Neurocomputing 471 (2022) 194–207

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# STCM-Net: A symmetrical one-stage network for temporal language localization in videos

Zixi Jia *, Minglin Dong, Jingyu Ru *, Lele Xue, Sikai Yang, Chunbo Li

*Faculty of Robot Science and Engineering, Northeastern University, Shenyang 110819, China*

ARTICLE INFO

*Article history:*
Received 26 March 2021
Revised 22 October 2021
Accepted 7 November 2021
Available online 16 November 2021

*Keywords:*
Computer vision
Video segment localization
Sentence semantic mining
Video understanding

ABSTRACT

The task of temporal language localization in the video is to locate a video segment through natural language description for an untrimmed video. Compared with the general video localization task, it is more flexible and complex, which can accurately locate various scenes described by any natural language without making video labels in advance. It can be widely used for the field such as video retrieval and robot intelligent cognition. The main challenges of this task are the extraction of sentence semantics and the integration of contextual information in videos. Among them, contextual video integration can be optimized through the two-dimensional temporal adjacent network. Therefore, complete extraction of the potential information in the query sentence is necessary to solve the task more granularly. At the same time, we found a large amount of time-related information in the query sentence, which helps improve the localization accuracy. Thus, in this paper, we first define the time concept in a sentence and then propose a Sentence Time Concept Mining Network (STCM-Net), an symmetrical one-stage network. Can effectively extract the time concept contained in the query sentence, it can optimize the process of target localization and improve the localization performance. We also evaluate the proposed STCM-Net on three challenging public benchmarks: Charades-STA, ActivityNet Captions, and TACoS. Our STCM-Net gets encouraging improvements compared with the state-of-the-art approaches.



## 1. Introduction

As shown in Fig. 1, temporal language localization in the video is a foundational task of video understanding, which is significant for massive video retrieval and question answering.

In recent years, feature extraction mainly sentences semantics exploration, has been a hot research issue in this task. Many scholars have done relevant work in this direction, such as (R. Ge et al.2019) [1] proposed to match the "language activity concept" and "video activity concept" to improve the accuracy of localization. (M. Liu et al.2018) [2] took the sentence twice into the network to extract more semantic information and improve the accuracy of the final localization.

However, the mentioned methods rely mainly on increasing the network depth or multiple inputs to improve the effect. It did not extract the detailed semantic information in the sentence and did not consider visual information during sentence extraction.

Then, many researchers began to explore some particular information in sentences. (Dong. J et al. 2021)[3]proposed a simple and effective dual encoding method with hyperspace learning to make the excellent interpretability of the concept space. (X. Qu et al. 2020)[4] extract some semantics in the sentences through the bi-directional attention guidance network, but the lack of targeted mining limits their effectiveness.

At the same time, we find in the public datasets ActivityNet Captions and TACoS, more than 60% of the sentences contain obvious time-related information. (The experiment details show in Section 2). Precisely, some particular words and expressions in the query sentence can imply the period of the described segment. If the general information of the entire video is known in advance, the period of the target segment can be obtained very accurately. Therefore, we define the time concept in sentences, which refers to a characteristic that some particular words and phrases in the query sentence combined with the specific scene and video information, which can vary imply the period of the described segment. The experiments show that mining the time concept can effectively make up for the shortcomings of insufficient use of sentence information.

* Corresponding authors.
*E-mail addresses:* jiazixi@mail.neu.edu.cn (Z. Jia), dongminglin1234@163.com (M. Dong), rujingyu@mail.neu.edu.cn (J. Ru).

https://doi.org/10.1016/j.neucom.2021.11.019
0925-2312/

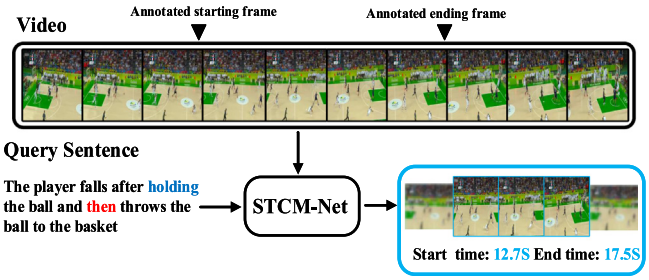



**Fig. 1.** An illustrative example of the temporal language localization task. Given a video and a query sentence, it aims to identify the starting and end time of the video segment described by the query.

However, mining the time concept is not a simple task, and two critical problems need to resolve. The one is how to quantify the vague concept of time. The mining of time concept is not equal to the simple analysis of representative words in a sentence. The other is that the time concept is relative, which means the same word corresponds to the different periods in different videos. Furthermore thus, the mining of time concept must be fully integrated with visual information.

Therefore, we propose the STCM-Net that can extract the time concept in the query sentence. With the fusion of visual information through a time mining network module, it can extract the keywords in the sentence in a targeted manner and convert the period results into a two-dimensional time probability map to optimize the results.

