/2024/5-ART
https://doi.org/10.1145/3663368

ACM Trans. Multimedia Comput. Commun. Appl.
localization by designing complicated cross-modal interaction operations. Although they have achieved excellent results on several public datasets, only a few [1, 18] take notice of the test-time cost. In fact, the test-time is an essential metric in many practical applications that need an efficient mechanism to respond quickly when receiving a natural language query. For instance, when enjoying intelligent robot service, the customer may feel impatient or even angry due to slow response and may not employ similar services in the future. Moreover, the response time is more critical in autonomous driving, which always means whether it is safe or not.

Recently, fast video temporal grounding (FVTG) [18] is proposed for accurate temporal localization and an efficient test process. Note that the current VTG pipeline can be divided into three components: video encoder, text encoder, and cross-modal interaction module. For the video/text encoders, to obtain more effective information, some traditional modules are widely adopted in most VTG methods for encoding different modality information, e.g., I3D [4] C3D [60] for visual encoding and BiLSTM [33], GRU [9] for text encoding.

The most significant component of the VTG is the interaction between different modalities. As a result, existing methods utilized attention [62] or transformer [83], graph neural networks [82, 85] and temporal adjacent networks [84] to conduct interaction. Although bringing rich cross-modal interaction information, this module always consumes the majority of the test-time due to complex feature matrix interaction operation [48, 77, 79] or transformations [25]. Different from the above approaches, to calculate the similarity scores between video moments and texts, common space is utilized in FVTG [18], where the efficient vector operations like dot production between different modality features are conducted. As a result, the common space-based approaches can achieve a significant test speed. For example, Gao and Xu [18] propose a fast video moment retrieval approach, which achieves not only about 35 times faster than the state-of-the-arts in the cross-modal learning process but also favorable results by employing a fine-grained semantic distilling framework. However, such common space-based methods still have a problem restricting their localization accuracy. That is, they cannot establish the interaction process explicitly between two different modalities, which leads to the features learned in common space hardly capturing the complex relation between them.

Intuitively, without a well-designed cross-modal interaction, it is difficult to ground a textual query onto the video effectively since the information deficiency is brought by inadequate interactions between visual and textual modalities. This deficiency further leads to the lack of learning discrimination of different modality features. Therefore, how to effectively mine the complex cross-modal relation while still maintaining efficient moment-text alignment in the common space is a key issue in FVTG task.

We notice that humans can accurately capture this kind of complex cross-modal relationship in the real world. The reason is that humans have the ability to comprehend and connect the video with the text by using their
commonsense knowledge of the world learned through experience, which can be expressed by core semantic concepts and the correlations between them. As shown in Figure 1, given a query sentence, “A man enters the kitchen, removes a frying pan from the drawer and places it on the stove”. When the word “kitchen” appears, the words “pan” and “stove” are more likely to appear in the text. Moreover, seeing such words in a sentence also brings people’s minds to visualize objects and their associations with related entities. e.g., when seeing the word “pan”, it always reminds us to imagine the visual appearance of “kitchen”, “stove” and other similar words. That is, the co-occurrence of “kitchen”, “pan”, and “stove” consists of a kind of commonsense knowledge in both visual and textual modalities. Until now, such commonsense knowledge has not been studied and exploited for the VTG task.

Motivated by the above observation, we propose a Commonsense-aware Cross-modal Alignment Network (C2AN) for the fast video temporal grounding task, which incorporates commonsense-guided visual and text representations into a complementary common space and learns efficient commonsense-aware cross-modal alignment. To obtain appropriate commonsense knowledge concepts, we utilize the natural language queries to perform statistics and select a set of representative words as concepts. Besides, the relations between these concepts are also considered as a graph, and a concept extractor is employed to obtain commonsense concept features. After that, a commonsense-aware interaction module is designed on both moment and text sides, which can obtain commonsense-guided visual and text features by adaptively and attentively using concept features to bridge the cross-modal gap and make up for the information deficiency in further common space learning. The learned concept representation could be pre-extracted and stored in an offline manner, which does not affect the test time. To maintain the original semantic information of textual queries, a cross-modal complementary common space is learned to obtain matching scores for all moment proposals and rank them to select the best matching ones. Experiments on two benchmarks show that our proposed framework outperforms other state-of-the-art competitors on overall performance comparison and microscope analyses.

The main contributions of this paper are summarized as follows:

(1) To the best of our knowledge, we are among the first to exploit commonsense information for fast VTG task without adding additional cross-modal interactions while achieving favorable performance.

(2) A commonsense-aware interaction module is designed to efficiently leverage the extracted commonsense information, which could serve as a bridge between visual and text modalities. Here, two concept-guided attention modules are utilized to enhance visual and text features for cross-modal alignment, respectively.

(3) Extensive experiments are conducted on three challenging datasets: TACoS, ActivityNet Captions and Charades-STA. The experimental results demonstrate that our proposed method performs superior to the state-of-the-art methods and almost has the highest speed.

2 RELATED WORK

2.1 Video Temporal Grounding

Video Temporal Grounding task aims to localize the most relevant video segment corresponding to the given sentence query. VTG has been gaining popularity in the past few years due to its enormous potential applications in video comprehension, to name a few [6, 17, 20, 82]. Existing state-of-the-art VTG methods can be divided into three categories: proposal-based methods, proposal-free methods and reinforcement-learning-based method.

Proposal-based Methods. Most Proposal-based methods [15, 16, 23, 43, 71, 73, 77, 80, 81, 84, 85] are based on several well-designed dense sample strategies, which obtain a set of video segments as candidate proposals and rank them according to the similarity scores calculated between the proposals and the query to select the best matching pairs. Sliding windows is widely used in previous work like CTRL[15], MCN[1], ACRN[44] and ACL[23]. Gao et al. [15] first conceive this problem and propose a Cross-modal Temporal Regression Localizer (CTRL) framework by utilizing sliding windows to generate temporal proposals for localization. To improve the
quality of the generated moment proposals, Yuan et al. [77] propose a Semantic Conditioned Dynamic Modulation (SCDM), which can dynamically adjust the temporal convolution according to the query semantics. Besides, rich temporal information is taken into consideration in some work. Zhang et al. [81] propose a Moment Alignment Network (MAN) to model complex temporal relations in a video by explicitly establishing relations between different moments and structuring them into a graph for localization in an end-to-end manner. Zhang et al. [84] make full use of the temporal context from different moment proposals by structuring a 2D temporal map to capture the temporal relations between different video moments. For more fine-grained interaction, Zhang et al. [85] propose a Cross-Modal Interaction Network (CMIN), which leverages syntactic structures for fine-grained feature learning and utilizes multi-stage cross-modal interaction to obtain the potential relations between visual and text modalities. In addition, Gao et al. [16] regard VTG task as video reading comprehension and propose the Relation-aware Network (RaNet). Currently, most proposal-based methods are time-consuming due to the large number of proposal-query interactions.

**Proposal-free Methods.** In fact, the impressive performance achieved by proposal-based methods largely depends on the quality of the sampled proposals. Instead of generating numbers of moment proposals as candidates, proposal-free methods [6, 25, 39, 41, 48, 56, 62, 78, 79, 88] directly regress or predict the starting and ending time of the target moment to reduce the extra computational cost brought by the generation of proposal features, such as [48, 79]. Also, taking local and global information into account, Liu et al. [42] propose a Context-aware Biaffine Localizing Network (CBLN) that incorporates local and global contexts into the boundaries with position information for biaffine-based localization. Besides, Wang et al. [64] aggregate contextual information by obtaining the relations between the current segment and its neighbor segments and propose a Contextual Boundary-aware Prediction (CBP). By addressing VTG task with a span-based QA method, Zhang et al. [82] propose a Video Span Localizing Network (VSLNet), which utilizes a query-guided highlighting strategy for matching video segment in the highlighted region. Considering the off-balance data distribution, Liu et al. [41] propose a Memory-Guided Semantic Learning Network (MGSL-Net) to alleviate the forgetting issue by a memory argumentation module. Although these proposal-free methods are more likely to perform effectively and accurately due to the excellent cross-modal interaction and regression operation, most of them regress only one target moment, which does not match the requirements of practical tasks.

To improve the localization performance, a lot of existing work tends to apply more complex cross-modal interaction operations. Wang et al. [62] propose a Structured Multi-level Interaction Network (SMIN) that utilizes three submodules for the interaction between the visual proposals and different scales of the text representations and uses them iteratively. Zhang et al. [83] propose a Multi-stage Aggregated Transformer Network (MATN) in which a multi-stage cross-modal interaction module is designed by complex attention operation. By reformulating VTG task as a set prediction task, Cao et al. [2] propose a multi-modal transformer framework (GTR) by leveraging a visual-language transformer-based backbone to obtain discriminative representations for moment localization. In addition, fine-grained semantic information has been gradually explored in VTG task. Chen et al. [8] explicitly structure a sentence in three different semantic levels and a graph neural network is used to obtain different level semantic information. Though methods like MATN, SMIN, and GTR have achieved excellent performances on several public datasets, the complex cross-modal interaction operations of them also result in a huge time cost during testing.

**Reinforcement-learning-based Method.** With reinforcement learning gradually becoming more popular in recent years, several new reinforcement-learning-based VTG methods [30, 68, 70] have also been proposed. Wu et al. [70] formulate a Tree-Structured Policy based Progressive Reinforcement Learning framework (TSP-RPL), which consecutively regulates the predicted boundaries by an iterative refinement process. Besides, the Semantic Matching Reinforcement Learning model (SM-RL) is proposed [68] to obtain semantic concepts by reinforcement learning and fuse them with context features. Motived by Visual Question Answering tasks (VQA), He et al. [30]...
regard VTG task as a problem of sequential decision and design a RWM-RL method to regulate the boundaries of predicted results based on its policy.

Nevertheless, with the improvement of performance, larger and more complex architectures inevitably result in higher computational cost during the test phase. Therefore, in this paper, a common space is learned in our proposed method for fast video temporal grounding.

### 2.2 Fast Video Temporal Grounding

Recently, fast video temporal grounding (fast VTG) has been proposed for more practical applications. VTG task usually requires methods to efficiently localize target video segments in thousands of candidate proposals. In fact, several early algorithms, e.g., common space-learning methods [1, 32] and skip scanning-based method [28], make some contribution to reducing computational cost. Gao et al. [18] explore the fast VTG formally and achieve great performance. According to [18], the standard VTG pipeline can be divided into three components. The visual encoder and the text encoder are proved to have little influence in model testing due to the features pre-extracted and stored at the beginning of the test, and cross-modal interaction is the key to reducing the test-time. In [18], a common space learning paradigm is designed to speed up the model. Moreover, a fine-grained semantic distillation framework is utilized to leverage semantic information for improving performance. However, though fine-grained semantic information is used to strengthen the learning of common space, FVMR cannot establish the interaction process explicitly between two different modalities, which leads to the information deficiency in cross-modal relationships. In our proposed method, commonsense knowledge is utilized to obtain bridged visual and text representations. Furthermore, we also design a complementary common space learning strategy to make our model effectively compute matching scores while maintaining the original semantic information of the two modalities. It is the commonsense knowledge and the designed complementary common space learning strategy that make the visual and the textual modalities have the ability to promote each other and make our method perform better.

### 2.3 Knowledge Based Visual Related Tasks

Recently, the usage of knowledge has become a very active field of computer vision by providing external information and efficiently improving the capability of the model [21, 22, 26, 27, 31, 34–36, 54, 65, 72, 87, 89]. Knowledge information has been widely used in computer vision tasks. For scene understanding, Gu et al. [26] utilize external knowledge and image reconstruction loss to overcome the noisy and missing annotations in datasets. Zheng et al. [89] obtain knowledge graphs by extracting abundant visual concepts and then combining DNN structures with professional knowledge for scene understanding. To achieve fine-grained image classification, He et al. [31] propose a Knowledge Graph Representation Fusion (KGRF) framework by using prior knowledge. Besides, Rambhatla et al. [54] propose a working and semantic memory framework to discover unknown categories when prior knowledge is known. Moreover, a lot of cross-modal comprehension tasks also pay attention to it, such as VQA [46, 65] and image classification [47]. There also has been a growing focus on video-text retrieval [3, 13, 14, 24, 63, 66, 67, 74]. Some of them acquire external knowledge from pre-trained models [13, 24, 74]. Xu et al. [74] propose a transformer-based framework called VideoCLIP, which aims to achieve great zero-shot video understanding by establishing fine-grained associations between video and text. To explicitly model the temporal dynamics of video inputs, Fu et al. [13] design a video-language transformer with a new pre-training task named Masked Visual-token Modeling. Moreover, Ge et al. [24] design a novel pretext task named Multiple Choice Questions for video-text pre-training. Different from them, Cao et al. [3] explore the visual consensus by structuring them into a graph, and propose a Visual Consensus Modeling (VCM) framework. In image-text matching, Wang et al. [63] propose a Consensus-aware Visual-Semantic Embedding (CVSE) model to mine consensus information in image-text retrieval. Compared with VCM, we extract commonsense concepts in an offline manner from the text set structured by all the annotations from a specific training dataset, and the process doesn’t require any visual information. Moreover, VCM utilizes an online manner to extract the consensus.
Fig. 2. Overview of our proposed C2AN framework. C2AN mainly consists of three components: multi-modal feature extractor, commonsense-aware interaction module, and complementary common space. We utilize the multi-modal feature extractor to extract visual, text, and commonsense concept features. Then in the commonsense-aware interaction module, two attention-based structures are used to obtain commonsense-guided visual and text features. After that, we map these guided features into two common spaces to calculate the matching scores for each proposal, and we leverage a residual mechanism to obtain final scores. Finally, we rank the scores of all proposals, and a BCE loss is used to optimize the whole framework.

representations from each video-text pair input, which means more time-consuming. Compared with CVSE, our proposed C2AN method is the first one to leverage commonsense knowledge for temporal modeling in the video temporal grounding task. C2AN takes the temporal information in video and text into consideration, while CVSE only fuses consensus features with visual and textual features, respectively. Besides, the commonsense concepts extracted in C2AN are from the text set structured of the text annotations of the specific training datasets, while CVSE utilizes large-scale external knowledge to obtain richer knowledge information. Moreover, CVSE categorizes concepts into three types for more detailed information, while C2AN selects concepts only based on their frequencies.

3 METHODOLOGY

In this section, we first introduce the problem formulation of video temporal grounding task and the general scheme of our proposed framework. Then, we present each component of our framework, including multi-modal feature extractor, commonsense-aware interaction module, and complementary common space, as shown in Figure 2. Finally, the training and inference settings of our framework are present.

3.1 Problem Formulation

Given an untrimmed video \( V \) and a natural language query \( Q \), the goal of video temporal grounding is to localize the video segment \((t_s, t_e)\) that is most relevant to the query, where \( t_s \) and \( t_e \) denote the start and the end time of the target moment. The video is denoted as a sequence of frames \( V = \{v_i\}_{i=1}^T \), where \( T \) is the number of all the frames, and the corresponding text query is denoted as \( Q = \{w_j\}_{j=1}^L \), where \( L \) is the length of the query and \( w_j \) is the \( j \)-th word in the query.
3.2 General Scheme
Aiming to solve the problem that existing VTG methods always resort to complex and time-consuming cross-modal interaction modules to obtain rich information, we propose a commonsense-aware cross-modal interaction framework to make full use of commonsense knowledge. In this way, we can obtain more accurate results on the VTG task without adding additional cross-modal interactions. Figure 2 illustrates the architecture of our proposed C2AN method. Our approach consists of three components: a multi-modal feature extractor, a commonsense-aware interaction module, and a complementary common space. We first utilize three encoders to extract video, query, and commonsense features. Then a commonsense-aware interaction module is designed to generate commonsense-guided visual and textual representations. After that, we leverage two different attention-based structures to adaptively interact the extracted commonsense features with the video and text features, respectively. Finally, a complementary common space is learned to obtain commonsense-guided moment-query matching scores for temporal localization.

3.3 Multi-modal Feature Extractor
In the multi-modal feature extractor, three types of encoders are utilized to extract visual, text and commonsense features, respectively.

**Visual encoder.** We firstly generate a set of video moment candidate proposals denoted as \( P = \{p_n\}_{n=1}^N \), where \( p_n \) is the \( n \)-th proposal and \( N \) means the number of all generated proposals. Here, we simply adopt the commonly used 2D-temporal proposal generation approach [84]. Then, a pre-trained CNN model (e.g. I3D, C3D) is utilized to extract visual features for each proposal as follows:

\[
P = \text{visEncoder}(P) = \{p_1, p_2, ..., p_N\},
\]

where \( P \in \mathbb{R}^{N \times d^V} \) and \( d^V \) represents the dimension of the extracted feature. \( p_n \) means the visual features of proposal \( p_n \).

**Text encoder.** For the text query, we leverage a Bi-LSTM [33] to integrate the sequential information of the word list as follows:

\[
Q = \text{BiLSTM}(Q) = \{q_1, q_2, ..., q_L\},
\]

where \( Q \in \mathbb{R}^{L \times d^Q} \) and \( d^Q \) means the dimension of the extracted query feature. We obtain the word-level feature \( q_i \), where \( q_i = [\overrightarrow{q_i}; \overleftarrow{q_i}] \) through the concatenation of hidden states in both directions of Bi-LSTM, and calculate the sentence-level feature \( q = [\overrightarrow{q}; \overleftarrow{q}] \) by concatenating the last hidden state of both forward and backward LSTM.