The contributions of this paper are as follows:

1. We study and define the character of the time concept in the sentence and a symmetrical one-stage network STCM-Net is proposed in Section 3. Compared with the previous work, it can extract the time-related information in the sentence more pertinently and make up for the defect of insufficient use of the sentence in the previous work.
2. We design a visual-guided attention mechanism to solve the problem of the quantification of time concept, which can fully pay attention to the time-related keywords with the combination of visual information.
3. We conduct experiments on three public datasets: ActivityNet Captions, TACoS, and Charades-STA, and our STCM-Net significantly outperforms state-of-the-art by a large margin.

## 2. Motivation

### 2.1. The details in the sentence

In recent years, significant progress has been achieved in feature extraction and sentence semantic extraction. The previous work focused mainly on understanding the entire event described in the sentence. However, it lacks attention to the details in the sentence, such as the description of the action, the dress, the tone, and the time-related information, especially some time-related description can be very effective in helping to locate the segment.

### 2.2. Our discovering

In the query sentence, many representative words can indicate the period of the corresponding segment. As shown in Fig. 2, phrases such as "turn off/on the light" indicate instantaneous action, and the corresponding period is relatively short in most cases, whereas "again/several more times/standing up/reading a book" should have a relatively long period corresponding to the description. Given the video and sentence, it can make full use of the potential information to predict the duration of the relevant segment.

### 2.3. Datasets analysis

After analyzing the three public datasets ActivityNet Captions, TACoS, and Charades-STA, respectively. We calculate the average duration of all videos in each dataset. Those with a duration longer than 70% of the average duration are marked as long videos, and less than 30% of the average duration is marked as short videos.

We count the proportions of different durations and the typical words. The results are shown in Table 1, and it can be found that for the ActivityNet and TACoS, the sum of long and short videos exceeds 70%. It indicates that the period of the target segment is not distributed evenly but prefers shorter or longer, and the duration and some typical words have an apparent relationship. In the Charades-STA, the duration is relatively short and fixed, and we believe that this is mainly because of the single scene of the dataset.

Many typical words (excluding prepositions and conjunctions, such as "person/the/a/and") can reflect the duration of the target segment. When the video and sentence are known, logically, humans can roughly predict the duration of the target segment. Due to the ActivityNet Captions has an enormous amount of data, so we choose ActivityNet as an example.

After removing the prepositions and conjunctions, we sort the top-10 most frequently occurring words in long and short videos, as shown in Fig. 3. We can see that the typical words in long videos are more words that can represent time duration, such as the "then/while/around," and words that describe instantaneous actions such as "seen/walks/appears" appear more in short videos.

In more detail, as shown in Fig. 4, the number of the word "see" and "then" of the training set, validation set, and test set in the ActivityNet Captions are listed in detail. It also proves that a correlation between typical words and the duration of the segment exists.

At the same time, to more clearly show the proportion of different words in different kinds of videos, we calculated the data in the following Table 2, which contains the number of different videos. It can be seen that the proportion of "then" in long videos is more than that of "see," but the opposite is true in short videos. Not only the proportion of typical words in each video can explain the problem, the change in the proportion of the same word in long and short videos can also further support our hypothesis.

Meanwhile, the main reason for using the ActivityNet dataset as a statistical sample is that it contains many instances and has a certain degree of representativeness. However, our model below does not improve the localization result due to the relatively complex scene described on the ActivityNet dataset. To further support our hypothesis, we additionally count the distribution of two typical words on the TACoS dataset in Table 3.

### 2.4. Challenge

The above analysis shows that there are some special pieces of information in the sentence, primarily the time-related information that can be used to locate the segment. But the exploration of the time concept is not easy, and the main challenges are as follows:

1. The mining of the concept of time is not a simple selection and analysis of typical words, but in-depth mining of various time-related information in the sentence, such as the implicit information about the occurrence of events, the impact on the time dimension. How to effectively capture this information and ingeniously transform the mining results is crucial.

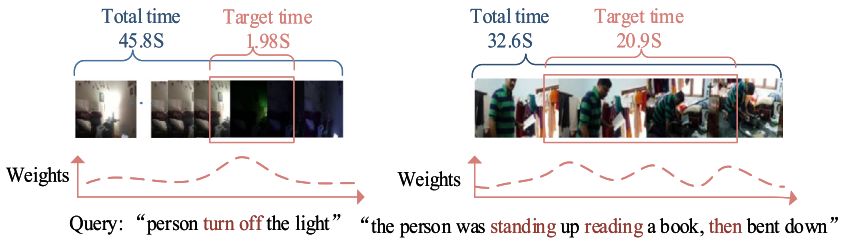



**Fig. 2.** An illustrative example of time concept.

**Table 1**
Duration analysis of datasets.

| Dataset | Average video duration | Average clip duration | Short video rate | Long video rate |
|---|---|---|---|---|
| ActivityNet Captions | 117.3 S | 37.8 S | 30.8% | 46.2% |
| TACoS | 233.6 S | 29.84 S | 34.6% | 36.0% |
| Charades-STA | 30.8 S | 8.00 S | 0.55% | 75.3% |

2. Exploring the time concept must be based on the combination of video, and how to effectively use visual information is crucial. It should make full use of visual communication as prior knowledge.

In summary, mining and applying the particular information in the sentence, especially the time-related semantics, is challenging. For this reason, we design the following network to realize the mining of time-related concepts in sentences.

## 3. System overview

In this section, we formulate the problem, introduce the overall system framework and the representation of language and video.

### 3.1. Problem formulation

Given an untrimmed video $V$ and a query sentence $S$, it aims to locate the segment $V_R$ that best matches the query sentence, get its start time $t_{start}$ and end time $t_{end}$ . In particular, take the video $V$ as a sequence arranged by frame, represented as $V = \{V_i\}_{i=0}^{l_v-1}$ ,where $V_i$is the i-th frame in the video and $l_v$is the total number of frames, the sentence $S$is represented as $S = \{S_j\}_{j=0}^{l_s-1}$ where $S_j$ is the word in the sentence and $l_s$ is the total number of all words in the sentence, and the target segments are $V_R = \{V_i\}_{i=\mathrm{m}}^{n}$,where the m is the start frame of the target segment, n is the end frame of the target segment.

To better integrate the video context information, we refer to the 2D adjacent network and use the two-dimensional map to represent the one-dimensional time. As shown in Fig. 5, the vertical axis START means the start time of a segment, ranging from the first second to the last second of the video, and the horizontal axis

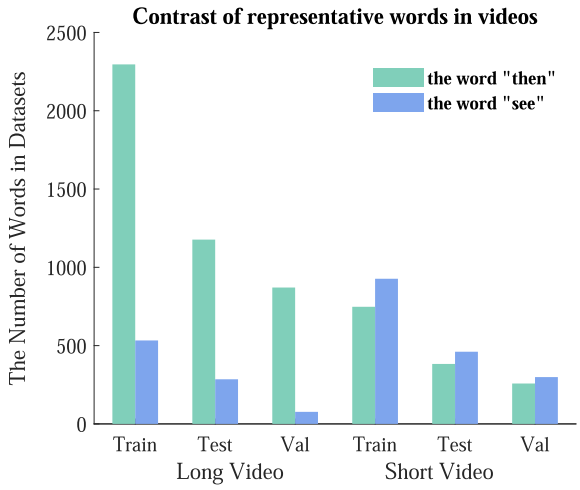



**Fig. 4.** Number of Typical Words in Long and Short Videos.

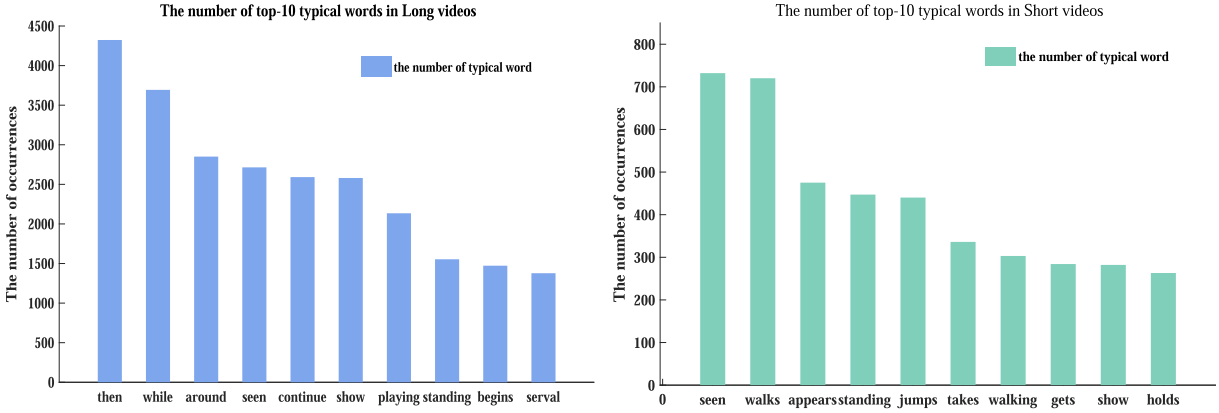



**Fig. 3.** Top-10 Typical Words in Long and Short Videos.

**Table 2**
The Contrast of the typical words in long and short videos of ActivityNet.

| Video dataset | | Long train | Long test | Long val | Short train | Short test | Short val |
|---|---|---|---|---|---|---|---|
| Video nums | | 16938 | 8179 | 7973 | 11359 | 4747 | 5396 |
| Then | Nums | 2296 | 871 | 1177 | 748 | 383 | 258 |
| | Proportion | **13.56%** | **14.39%** | **10.92%** | 6.59% | 8.07% | 4.81% |
| See | Nums | 533 | 285 | 77 | 927 | 461 | 299 |
| | Proportion | 3.14% | 3.48% | 0.97% | **8.16%** | **9.71%** | **5.54%** |