**Commonsense Concept Encoder.** To capture the high-level relation information as in human reasoning, which is referred to as "commonsense knowledge concept", we choose some high-frequency words to serve as our commonsense concept. Specifically, we follow [12] to select \( M \) frequent concepts as commonsense concepts in the language corpus. After getting these concepts, we use Glove-300 [51] to initialize the concepts denoted as \( C = \{c_1, c_2, ..., c_M\} \), where \( M \) is the number of concepts. According to [63], we use both normalized co-occurrence frequency and word embedding similarity to initialize the concept relation graph \( G \). Finally, a two-layer Graph Convolution Network (GCN) is employed to extract commonsense concept features as follows:

\[
H^{(l+1)} = f(H^{(l)}, A) = \sigma(AD^{-1/2}GW^{(l)}H^{(l)}),
A = D^{-1/2}GD^{-1/2},
\]

where \( D \) represents degree matrix which is used to normalize \( G \). \( W^{l} \) is a learnable weight matrix. \( \sigma \) is a non-linear activation function, such as ReLU.
After GCN, the node representation $\mathbf{H}^{(l+1)} \in \mathbb{R}^{M \times d^C}$ produced by the last graph convolutional layer is regarded as commonsense concept features $\mathbf{C}$, where $\mathbf{C} = \mathbf{H}^{(l+1)} = \{c_1, c_2, ..., c_M\}$ and $d^C$ means the dimension of concept features.

### 3.4 Commonsense-aware Interaction Module

In traditional approaches, there is no mechanism to explore the commonsense knowledge for the cross-modal bridging in the VTG task. To address this drawback, we construct a commonsense-aware interaction module to achieve the goal by associating the extracted commonsense concept with both visual and textual representations. After encoding the moment proposals, sentence, and commonsense concepts, two adaptive attention-based Visual-Commonsense Interaction Module and Text-Commonsense Interaction Module are designed for constructing cross-modal alignment.

#### Visual-Commonsense Interaction Module

Recently, the attention mechanism is widely used in cross-modal learning tasks for modality interaction. However, with the complex temporal relations between different proposals and the high-level relations in commonsense concepts, a simple soft attention mechanism that captures interaction from one specific attention space may not be enough to achieve comprehensive information passing between different modalities. Therefore, a visual-commonsense interaction module is designed for robust relation modeling and better interaction between visual proposals and commonsense concepts. Specifically, we first concatenate moment proposals $\mathbf{P}$ with concept features $\mathbf{C}$ into a unified sequence:

$$F^{\text{Cat}} = \text{Concat}(\mathbf{P}, \mathbf{C})$$

where $F^{\text{Cat}}$ means that $\mathbf{P}$ and $\mathbf{C}$ are concatenated in their spatial dimension, $F^{\text{Cat}} \in \mathbb{R}^{(N+M) \times d^V}$. Specially, $d^C$ is set the same as $d^V$. After that, a multi-head self-attention module is employed to process this long sequence appropriately and efficiently, which could allow our model to jointly attend to the information from different positions and capture richer characteristics from diverse modalities.

$\mathbf{A}^{\text{Cat}} = F^{\text{Cat}} \mathbf{W}^{Q}_{i} (F^{\text{Cat}} \mathbf{W}^{K}_{i})^{\top}$,

$\text{head}_{i} = \text{softmax}(\frac{\mathbf{A}^{\text{Cat}}}{\sqrt{d^{\text{avg}}}}) F^{\text{Cat}} \mathbf{W}^{V}_{i}$,

$$F^{\text{mul}} = \text{Concat}(\text{head}_1, \text{head}_2, ..., \text{head}_n) \mathbf{W}^{\text{mul}}$$

where $\mathbf{W}^{Q}_{i}, \mathbf{W}^{K}_{i}$ and $\mathbf{W}^{V}_{i} \in \mathbb{R}^{d^V \times d^{\text{avg}}}$ are learnable weight matrices in the $i$-th head of the multi-head self-attention, and $d^{\text{avg}} = (N + M)/n$, $n$ is the number of the parallel heads. $F^{\text{mul}}$ represents the output of the multi-head structure, and $\mathbf{W}^{\text{mul}} \in \mathbb{R}^{d^{\text{avg}} \times d^V}$ is also a learnable weight matrix.

After exchanging information between our concatenated feature $F^{\text{cat}}_{i}$ at all spatial positions and obtaining $F^{\text{mul}}$, we further leverage two linear layers and a layer normalization operation to generate the high-level multi-head feature, which serves as a residue to $F^{\text{mul}}$. Finally, we can obtain visual-commonsense feature $F^{\text{cg}}$ as follows:

$$F^{\text{cg}} = F^{\text{mul}} + \text{LayerNorm}(\mathbf{W}^{\theta}(\text{ReLU}(\mathbf{W}^{\vartheta}F^{\text{mul}}) + \mathbf{b}_{p}) + \mathbf{b}_{q})$$

where $\mathbf{W}^{\theta}, \mathbf{W}^{\vartheta}, \mathbf{b}_{p}$ and $\mathbf{b}_{q}$ are the learnable parameters of two linear layers.

After normalizing the visual-commonsense feature, we select the top-$N$ dimension of $F^{\text{cg}}$ to generate commonsense-guided moment proposal features, denoted as $\hat{\mathbf{P}}$:

$$\{\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2, ..., \hat{\mathbf{p}}_N\} = \hat{\mathbf{P}} = \text{Norm}(F^{\text{cg}})_{[: N, :]}$$

here, $\text{Norm}()$ means L2 normalization.

#### Text-Commonsense Interaction Module

Different from the visual side that has a large number of moment proposals, there is only one text query for a VTG task. Considering different modalities with their specific contents...
and relation patterns, we design a text-commonsense interaction module to obtain commonsense-guided text representation. Concretely, we first calculate the attention weights $A$, which represents the pair-wise relations between text and concepts:

$$A = qW^Q(CW^K)^\top,$$

where $W^Q \in \mathbb{R}^{d^Q}$ and $W^K \in \mathbb{R}^{d^C}$ are learnable parameter matrices. Then, a softmax function is employed to obtain commonsense-guided query representation:

$$\text{ConAttn}(q, C, C) = \text{softmax}(\frac{A}{\sqrt{d_{avg}}})CW^V,$$

(9)

The obtained representation is then normalized as the final query representation, denoted as $\hat{q}$, to calculate matching scores in further common space learning:

$$\hat{q} = \text{Norm}(\text{ConAttn}(q, C, C)).$$

(10)

### 3.5 Complementary Common Space

The performance of [18] shows that replacing complex cross-modal interaction with common space learning has a significant effect on reducing test-time cost. Nonetheless, common space learning cannot explicitly establish the interaction process between two different modalities, leading to the features learned in common space hardly capturing the complex relations between these modalities.

To solve this problem, commonsense knowledge has been learned and fused with visual and text information, which can efficiently enhance the discriminative ability of different modality features. Nevertheless, the commonsense-guide information may lose the global information of the query sentence since commonsense consists of separate concepts, which leads to incomplete VTG learning. Therefore, in this section, we design a complementary common space to effectively compute matching scores while maintaining the original semantic information of the video and query. To reduce the number of parameters and the cost of time, we simply adopt two feature transformation modules, $\phi_1, \phi_2$ to project commonsense-guided visual features for both the original query features and commonsense-guided query features, respectively. Then the final matching scores can be calculated adaptively as follows:

$$m_i = \phi_1(\hat{p}_i)^\top q,$$

$$n_i = \phi_2(\hat{p}_i)^\top \hat{q},$$

$$a_i = \gamma m_i + (1 - \gamma) n_i,$$

(11)

where $\phi_1$ and $\phi_2$ are MLPs. For each proposal feature $\hat{p}_i$, $m_i$ represents the matching score between $\hat{p}_i$ and original query feature $q$, and $n_i$ represents the matching score between $\hat{p}_i$ and commonsense-guided query feature $\hat{q}$. Then, a learnable parameter, $\gamma$ is utilized to adaptively balance the weight of the two types of matching score. Finally, $a_i$ means the matching score of the $i$-th moment proposal for temporal localization.

### 3.6 Training and Inference

**Training:** With the above calculated matching scores, a binary cross entropy loss is used to optimize our model as follows:

$$L(y_i, a_i) = -\frac{1}{N} \sum_{i=1}^{N} y_i \log a_i + (1 - y_i) \log(1 - a_i),$$

(12)

where the soft label $y_i$ is generated by thresholding the overlap ratio between the $i$-th moment proposal and the groundtruth temporal segment.

**Inference:** During testing, we employ the matching scores $a$ of each proposal for video temporal grounding. The time cost of our proposed $C_2$AN method only consists of the calculation of text encoder, text-commonsense
interaction and matching scores since the moment proposal features can be pre-calculated and stored in a gallery database.