**Table 3**
The Contrast of the typical words in long and short videos of TACoS.

| Video dataset | | Long train | Long test | Long val | Short train | Short test | Short val |
|---|---|---|---|---|---|---|---|
| Video nums | | 1784 | 768 | 969 | 808 | 361 | 500 |
| Then | Nums | 138 | 65 | 95 | 16 | 7 | 13 |
| | Proportion | **7.74%** | **8.46%** | **9.80%** | 1.98% | 1.94% | 2.60% |
| Takes | Nums | 93 | 34 | 55 | 53 | 23 | 38 |
| | Proportion | 5.55% | 4.43% | 5.68% | **6.56%** | **6.37%** | **7.60%** |

indicates the END time; the range is also from the first second to the last second of the video. So each point in the map represents a video segment with a different start time and end time. The score of each point in the map means the predicted score of the corresponding video segment.

### 3.2. System framework

Fig. 6 shows the overall system framework. First,the video feature $F^V$ get through the C3D [5] model, and then use the 2D-TAN [6] method to generate a two-dimensional feature map $F^M$.The word vector $Q$ and sentence-level vector $F^S$ are generated through the Glove model and GRU. Then, on the one hand, merging the $F^M$ and $F^S$ to obtain the target prediction map $F^{MP}$ through the segment localization module, which represents the preliminary results of the target segment prediction. On the other hand, we can obtain the duration prediction $T_p$ of the target segment through the time concept mining module. Finally, we use the duration prediction map $F^{MT}$ to optimize the $F^{MP}$ and gain the final target prediction map $F^P$, in which each point in the map represents the prediction score of the corresponding segment.

### 3.3. Representation for language and video

**Language Representation.** To embed the query sentence, it first uses the Glove word2vec model [7] to encode each word $S_i$ to obtain the word vector $Q = (q_1, q_1, \ldots q_{l_s})$ , then put the word vector into a three layer bi-directional GRU [8] network, and take the last hidden layer of the network as the final feature vecto $F^S \in R^{d_S \times l_S}$ of query sentence, which $d_S$ is the feature vector dimension. The final feature of the sentence $F^S$ is obtained through fullly connected layers.

**Video Representation.** To obtain the context information in the video, we use the pre-trained 3D convolution network C3D to get the video feature $F^V \in R^{d_V \times N}$, where $d_V$ is the dimension of the video feature.

## 4. Our approach

### 4.1. Segment localization

From $F^V$ to $F^M$, the input vector of dimension $d_V \times N$ is transformed into a two-dimensional vector of dimension $d_V \times M \times M$ through combination and split based on 2D-TAN [6] method. The $F^S$ and $F^M$ are merged to get the target prediction map $F^{MP}$. Specifically, after joining $F^S$ and $F^M$ by the Hadamard product, use the maximum pooling method for convolution calculation. By the way, we also test the multi-modal fusion through vector splicing, and the results show that the direct vector multiplication is more concise and efficient.

From $F^{MP-first}$ to $F^{MP}$, the information of the context can be better integrated through two-dimensional convolution, and the score of each point in $F^{MP}$ is the probability that the moment is the target segment.

### 4.2. Generate the new sentence feature $F^{S-new}$

After the feature $F^S$ is generated through the *GRU* network, we need to pay more attention to the information that can reflect the time-related information in the sentence and strengthen the critical information before predicting the period.

To solve the challenges mentioned above, we design the sentence self-attention network and a visually guided attention network to explore the implicit information in sentences.

As shown in Fig. 7, we first get the sentence feature $F^S_{self}$ through the self-attention network, and then get the new feature $F^{S-new}$ under the guidance of visual information $F^V$, in the $F^{S-new}$, which the words related to time concept have a higher weight.