4 EXPERIMENTS

In this section, we evaluate our approach for video temporal grounding task on three public datasets: ActivityNet Caption [38], TACoS [55] and Charades-STA dataset [15]. Note that the bias in the annotations and the distributions of the target moments in the Charades-STA dataset provide trivial priors on moment locations according to [29, 50], which means that the performance of the model on the Charades-STA dataset does not have a strong reference value. As a result, we evaluate our model on Charades-STA dataset as an assisted experiment.

4.1 Datasets and Evaluation

TACoS. The TACoS dataset is widely used on VTG task, and it is collected by Regneri et al. [55], which consists of 127 videos from different cooking scenarios. The video's average duration is about 4.79 minutes. TACoS is a challenging dataset for VTG due to the multi-level activities contained by the query sentences in the dataset. Following the standard split [15], TACoS has 10146, 4586, and 4083 moment-query pairs for training, validation, and testing, respectively.

ActivityNet Captions. The ActivityNet Captions dataset is a popular benchmark dataset containing around 20K videos with 100K annotations. Currently, it is the largest dataset for video moment retrieval task. Following [79, 84], we use the first validation set for validation and the second validation set for testing. The total length of all videos is over 648 hours, and the videos are associated with more than 200 types of daily activities. The videos contain 3.65 moment-sentence pairs on average, and the average length of the descriptions is 13.48 words.

Charades-STA. The Charades-STA dataset consists of approximately 7K daily life videos, and most of them describe indoor activities. There are 12408 moment-sentence pairs for training and 3720 pairs for testing in the dataset. The average length of videos is 29.8s and the average duration of the moment is 8.2s. There are roughly 2.4 annotated moments per video.

Metrics. Following previous work [15, 79], the evaluation metrics "R@n, IoU=m" is utilized to evaluate the ability of our approach. "R@n, IoU=m" represents the percentage of at least one of the top-n predicted segments which have Intersection over Union (IoU) larger than m. Specifically, we set n∈{1, 5} and m∈{0.1, 0.3, 0.5, 0.7} for ActivityNet Caption dataset and TACoS dataset, and n∈{1, 5} and m∈{0.5, 0.7} for Charades-STA dataset.

4.2 Implementation Details

Feature Extractor. The same visual features (i.e., VGG [57], C3D [60] and I3D [4]) as previous approaches are employed to produce a fair and detailed comparison to reduce the influence of different visual encoders. We adopt VGG, C3D and I3D to make comparisons on TACoS, the latter two on ActivityNet Captions, and the former on Charades-STA.

Concept Relation Graph. For concept selection in the training set, we first collect all the annotations of a specific training dataset (the ActivityNet Captions dataset or the TACoS dataset) to construct a text set. From this text set, we statistic the frequencies of each word, and select those greater than or equal to 3 on TACoS, 5 on ActivityNet Captions and 4 on Charades-STA as our commonsense concepts. Following [63], we expand the set of commonsense concepts. Subsequently, we conduct a statistical computation to obtain the correlations of the concepts which belong to training set to generate the concept relation graph. For concepts not present in the training set, we calculate their cosine similarities with other concepts, and then make similarities as the weight in relation graph. Finally, we obtain 624 commonsense concepts on TACoS, 3152 on ActivityNet Captions and 392 on Charades-STA.
Table 1. Speed-accuracy analysis on two datasets. **TE**: time cost of query (Text) Embedding generation. **CML**: time cost of the Cross-Modal Learning for VMR. **ALL**: The total time cost of TE and CML. We report the accuracy (**ACC**) of R@1, IoU=0.5 and the sum of the accuracy (**sumACC**), the sum of the accuracy of IoU=0.3 and IoU=0.5 for comparison.

| Methods | TACos | ANetCap |
|-------|-------|--------|
|       | TE    | CML   | ALL | ACC | sumACC |
|       | TE    | CML   | ALL | ACC | sumACC |
| TMLGA | 1.14  | 11.37 | 12.51 | 21.65 | 46.19 |
| VSLNet| 3.58  | 5.02  | 8.59 | 24.27 | 53.88 |
| LGI   | -     | -     | -   | -   | -     |
| DRN   | 4.67  | 22.13 | 26.81 | 23.17 | -     |
| CTRL  | 4.32  | 534.23 | 538.55 | 13.30 | 31.62 |
| SCDM  | 3.65  | 780.0 | 783.65 | 21.17 | 47.28 |
| CBP   | 3.17  | 2659.01 | 2662.18 | 24.79 | 52.10 |
| 2D-TAN| 1.72  | 135.84 | 137.56 | 25.32 | 62.61 |
| FVMR  | 3.51  | 0.14  | 3.65 | 29.12 | 70.60 |
| Ours  | 2.33  | 0.29  | 2.62 | 32.83 | 78.13 |

**Architecture settings.** As for text encoding, we set the max length of word sequence to 30, and utilize the pre-trained 300-dim Glove embeddings [51] to initialize the words in the query and the consensus concepts. Besides, a two-layer bi-directional LSTM is adopted with 512 hidden state dimensions. We use a two-layer Graph Convolution Network (GCN) with 512 and 1024 embedding dimensions for concept feature extraction. The feature dimensions $d^V$, $d^Q$ and $d^C$ are all set to 512.

**Training and Inference settings.** We follow [84] to generate candidate moment proposals by adopting sliding windows to random select $N$ consecutive clips, which is structured as 2D feature map. The window size $N$ is set to 128 for TACoS, 64 for ActivityNet Captions and 16 for Charades-STA, respectively. Then, non-maximum suppression (NMS) is used on our predicted temporal segments. The NMS threshold is set to 0.49 for all experiments. Our proposed method is trained by an Adam optimizer [10] with a learning rate of 0.0001. Our model is trained for 50 epochs on TACoS dataset and 30 epochs on ActivityNet Captions dataset and Charades-STA dataset. The batch size is set to 64 for TACoS and 32 for ActivityNet Captions and Charades-STA. All our experiments are implemented in PyTorch toolkit with 2 NVIDIA Geforce RTX 3090 GPUs.

4.3 Comparison with State-of-the-art Methods

We compared our proposed C2AN approach with the following state-of-the-art baselines on three benchmark datasets:

- **Proposal-based Methods:** CTRL [15], MCN [1], MAN [81], SCDM [77], SAP [7], TGN [5], ACRN [44], QSPN [75], CMIN [85], FIAN [53], 2D-TAN [84], CMAS [76].

- **Proposal-free Methods:** ABRL [78], TMLGA [56], LGI [48], DRN [79], VSLNet [82], DEBUG [45], ExCL [25], CBP [64], CBLN [42], CPN [88], SSCS [11], MGSL-Net [41], GTR [2], SDN [37].

- **Reinforcement-learning-based Methods:** RWM-RL [30], SM-RL [68], TSP-RPL [70], TripNet [28].

In the following, the best performance is highlighted in **bold** and the second-best underlined.

**Overall Speed-Accuracy Analysis.** Considering that fast VTG task pays the same attention to the speed as the accuracy, following [18] we evaluate the time cost of Text Encoding (TE) for query embedding generation and Cross-Modal Learning for moment localization (CML). Table 1 shows the performance. Besides, we also calculate the sum of the accuracy in terms of "IoU=0.3" and "IoU=0.5", named sumACC to evaluate the whole performance.
Table 2. Comparison results on TACoS.

| Method   | R@1 IoU=0.1 | R@1 IoU=0.3 | R@1 IoU=0.5 | R@1 IoU=0.7 | sumACC | R@5 IoU=0.1 | R@5 IoU=0.3 | R@5 IoU=0.5 | R@5 IoU=0.7 | sumACC |
|----------|--------------|--------------|--------------|--------------|--------|--------------|--------------|--------------|--------------|--------|
|          |              |              |              |              |        |              |              |              |              |        |
| VGG features |              |              |              |              |        |              |              |              |              |        |
| MCN      | 14.42        | -            | 5.58         | -            | 37.35  | 10.33        | -            |            | -            |        |
| SM-RL    | 26.51        | 20.25        | 15.95        | -            | 36.20  | 50.01        | 38.47        | 27.84        | -            | 66.31  |
| SAP      | 31.15        | -            | 18.24        | -            | -      | 53.51        | -            | 28.11        | -            |        |
| Ours     | 54.55        | 43.25        | 30.52        | 17.02        | 73.77  | 78.38        | 65.97        | 52.78        | 31.88        | 118.75 |

| Method   |              |              |              |              |        |              |              |              |              |        |
| C3D features |              |              |              |              |        |              |              |              |              |        |
| TGN      | 41.87        | 21.77        | 18.90        | 11.88        | 40.67  | 53.40        | 39.06        | 31.02        | 15.26        | 70.08  |
| ACRN     | -            | 19.52        | 14.62        | -            | 34.14  | -            | 34.97        | 24.88        | -            | 59.85  |
| DEBUG    | 41.15        | 23.45        | -            | -            | -      | -            | -            | -            | -            |        |
| DRN      | -            | -            | 23.17        | -            | -      | -            | -            | 33.36        | -            |        |
| CTRL     | 24.32        | 18.32        | 13.30        | -            | 31.62  | 48.73        | 36.69        | 25.42        | -            | 62.11  |
| QSPN     | 25.31        | 20.15        | 15.23        | -            | 35.38  | 53.21        | 36.72        | 25.30        | -            | 62.02  |
| ACL      | 31.64        | 24.17        | 20.01        | -            | 44.18  | 57.85        | 42.15        | 30.66        | -            | 72.81  |
| SCDM     | -            | 26.11        | 21.17        | -            | 47.28  | -            | 40.16        | 32.18        | -            | 72.34  |
| CBP      | -            | 27.31        | 24.79        | -            | 52.10  | -            | 43.64        | 37.40        | -            | 81.04  |
| 2D-TAN   | 47.59        | 37.29        | 25.32        | -            | 62.61  | 70.31        | 57.81        | 45.04        | -            | 102.85 |
| FIAN     | 39.55        | 33.87        | 28.58        | -            | 62.45  | 56.14        | 47.36        | 39.16        | -            | 86.92  |
| CBLN     | 49.16        | 38.98        | 27.65        | -            | 66.63  | 73.12        | 59.96        | 46.24        | -            | 106.20 |
| CMN      | -            | 24.64        | 18.05        | -            | 42.69  | -            | 38.46        | 27.02        | -            | 65.48  |
| TripNet  | -            | 23.95        | 19.17        | 9.52         | 43.12  | -            | -            | -            | -            |        |
| ABLR     | 34.70        | 19.50        | 9.40         | -            | 28.90  | -            | -            | -            | -            |        |
| BPNet    | -            | 25.96        | 20.96        | -            | 46.92  | -            | -            | -            | -            |        |
| MGSN-Net | -            | 42.54        | 32.27        | -            | 74.81  | -            | 63.39        | 50.13        | -            | 113.52 |
| SSCS     | -            | 41.33        | 29.56        | -            | 70.89  | -            | 60.65        | 48.01        | -            | 108.66 |
| CMAS     | -            | -            | 31.37        | 16.85        | -      | -            | 51.42        | 27.68        | -            |        |
| GTR      | -            | 40.39        | 30.22        | 16.35        | 70.61  | -            | 61.94        | 47.73        | -            | 109.67 |
| FVMR     | 53.12        | 41.48        | 29.12        | 16.35        | 70.60  | 78.12        | 64.53        | 50.00        | 30.15        | 114.53 |
| Ours     | 54.38        | 43.30        | 31.08        | 18.05        | 74.38  | 77.68        | 64.03        | 52.78        | 32.83        | 116.47 |

| Method   |              |              |              |              |        |              |              |              |              |        |
| B3D features |              |              |              |              |        |              |              |              |              |        |
| ExCL     | -            | 45.50        | 28.00        | 13.80        | 73.50  | -            | -            | -            | -            |        |
| TMLGA    | -            | 24.54        | 21.65        | 16.46        | 46.19  | -            | -            | -            | -            |        |
| VSLNet   | -            | 29.61        | 24.27        | -            | 53.88  | -            | -            | -            | -            |        |
| Ours     | 54.38        | 43.30        | 31.08        | 18.05        | 74.38  | 77.68        | 64.03        | 52.78        | 32.83        | 116.47 |

of each model. Due to the learning of consensus knowledge and the simple but efficient dot production between different modal feature vectors in learned common space, our proposed method achieves great performance with both high speed and efficiency. Obviously, TE has no effect on the test-time, and each method spends a similar time (~3ms) on TE module because of the limited capability of text encoders like LSTM. However, a vast difference appears in the time cost of CML between different approaches. For the CML, we can find that our proposed method is at least fifty times faster than state-of-the-arts that do not learn common space. Compared with FVMR, with the similar time cost, our proposed method outperforms on "R@1, IoU=0.5" by gains of 3.71% on TACoS dataset and 2.90% on ActivityNet Captions dataset. According to sumACC, we can find that our proposed C3AN outperforms other state-of-the-art methods by gains of at least 7.53% on TACoS and 6.66% on ActivityNet Captions. It demonstrates that the learning of consensus knowledge can obtain more discriminative features
for better performance. Though these proposal-free approaches such as LGI, VSLNet, and TMLGA also achieve favor-able performance with low computational expenses, our proposed method still outperforms them by gains of 6.60%, 4.89%, 15.07% on ActivityNet Captions, respectively. Moreover, most proposal-free methods can only regress one temporal location for VTG, which is not suitable in practical applications. The above comparison illustrates that our method has significant speed and accuracy advantages. Although it can be pre-extracted, we also report the time cost of visual encoding and visual-commonsense interaction modules as follows: We choose C3D for visual encoding, and the time cost is approximately 54.37ms on average for Charades-STA, 513.18ms for TACoS, and 307.88ms for ActivityNet Captions. Moreover, the time cost of visual-commonsense interaction is 12.14ms for Charades-STA, 97.12ms for TACoS, and 88.54ms for ActivityNet Captions.

**Results on TACoS and ActivityNet Captions.** We compare the performance of our proposed method against extensive video temporal grounding models on two benchmark datasets. As shown in Table 2, we can observe that our method performs better than other methods in most metrics. On the TACoS dataset, our proposed C2AN achieves significant performance on all the three types of visual features. Compared with FVMR, though our method achieves (1.52%, 0.15%) lower than FVMR on metric “R@5, IoU=[0.1, 0.3]” on TACoS, it outperforms FVMR in all other metrics, especially on the metrics “R@1, IoU=[0.5, 0.7]” and “R@5, IoU=[0.5, 0.7]” by gains of (3.71%, 1.72%) and (2.68%, 2.95%) respectively. Note that IoU=0.7 is a more crucial criterion to determine whether a VTG model is accurate or not. The comparison of performance on ToU=0.7 shows that our method can predict results with higher quality. Moreover, as shown in Table 3 our proposed C2AN also surpasses FVMR on ActivityNet Captions by (2.90%, 2.22%) in terms of “R@1, IoU=[0.5, 0.7]”. ActivityNet Captions has larger scales than the TACoS dataset. The results indicate that our method also performs well in a more complex visual-text environment.

Then, we compare our model with much more VTG methods in more detail. Firstly, we compare C2AN with previous proposal-based methods: CTRL, MCN, MAN, ACL, TGN, QSPN, CMIN, CBP, SCDM, 2D-TAN and CMAS. From the results in Table 2 and Table 3, we observe that our C2AN achieves great performance compared with the aforementioned methods on most of the metrics. Part of previous work ignores the temporal context information as well as making inadequate cross-modal interaction. Meanwhile, our C2AN model replaces complex cross-modal interaction with commonsense knowledge, which makes the generated features more discriminative, and saves much more time cost. The experimental results demonstrate the effectiveness of C2AN in capturing rich cross-modal information and characterizing the complex associative patterns between different modalities.

Moreover, we compare our method with previous proposal-free methods: TGN, CMIN, CBP, SCDM, DRN, LGI, CBLN, SSCS, MGSL-Net, GTR and SDN. Due to the fine-grained interaction and the regression product, the mentioned methods have achieved great results in recent years. DRN mainly regresses the distances from each frame to the temporal boundaries, LGI combines local and global information, while MGSL-Net pays more attention to the off-balance data distribution. In addition, CMAS focuses on modeling the temporal dependencies between its generated proposals, while SSCS tends to make samples more discriminative in a contrastive manner. Compared with them, our proposed C2AN method achieves better performance. On ActivityNet Captions, we outperform LGI by gains of (2.06%, 4.68%, 5.80%) in terms of “R@1, IoU=[0.3, 0.5, 0.7]”, and outperform DRN by gains of (0.74%, 4.51%) in terms of “R@1, IoU=[0.5, 0.7]”. Besides, our method outperforms SDN by gains of (2.15%, 5.70%, 5.46%) on the metrics “R@1, IoU=[0.3, 0.5, 0.7]”, and reaches higher scores with an approximately 2.40% performance improvement than SSCS concerning “R@1” metric. We also surpass CMAS by gains of (1.88%, 0.06%, 2.28%, 4.57%) on the metrics “R@1[1,5], IoU=[0.5,0.7]”. On TACoS, compared with MGSL-Net, our method outperforms it by gains of (2.76%, 0.56%, 0.99%, 2.55%) on the metrics “R@1[1,5], IoU=[0.3, 0.5]”. Besides, it also surpasses the recent work CBLN with an average 6.11% improvement on the metrics “R@1[1,5], IoU=[0.1, 0.3, 0.5]”. Moreover, our C2AN outperforms CMAS by gains of (1.46%, 1.22%, 1.26%, 5.42%) in terms of “R@1[1,5], IoU=[0.5, 0.7]”, and outperforms SSCS with (3.97%, 3.27%, 3.73%, 4.67%) improvements on the “R@1[1,5], IoU=[0.3, 0.5]” metrics. Note that other state-of-the-art approaches hardly show their results on the metric “IoU=0.7” on TACoS,
Table 3. Comparison results on ActivityNet Captions.