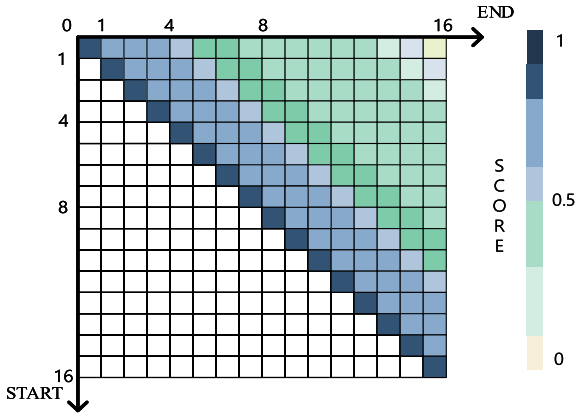



**Fig. 5.** Timespan prediction two-dimensional diagram.

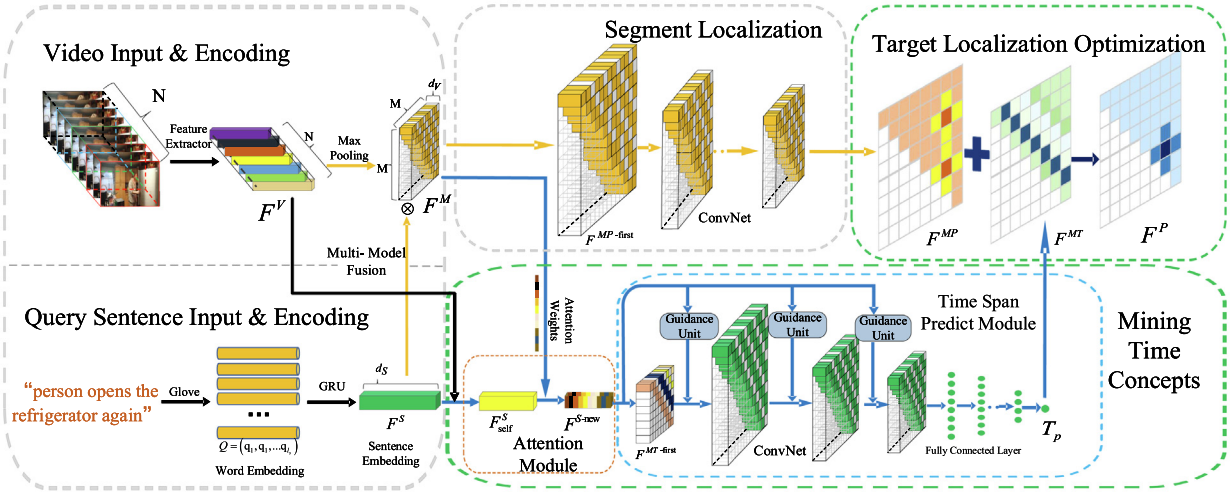



**Fig. 6.** System Frame Diagram.

In the process of calculating $F^{S-new}$, the specific calculation formula is as follows.We use the basic self-attention calculation method to obtain $F^{S}_{self}$

$$Q = W_Q F^S$$
$$K = W_K F^S$$
$$V = W_V F^S$$

where $W_Q, W_K$ and $W_V$ are the learnable parameter matrix, and then we use the formula to obtain the output vector $F^{S}_{self}$

$$\mathrm{F}^{S}_{\mathrm{self}\,i} = att((K,V),q_i) = \sum_{j=1}^{N} \alpha_{ij} v_j = \sum_{j=1}^{N} soft\max(s(k_j,q_i)) v_j \tag{1}$$

Where the attention scoring function $s$ use scaled dot product model as follows

$$s(k_j,q) = \frac{k_j^T q}{\sqrt{d}} \tag{2}$$

From $F^{S}_{self}$ to $F^{S-new}$, we still use the above formula 1 to guide the calculation process. We change the query vector from the vector itself to the visual information vector $F^M$ and finally get $F^{S-new}$

Specifically, as shown in Fig. 8, the weight of keywords such as the noun "person," the verb "writing," and "sitting" change significantly after the trained self-attention network.

The visually guided attention network, it can be seen from Fig. 9, the corresponding word weight that matches the video information will have a specific change, which shows that the guidance of visual information has a positive effect on the optimization of feature vectors.

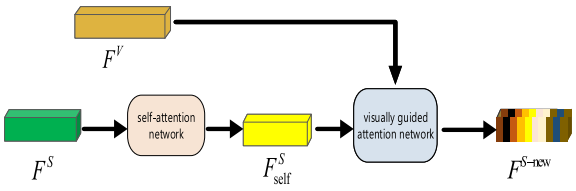



**Fig. 7.** generate new sentence feature.

### 4.3. Predict the period

After getting the new feature $F^{S-new}$, we focus on predicting the period joint with the guidance unit, which is to add the sentence guidance in the process of convolution and finally get the duration prediction $T_p$.

To apply the results, we convert the $T_p$ into a 2D time-predicted map $F^{MT}$ through a conversion function. Each point on the map represents the probability that the segment has the same duration as the target segment.

We merge $F^M$ and $F^{S-new}$ through the Hadamard product. Specifically, multiply each valid eigenvector in them, and the calculation formula is as follows:

$$F^{MT-\mathrm{first}} = \left\| \left(W_{MT} \cdot F^M\right) \odot \left(W_{ST} \cdot F^{S-new}\right) \right\|_{F_2} \tag{3}$$

where $W_{MT}$ and $W_{ST}$ is the weight matrix and $\|\ldots\|_{F_2}$ represents Frobenius regularization.

We use the attention unit to guide the fusion process again to strengthen the mining of semantic information. Because the start time is larger than the end time in the bottom left of the diagonal line in the two-dimensional map, it is an invalid area. All invalid points in the feature map are set to 0, a guidance unit based on the $F^{S-new}$ is used to guide each convolution process. We get the duration prediction $T_p$ through the L-layer fully connected layer.