| Method  | R@1 | IoU=0.3 | IoU=0.5 | IoU=0.7 | IoU=0.3 | IoU=0.5 | IoU=0.7 |
|---------|-----|---------|---------|---------|---------|---------|---------|
| C3D features |     |         |         |         |         |         |         |
| MCN     | 39.35 | 21.36  | 6.43   | 68.12  | 53.23  | 29.70  |
| TGN     | 47.43 | 29.01  | 10.34  | 75.32  | 59.17  | 37.54  |
| ACRN    | 49.70 | 31.67  | 11.25  | 76.50  | 60.34  | 38.57  |
| DEBUG   | 55.91 | 39.72  | -      | -      | -      | -      |
| GDP     | 56.17 | 39.27  | -      | -      | -      | -      |
| ABRL    | 55.67 | 36.79  | -      | -      | -      | -      |
| TripNet | 48.42 | 32.19  | 13.93  | -      | -      | -      |
| TSP-PRL | 56.08 | 38.76  | -      | -      | -      | -      |
| I3D features |     |         |         |         |         |         |         |
| ExCL    | 62.30 | 42.70  | 24.10  | -      | -      | -      |
| TMLGA   | 51.28 | 33.04  | 19.26  | -      | -      | -      |
| VSLNet  | 63.16 | 43.22  | 26.16  | -      | -      | -      |

and “IoU=0.7” is a stricter standard to define whether a localized moment is correct. It indicates that our method can localize the moment with higher quality. Our C2AN obtains more accurate results because the extracted commonsense knowledge bridges different modalities and makes the model have the ability to comprehend complex cross-modal relationships. In fact, the simple but effective use of commonsense knowledge replaces large and repetitive cross-modal interaction and reduces the time cost of capturing fine-grained information. It validates that C2AN can efficiently and effectively localize the target moment boundary.

**Results on Charades-STA.** We also compare the performance of our proposed C2AN with SOTA approaches on the Charades-STA dataset with VGG features. As shown in Table 4, we can find that our model achieves better performance than other approaches on most metrics. To be specific, our model outperforms FVMR by gains of (4.20%, 3.71%, 0.73%, 8.56%) in terms of “R@{1,5}, IoU=[0.5, 0.7]”. Compared with SSCS and CPN, our method surpasses them by gains of (3.41%, 2.31%) and (0.48%, 2.79%) in terms of “R@1, IoU=[0.5, 0.7]”. Though DRN and CBLN achieve better performance on “R@5, IoU=0.5” metric, we outperform them with the gains of (3.66%, 4.17%) and (2.89%, 3.41%) on the metrics “R@1, IoU=[0.5, 0.7]”, respectively. Moreover, our approach...
Table 4. Comparison results on Charades-STA.

| Method | R@1       | R@5       | IoU=0.5 | IoU=0.7 | IoU=0.5 | IoU=0.7 |
|--------|-----------|-----------|---------|---------|---------|---------|
| VGG features |           |           |         |         |         |         |
| MCN    | 17.46     | 8.01      | 48.22   | 26.73   |         |         |
| SAP    | 27.42     | 13.36     | 66.37   | 38.15   |         |         |
| MAN    | 41.21     | 20.54     | 83.21   | 51.85   |         |         |
| DRN    | 42.90     | 23.68     | 87.80   | 54.87   |         |         |
| 2D-TAN | 39.70     | 23.31     | 80.32   | 51.26   |         |         |
| CBLN   | 43.67     | 24.44     | 88.39   | 56.49   |         |         |
| CPN    | 46.08     | 25.06     | -       | -       |         |         |
| SSCS   | 43.15     | 25.54     | 84.26   | 54.17   |         |         |
| FVMR   | 42.36     | 24.14     | 83.97   | 50.15   |         |         |
| SDN    | 44.60     | 27.10     | -       | -       |         |         |
| Ours   | 46.56     | 27.85     | 84.70   | 58.71   |         |         |

Table 5. Ablation Studies on TACoS and ActivityNet Captions.

| Methods        | TACoS      | ANetCap    |
|----------------|------------|------------|
|                | R@1 IoU=0.1 | R@1 IoU=0.3 | R@1 IoU=0.5 | R@1 IoU=0.7 | R@5 IoU=0.1 | R@5 IoU=0.3 | R@5 IoU=0.5 | R@5 IoU=0.7 |
| backbone       | 50.42 26.55 | 77.47 46.33 | 60.08 24.53 | 86.51 61.92 | 57.00 32.83 | 75.60 44.38 | 65.15 32.83 | 87.06 64.82 |
| Ours(w/o. v-c) | 51.82 28.88 | 76.95 50.52 | 63.20 35.35 | 86.28 63.77 | 57.00 32.83 | 75.60 44.38 | 65.15 32.83 | 87.06 64.82 |
| Ours(w/o. t-c) | 53.45 30.30 | 76.42 50.05 | 63.99 38.72 | 86.45 64.37 | - - - - | - - - - | - - - - | - - - - |
| Ours(w/o. c-c) | 52.28 30.30 | 77.53 50.53 | 64.39 41.11 | 86.29 64.63 | 57.00 32.83 | 75.60 44.38 | 65.15 32.83 | 87.06 64.82 |
| Ours(full)     | 56.00 32.83 | 76.60 52.60 | 65.15 41.11 | 87.06 79.32 | 57.00 32.83 | 75.60 44.38 | 65.15 32.83 | 87.06 64.82 |

achieves (1.96%, 0.75%) improvements on "R@1, IoU={0.5, 0.7}" metrics when compared with SDN. The promising results in terms of "R@1" indicate that our framework can predict results with higher quality. Obviously, to some extent, the use of simple and effective commonsense knowledge in our model can achieve the comparable performance against the methods leveraging large and complex cross-modal interactions.

4.4 Ablation Study
In this section, we take in-depth ablation studies to investigate the contribution of each main component in our proposed C₂AN method on TACoS and ActivityNet Captions datasets. Specifically, to perform complete ablation studies, we divide the consensus-aware interaction module into two attention-based modules: the visual-commonsense interaction module and the text-commonsense interaction module. We train C₂AN with the following configurations:

- **backbone**: To prove the effect of each component, we directly map extracted visual and text features into common space to calculate similarity scores without any other modules.
- **w/o. v-c**: We remove the visual-commonsense interaction module to verify whether commonsense knowledge can strengthen visual information.
- **w/o. t-c**: We replace the text-commonsense interaction module with directly using encoded text features to further calculate matching scores.
- **w/o. c-c**: To investigate the complementary role of original semantic information to the performance of C₂AN, only one common space is used for obtaining matching scores (only \( n_i \) is used).
The "full" means the full C2AN model. Table 5 summarizes the localization results in terms of "R@1, IoU={0.1, 0.3, 0.5, 0.7}" for TACoS and "R@1, IoU={0.3, 0.5, 0.7}" for ActivityNet Captions. For the improvement of each branch, we have concrete analysis as follows:

**Effects of consensus-aware interaction module.** We evaluate the effectiveness of the commonsense-aware interaction module by training our model only using no commonsense-guided visual or text features. From the results in Table 5, we can observe that, compared with backbone, no matter which type of modality features are fused with concept information can improve the performance of the model. Obviously, each interaction module has a positive effect on the VTG task. On TACoS, the full model outperforms "w/o. t-c" by gains of (2.55%, 3.90%, 2.53%) on metrics "R@1, IoU={0.1, 0.3, 0.5}" and outperforms "w/o. v-c" by gains of (4.18%, 4.75%, 3.95%) on the same metrics. For the ActivityNet Captions, the full model exceeds "w/o. v-c" by (1.95%, 2.76%, 3.02%) in terms of "R@1, IoU={0.3, 0.5, 0.7}" while achieves an average 1.07% improvement in terms of "R@5, IoU=0.7". Besides, the full model also outperforms "w/o. t-c" by a large margin on all metrics. Obviously, the interaction between visual and commonsense knowledge leads to a improvement in performance, which proves that the exploitation of commonsense knowledge is significant to complementing visual information. Moreover, we can find that the text features are more effective when reinforced by commonsense knowledge. That is, commonsense knowledge can cooperate with visual and text features to obtain favorable results.