As shown in Fig. 10, we adopt a conversion function to convert it into a two-dimensional map to optimize the positioning results, $T_p$ is transformed into a time-predicted 2D map $F^{MT}$ through a formula:

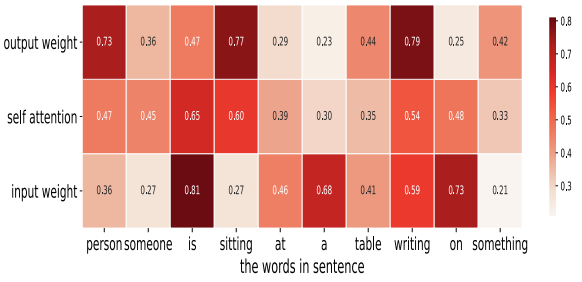



**Fig. 8.** Visualization on the changing of self attention weights.

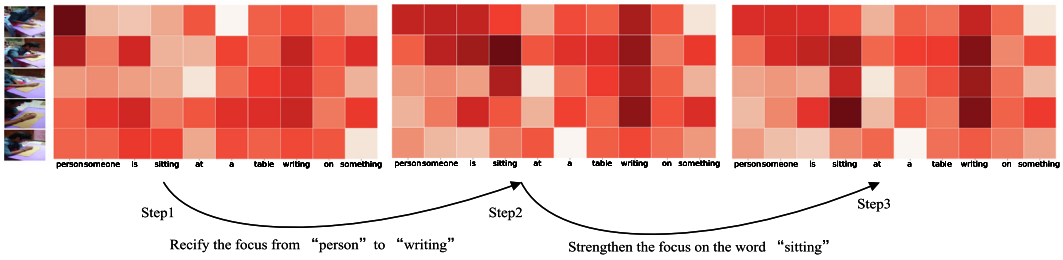



**Fig. 9.** Visualization on the changing of visually guided attention weights.

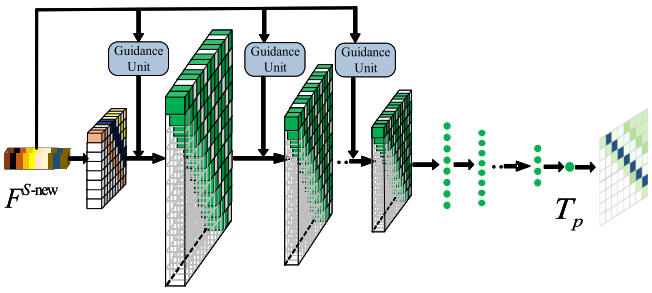



**Fig. 10.** Schematic diagram of time predict network.

$$t_{PM} = \lambda^{\|t_T - T_P\|} \tag{4}$$

Where $\lambda \in [0, 1]$ is the conversion factor, $T_P$ is the predicted duration, and $t_T$ is the real-time span corresponding to the video represented at each point in the 2D Map, the higher the score, the higher the confidence of the point. Meanwhile, because the period defined by the point on the same slash is identical, the scores of all points on the same slash are consistent.

### 4.4. Target localization optimization

The target prediction map $F^{MP}$ and the time-predicted map $F^{MT}$ are superimposed to determine the final target prediction map $F^{P}$. As shown in Fig. 11, the weights of all points in $F^{MP}$ are further optimized, especially the points with a high score but with a large duration error were excluded to improve the accuracy of final localization. All prediction scores are integrated to obtain $P = \{p_i\}_{i=1}^{C}$, where $p_i$is the i-th predict score and $C$ is the total number of all candidate segments. Arrange all scores from high to low, and the highest score is the video segment that most match the query text.

### 4.5. Loss function

During the training of the STCM-Net network, we design two $Loss_\alpha$ for segment localization module $Loss_\alpha$ and a mining time concept module $Loss_\beta$ .

**Localization Loss** In the segment localization module, the value $IoU$ is used as a measure, specifically for each candidate segment. The label value is calculated as follows:

$$y_i = \begin{cases} 0 & IoU_i < T_{\min} \\ (IoU_i - T_{\min})/(T_{\max} - T_{\min}) & T_{\min} \leqslant IoU_i \leqslant T_{\max} \\ 1 & IoU_i > T_{\max} \end{cases} \tag{5}$$

where $T_{max}$ and $T_{min}$ are the upper and lower thresholds, respectively, and $y_i$ are the loss label, where $Loss_\alpha$ is calculated by the following loss:

$$Loss_\alpha = \frac{1}{C}\sum_{i=1}^{C} y_i \log p_i + (1 - y_i)\log(1 - p_i) \tag{6}$$

where $C$ is the total number of candidate segments, and $p_i$ is the forecast score for each target segment.

**Mining Time Concepts Loss** In the time concept mining module, the root mean square error is used to calculate the loss value:

$$Loss_\beta = \frac{1}{C}\sum_{i=1}^{C}(x_i - t_T)^2 \tag{7}$$

where $t_T$ is the actual time duration of each candidate segment, $x_i$ is the predicted time duration for each given video and sentence, and $C$ is the total number of all candidate segments.