**Effect of the complementary common space.** To obtain discriminative representations, a complementary common space is designed to effectively compute matching scores while maintaining the original semantic information of the video and query. Here, a baseline "w/o. c-c" is designed to study the influence of the original information. From Table 5, the full model achieves a (3.72%, 2.92%, 2.65%) improvement compared with "w/o. c-c" in terms of "R@1, IoU={0.1, 0.3, 0.5}" on TACoS and achieves a (0.76%, 1.00%, 1.84%) in terms of "R@1, IoU={0.3, 0.5, 0.7}" on ActivityNet Captions. It proves that the original query can be a type of supplement for the consensus concept due to the fact that concepts are only words selected from the query.

**Effect of common space dimension.** We investigate the influence of different dimensions of the learned common space on TACoS. As shown in Figure 3(a), a too large dimension would lead to a higher cost of memory and times, which would also reduce the performance of our model. By contrast, a too small dimension always results in the lack of representation capability of learned common space. Therefore, a modest value of the common space dimension gets better performance.

**How many consensus concepts should be selected?** In section III, we introduce the extraction of the commonsense concept by selecting the words with frequencies larger than $O_{con}$. To explore the influence of the number of concepts, we validate the model performance with different $O_{con}$ on TACoS. As shown in Figure 3(b), it is obvious that with the addition of the threshold value, the performance of our model does not always improve. As a result, we set the threshold to 3.
Table 6. Comparison results on TACoS dataset and ActivityNet Captions dataset with different knowledge.

| Method     | TACoS   | ActivityNet Captions |
|------------|---------|-----------------------|
|            | R@1 IoU=0.3 IoU=0.5 IoU=0.7 | R@1 IoU=0.3 IoU=0.5 IoU=0.7 |
| C3D features | 38.23 26.80 15.42 | 62.60 50.15 32.05 |
| Ours_COCO  | 38.23 26.80 15.42 | 62.60 50.15 32.05 |
| Ours_F30K  | 37.80 25.78 14.72 | 62.10 48.33 29.49 |
| Ours       | 45.30 32.83 18.07 | 65.15 48.11 29.54 |

Different Commonsense Sources. A large-scale external language corpus is indeed important to get rich commonsense knowledge. But the knowledge extracted from the specific datasets is more consistent with the semantic environment, while the large-scale low-related corpus would contain more knowledge unrelated to the current dataset, which also generates noise in the training process. To investigate the reasonability and suitability of utilizing the commonsense knowledge from specific datasets, we replace our extracted commonsense knowledge with the knowledge obtained from the MSCOCO dataset and Flickr30K dataset utilized in CVSE. The results of the comparisons are as follows:

In Tab 6, Ours_COCO and Ours_F30K represent that the model utilizes the commonsense knowledge from the MSCOCO dataset and the Flickr30k dataset, respectively. We can find that our approach, leveraging the commonsense knowledge from the specific datasets, outperforms Ours_COCO and Ours_F30K on all metrics. The results in Tab 6 indicate that the knowledge consistent with the semantic environment is more suitable for learning cross-modal information than that extracted from a low-related large-scale language corpus.

In fact, the COCO dataset and Flickr30k dataset are both domain-specific datasets, with the former focusing on entry-level categories [40] while the latter focusing mainly on people and animals [52]. This indicates that there is more related commonsense knowledge in the TACoS and ActivityNet Captions datasets that does not exist in the COCO and Flickr30k datasets. Therefore, the knowledge extracted from the two datasets does not work well with the model.

4.5 Qualitative Results

To qualitatively validate the efficiency of our proposed C2AN method, several typical examples are shown in Figure 4 for qualitative comparison. Besides the C2AN method, three baselines, i.e., C2AN (w/o. v-c), C2AN (w/o. t-c), and C2AN (w/o. c-c), are also validated for deep insight into the impact of different main components on localization performance. In Figure 4, examples 1, 3, and 5 are from TACoS, while examples 2, 4, and 6 are from ActivityNet Captions. In examples 1-4, C2AN (full) performs the results almost the same as ground truth, while the other three baselines output worse results than the full model. The performance in complex environments validates that our proposed C2AN can accurately capture and comprehend the rich cross-modal relationship to a certain extent. Obviously, the ablation models predict the boundaries with a larger distance to ground truth compared with the full model. Since they either lack the crucial commonsense knowledge as a bridge between visual and text modalities or do not have a limit from original semantics. The result shows that the learning of commonsense knowledge plays an essential role in enhancing the discriminative ability of different modalities, and the original semantics are essential for better performance.

While our method can localize the correct temporal segment for most videos, failed samples also exist in some cases. As shown in Figure 4, examples 5 and 6 present the cases of failing to localize the correct video moments. In example 5, C2AN wants to localize the video segment that describes “The onion is cut partway through”. As a whole, the query has too less semantic information for retrieval, which leads to an incorrect result. Moreover, “partway” is not a commonsense concept of TACoS. That is, C2AN only pays more attention to the
In this paper, we propose a novel Commonsense-aware Cross-modal Alignment Network (C2AN) to achieve fast video temporal grounding, which incorporates commonsense-guided visual and text representations into a complementary common space and learns efficient commonsense-aware cross-modal alignment. We first utilize a multi-modal feature extractor to obtain different modality features. Specially, we select a set of representative words as commonsense concepts according to their occurrence frequencies. After that, a commonsense-aware interaction module is designed to obtain discriminative visual and text representations. Finally, we leverage a complementary common space to align visual and text representations to calculate their matching scores for temporal grounding. Experimental results of our proposed C2AN method demonstrate that it achieves competitive performance when compared with other state-of-the-art methods on two benchmark datasets.

Three perspectives will be considered for future work. First, considering the common disadvantage of proposal-based methods, it is necessary to regress the boundaries of the predicted results, which could improve the accuracy of the prediction. Second, We believe that it is very challenging to localize such complex moments without richer commonsense knowledge. To improve the efficiency of knowledge utilization and the generalization of our model, knowledge graph, external commonsense database, and the pre-training process could also be taken into account.
consideration. Finally, more sufficient and appropriate interactions between commonsense knowledge and visual or text features need to be explored such as modulated attention mechanism and casual reasoning.

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 ICCV. 5803–5812.
[2] Meng Cao, Long Chen, Mike Zheng Shou, Can Zhang, and Yuexuan Zou. 2021. On pursuit of designing multi-modal transformer for video grounding. arXiv preprint arXiv:2109.06085 (2021).
[3] Shuaiqiang Cao, Bairui Wang, Wei Zhang, and Lin Ma. 2022. Visual consensus modeling for video-text retrieval. In AAAI, Vol. 36. 167–175.
[4] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In CVPR. 6299–6308.
[5] Junyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. 2018. Temporally grounding natural sentence in video. In EMNLP. 162–171.
[6] Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, and Xiaolin Li. 2020. Rethinking the Bottom-Up Framework for Query-Based Video Localization. In AAAI. 10551–10558.
[7] Shaoxiang Chen and Yu-Gang Jiang. 2019. Semantic proposal for activity localization in videos via sentence query. In AAAI, Vol. 33. 8199–8206.
[8] Shuzhe Chen, Yida Zhao, Qin Jin, and Qi Wu. 2020. Fine-grained Video-Text Retrieval with Hierarchical Graph Reasoning. In CVPR. 10638–10647.
[9] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014).
[10] Kingma Da. 2014. A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).
[11] Xinpeng Ding, Nannan Wang, Shwei Zhang, De Cheng, Xiaomeng Li, Ziyuan Huang, Mingqian Tang, and Xinbo Gao. 2021. Support-set based cross-supervision for video grounding. In ICCV. 11573–11582.
[12] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xin 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).
[13] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lipan Wang, and Zicheng Liu. 2021. Violet: End-to-end video-language transformers with masked visual-token modeling. arXiv preprint arXiv:2111.12681 (2021).
[14] Junyu Gao, Mengyuan Chen, and Changsheng Xu. 2023. Collecting Cross-Modal Presence-Absence Evidence for Weakly-Supervised Audio-Visual Event Perception. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 18827–18836.
[15] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language grounding. In ICCV. 5267–5275.
[16] Jialin Gao, Xin Sun, Mengmeng Xu, Xi Zhou, and Bernard Ghanem. 2021. Relation-aware video reading comprehension for temporal language grounding. arXiv preprint arXiv:2110.05717 (2021).
[17] Junyu Gao and Changsheng Xu. 2020. CI-GNN: Building a category-instance graph for zero-shot video classification. IEEE Transactions on Multimedia 22, 12 (2020), 3088–3100.
[18] Junyu Gao and Changsheng Xu. 2021. Fast video moment retrieval. In ICCV. 1523–1532.
[19] Junyu Gao and Changsheng Xu. 2021. Learning video moment retrieval without a single annotated video. IEEE Transactions on Circuits and Systems for Video Technology (2021).
[20] Junyu Gao, Xiaoshan Yang, Yingying Zhang, and Changsheng Xu. 2020. Unsupervised video summarization via relation-aware assignment learning. IEEE Transactions on Multimedia 23 (2020), 3203–3214.
[21] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. 2018. Watch, think and attend: End-to-end video classification via dynamic knowledge evolution modeling. In ACM MM.
[22] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. 2019. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In Proceedings of the AAAI conference on artificial intelligence, Vol. 33. 8303–8311.
[23] Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. 2019. Mac: Mining activity concepts for language-based temporal localization. In WACV. IEEE. 245–253.
[24] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiahou Oie, and Ping Luo. 2022. Bridging video-text retrieval with multiple choice questions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 16167–16176.
[25] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann. 2019. ExCL: Extractive Clip Localization Using Natural Language Descriptions. In NACCL-HLT, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, 1984–1990. https://www.aclweb.org/anthology/N19-1198
[26] Juxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. 2019. Scene graph generation with external knowledge and image reconstruction. In CVPR. 1969–1978.
Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. 2019. Read, watch, and move: Reinforcement learning for video grounding. In ECCV.

Jiehang Hao, Haifeng Sun, Pengfei Ren, Jingyu Wang, Qi Qi, and Jianxin Liao. 2022. Can Shuling Video Beneit Temporal Bias Problem: A Novel Training Framework for Temporal Grounding. In MM.

Jonghwan Mun, Minsu Cho, and Bohyung Han. 2020. Local-Global Video-Text Interactions for Temporal Grounding. In AAAI.

Ke Ning, Ming Cai, Di Xie, and Fei Wu. 2019. Exploring Rich Semantics for Open-set Action Recognition. In IJCV.

Jun Hu, Bryan Hooi, and Bingsheng He. 2023. Efficient Heterogeneous Graph Learning via Random Projection. arXiv:2310.14481 [cs.LG].

Jun Hu, Shengsheng Qian, Quan Fang, Youze Wang, Quan Zhao, Huaixwen Zhang, and Changsheng Xu. 2021. Efficient Graph Deep Learning in TensorFlow with tf_geometric. In MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021. ACM, 3775–3778.

Yufan Hu, Junyu Gao, Jianfeng Dong, Bin Fan, and Hongmin Liu. 2024. Exploring Rich Semantics for Open-set Action Recognition. IEEE Transactions on Multimedia 26 (2024), 5410 – 5421.

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).

Zhonghao Wang, Xiaoye Qu, and Jianxin Liao. 2020. Memory-Guided Semantic Learning Network for Temporal Sentence Grounding. In AAAI.

Ziyue Wu, Junyu Gao, Shucheng Huang, Changsheng Xu, Ziyue Wu, Junyu Gao, Shucheng Huang, and Changsheng Xu. 2021. The Pursuit of Knowledge: Discovering and Localizing Novel Categories using Dual Memory. In ICCV. 9153–9163.

ACM Trans. Multimedia Comput. Commun. Appl.
[55] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. Transactions of the Association for Computational Linguistics 1 (2013), 25–36.

[56] Cristian Rodriguez, Edison Marrese-Taylor, Fatemeh Sadat Saleh, Hongdong Li, and Stephen Gould. 2020. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In WACV. 2464–2473.

[57] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. ICLR (2015).

[58] Haoyu Tang, Jihua Zhu, Meng Liu, Wan Gao, and Zhiyong Cheng. 2021. Frame-wise Cross-modal Matching for Video Moment Retrieval. IEEE Transactions on Multimedia (2021).

[59] Junya Teng, Xiankai Lu, Yongshun Gong, Xinrui Liu, Xiushan Nie, and Yilong Yin. 2021. Regularized two granularity loss function for weakly supervised video moment retrieval. IEEE Transactions on Multimedia 24 (2021), 1141–1151.

[60] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In ICCV. 4499–4507.

[61] Gongmian Wang, Xing Xu, Fumin Shen, Huimin Lu, Yanli J, and Heng Tao Shen. 2022. Cross-modal Dynamic Networks for Video Moment Retrieval with Text Query. IEEE Transactions on Multimedia (2022).

[62] Hao Wang, Zheng-Jun Zha, Liang Li, Dong Liu, and Jiebo Luo. 2021. Structured multi-level interaction network for video moment localization via language query. In CVPR. 7026–7035.

[63] Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. 2020. Consensus-aware visual-semantic embedding for image-text matching. In ECCV. Springer, 18–34.

[64] Jingwen Wang, Lin Ma, and Wenhao Jiang. 2020. Temporally grounding language queries in videos by contextual boundary-aware prediction. In AAAI, Vol. 34. 12168–12175.

[65] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. IEEE Transactions on Pattern Analysis and Machine Intelligence 40, 10 (2017), 2413–2427.

[66] Wei Wang, Junyu Gao, Xiaoshan Yang, and Changsheng Xu. 2020. Learning coarse-to-fine graph neural networks for video-text retrieval. IEEE Transactions on Multimedia 23 (2020), 2386–2397.

[67] Wei Wang, Junyu Gao, Xiaoshan Yang, and Changsheng Xu. 2022. Many hands make light work: Transferring knowledge from auxiliary tasks for video-text retrieval. IEEE Transactions on Multimedia (2022).

[68] Weiming Wang, Yan Huang, and Liang Wang. 2019. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In CVPR. 334–343.

[69] Yuechen Wang, Jiajun Deng, Wengang Zhou, and Houqiang Li. 2021. Weakly supervised temporal adjacent network for language grounding. IEEE Transactions on Multimedia (2021).

[70] Jie Wu, Guanbin Li, Si Liu, and Liang Lin. 2020. Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In AAAI, Vol. 34. 12386–12393.

[71] Ziyue Wu, Junyu Gao, Shucheng Huang, and Changsheng Xu. 2021. Diving into the relations: Leveraging semantic and visual structures for video moment retrieval. In 2021 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 1–6.

[72] Ziyue Wu, Junyu Gao, and Changsheng Xu. 2022. Weakly-supervised Video Scene Graph Generation via Unbiased Cross-modal Learning. In Proceedings of the 31st ACM International Conference on Multimedia. 4574–4583.

[73] Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. 2021. Boundary proposal network for two-stage natural language video localization. In AAAI, Vol. 35. 2986–2994.

[74] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. arXiv preprint arXiv:2109.14084 (2021).

[75] Hujian Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. 2019. Multilevel language and vision integration for text-to-caption retrieval. In AAAI, Vol. 33. 9062–9069.

[76] Xun Yang, Shanshan Wang, Jian Dong, Jianfeng Dong, Meng Wang, and Tat-Seng Chua. 2022. Video moment retrieval with text-to-clip retrieval. In AAAI, Vol. 33. 9062–9069.

[77] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. 2019. Semantic Conditioned Dynamic Modulation for Temporal Sentence Grounding in Videos. In NeurIPS. 534–544.

[78] Yitian Yuan, Tao Mei, and Wenwu Zhu. 2019. To find where you talk: Temporal sentence localization in video with attention based location regression. In AAAI, Vol. 33. 9159–9166.

[79] Runhao Zeng, Haoming Xu, Wenhong Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. 2020. Dense regression network for video grounding. In CVPR. 10287–10296.

[80] Yawen Zeng, Da Cao, Xiaochi Wei, Meng Liu, Zhou Zhao, and Zheng Qin. 2021. Multi-modal relational graph for cross-modal video moment retrieval. In CVPR. 2215–2224.

[81] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. 2019. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In CVPR. 1247–1257.

[82] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2020. Span-based Localizing Network for Natural Language Video Localization. In ACL.
[83] Mingxing Zhang, Yang Yang, Xinghan Chen, Yanli Ji, Xing Xu, Jingjing Li, and Heng Tao Shen. 2021. Multi-stage aggregated transformer network for temporal language localization in videos. In *CVPR*. 12669–12678.

[84] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, Vol. 34. 12870–12877.

[85] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. 2019. Cross-modal interaction networks for query-based moment retrieval in videos. In *SIGIR*. 655–664.

[86] Zijian Zhang, Zhou Zhao, Zhu Zhang, Zhijie Lin, Qi Wang, and Richang Hong. 2020. Temporal textual localization in video via adversarial bi-directional interaction networks. *IEEE Transactions on Multimedia* 23 (2020), 3306–3317.

[87] Wenqing Zhao and Lijiao Xu. 2024. Weakly supervised target detection based on spatial attention. *Visual Intelligence* 2, 1 (2024), 1–11.

[88] Yang Zhao, Zhou Zhao, Zhu Zhang, and Zhijie Lin. 2021. Cascaded prediction network via segment tree for temporal video grounding. In *CVPR*. 4197–4206.

[89] Wenbo Zheng, Lan Yan, Fei-Yue Wang, and Chao Gou. 2021. Progressive Knowledge-Embedded Unified Perceptual Parsing for Scene Understanding. In *CVPR*. 1633–1642.

Received 18 May 2023; revised 29 January 2024; accepted 17 April 2024

ACM Trans. Multimedia Comput. Commun. Appl.