**Joint Loss Training** We balance the ratio during training as follows:

$$Loss = Loss_\alpha + \eta Loss_\beta \tag{8}$$

where $\eta$is used to balance the two loss terms, the STCM-Net can be trained in an end-to-end manner by minimizing the total loss.

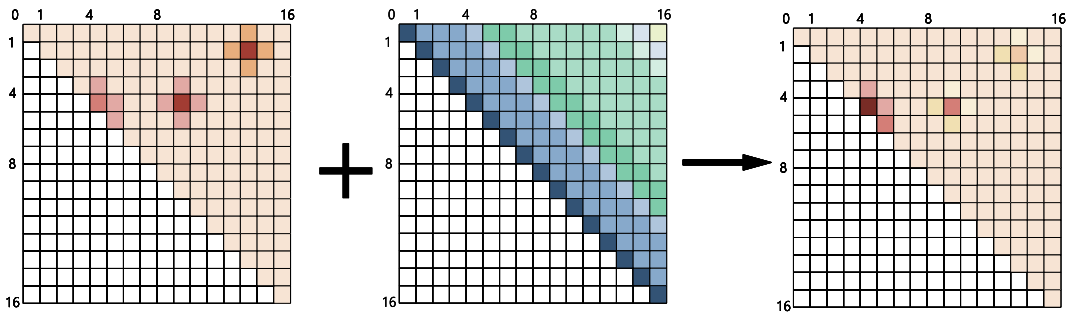



**Fig. 11.** Schematic diagram of positioning optimization.

## 5. Experiments

We test our model on three large-scale public datasets: Charades-STA (Sigurdsson et al. 2016), ActivityNet Captions (Krishna et al. 2017), and TACoS (Regneri et al. 2013).

### 5.1. Datasets introduction

**Charades-STA**. It contains 9,848 videos of daily indoor activities. Gao [9] extends the temporal annotation of this dataset with language descriptions and calls it Charades-STA. It contains 12,408 moment-sentence pairs in the training set and 3,720 pairs in the testing set.

**ActivityNet Captions**. It is a large dataset that contains 20 k videos with 100 k language descriptions, whose content is diverse and open. It contains 37,417, 17,505, and 17,031 moment-sentence pairs for training, validation, and testing, respectively.

**TACoS**. It consists of 127 videos selected from the MPII Cooking Composite Activities video corpus (Rohrbach et al. 2012), containing different activities in the kitchen room. Regnery et al. extend the sentence descriptions through crowd-sourcing. A standard split (Gao et al. 2017) consists of 10,146,4,589 and 4,083 moment-sentence pairs for training, validation, and testing, respectively. Table 4.

### 5.2. Evaluation metrics

Following the general method, "Rank@n, IoU@m" is used as the evaluation metrics, "Rank@n, IoU@m" is defined as the percentage of the language queries having at least one matched retrieval (IoU with the ground-truth moment is larger than m) in the top-n retrieved moments.

### 5.3. Implementation details

Different parameters were set for different datasets. Specifically, we report the results as $n \in \{1, 5\}$with $m \in \{0.5, 0.7\}$ for Charades-STA, $n \in \{1, 5\}$ with $m \in \{0.3, 0.5, 0.7\}$for ActivityNet Captions, and $m \in \{0.1, 0.3, 0.5\}$ for TACoS.

For a fair comparison with the state-of-the-art approaches, this paper uses the same visual feature extraction method as the previous work MAN[10], CMIN[11]. Specifically, we use the VGG feature for Charades-STA and the C3D feature for ActivityNet Captions and TACoS.

The dimension of the video feature $d^V$ is 512, and the dimension of the sentence feature $d^S$ is 768. The dimensions of the feature map $M$ are set to 64,64 and 128 for Charades-STA, ActivityNet Captions, and TACoS, respectively. The number of frames in a clip was 4 in Charades-STA and 16 in TACoS and ActivityNet Captions. The conversion factor $\lambda$is set to 0.1 in Charades-STA and TACoS and 0.6 in the ActivityNet Captions. The joint loss weight $\eta$is set to 0.01 in all three datasets.

### 5.4. Comparison of state-of-the-art methods

We evaluate our proposed method on three popular benchmarks: Charades-STA, TACoS, and the ActivityNet Captions. The results show in Tables 5–7, respectively, where we compare our method with three different frameworks methods.

One-stage single-shot approaches: The candidate segment is predicted directly after the video and sentence are merged. The representative methods are ABLR[12], 2D-TAN [6], SCDM [13], and FIAN [4].

Two-stage approaches: This framework selects candidate segments in two steps, first by generating candidate segments and then ranking the candidate segments. Many methods based on sliding window or anchor-based are proposed to solve this task, including MCN [14], CTRL [9], ACRN [2], ACL-K [1], VAL [15], SLTA [16].

Reinforcement learning-based(RL) approaches: Many models based on reinforcement learning have been proposed to solve temporal language localization, among which TripNet [17], and SM-RL [18] are representative.

Table 5 shows a comparison of the STCM-Net with the 2D-TAN model indicates the Rank1 evaluation metrics. Increased by an average of 2%, and the Rank5 evaluation metrics improved by 5% and 7%. Although the accuracy is not high enough compared to FIAN models, the improvement still fully demonstrates the positive effect of the mining time concept module.

For the FIAN model, it takes a fine-grained iterative attention network to extract bilateral query-video information. Our model improves the unidirectional interactions from video to query to extract better the information that matches the sentence and video. For relatively complex video scenes, it can achieve better results, especially on the Rank1 evaluation metrics. For the design intention, our model is more inclined to extract the specific semantic information and is more sensitive to the time-related information in the sentence.

However, this pipeline also has some significant shortcomings. First of all, although it uses the weight sharing method, the training cost will still be increase. Secondly, it will reduce the localization accuracy for the simple video scenes d. As shown, the follows performance on the TACoS dataset. In addition, due to its fine-grained iterative attention matching, the segment predicted by the network are relatively concentrated, so the Rank5 evaluation metrics will be reduced, as shown in the ActivityNet Captions Rank5@0.5 and Rank5@0.7.

Table 6 shows our model significantly affected the TACoS, especially with more than 10% improvement on the Rank5 evaluation metrics. The STCM-Net model performs better on this dataset than the FIAN model, which depends on the simple optimization of our time concept mining module. At the same time, because of the small size and relatively single scene of the TACoS, the prediction of the duration is relatively easy, leading to a more remarkable improvement in overall accuracy.

Table 7 shows that the improvement of ActivityNet Captions is not significant, but still have 1% to 1.5% improvement compared to the 2D-TAN model, and both Rank5@IoU0.5 and Rank5@IoU0.7 outperform other models. We think the reason is that the content involved in the dataset is comprehensive and extensive, which makes it relatively difficult to predict the duration of a sentence in different scenarios. However, it still improves on the dataset, especially on the Rank5 evaluation metrics, which proves the effectiveness of our model.

### 5.5. Qualitative results

The above results indicate that the addition of the time concept mining module positively affects this task, especially on the evaluation index of Rank5 evaluation metrics. If the overall framework is unchanged, the addition of the time concept mining module is reasonable to improve the Rank5 evaluation metrics. The duration prediction map $F^{MT}$ is equivalent to applying a higher score to the right candidate segments, which gives a higher score to the top 5 candidate segments, ultimately leading to a more significant improvement of the Rank5 evaluation metrics.

The simple scene localization and Rank5 evaluation metrics are still essential for the task. For this reason, we chose to use the current pipeline, and we plan to explore the bidirectional guide in the future.

**Table 4**
The video number & moment-sentence pairs of datasets.

| Dataset | Video number | Train pairs | Validation pairs | Test pairs |
|---|---|---|---|---|
| ActivityNet Captions | 20 K | 37417 | 17505 | 17031 |
| TACoS | 127 | 10146 | 4589 | 4083 |
| Charades-STA | 9848 | 12408 | – | 3720 |

**Table 5**
Comparison of state-of-the-art methods on Charades-STA.

| Framework | Method | Rank1 @IoU0.5 | Rank1 @IoU0.7 | Rank5 @IoU0.5 | Rank5 @IoU0.7 |
|---|---|---|---|---|---|
| RL | SM-RL | 24.36 | 11.17 | 61.25 | 32.08 |
| | TripNet | 36.61 | 14.50 | – | – |
| | RWM | 34.12 | 13.74 | – | – |
| Two-stage | MCN | 17.46 | 8.01 | 48.22 | 26.73 |
| | CTRL | 23.63 | 8.89 | 58.92 | 29.52 |
| | ACRN | 20.26 | 7.64 | 71.99 | 27.79 |
| | ACL-K | 30.48 | 12.20 | 64.84 | 35.13 |
| | SLTA | 22.81 | 8.25 | 72.39 | 31.46 |
| One-stage | ABLR | 24.36 | 9.01 | – | – |
| | FIAN | **58.55** | **37.72** | **87.80** | **63.52** |
| | 2D-TAN | 42.80 | 23.25 | 80.54 | 54.14 |
| | **Ours** | 44.30 | 25.99 | 85.08 | 61.34 |

**Table 6**
Comparison of state-of-the-art methods on TACoS.

| Framework | Method | Rank1 @IoU0.1 | Rank1 @IoU0.3 | Rank1 @IoU0.5 | Rank5 @IoU0.1 | Rank5 @IoU0.3 | Rank5 @IoU0.5 |
|---|---|---|---|---|---|---|---|
| RL | SM-RL | 26.51 | 20.25 | 15.95 | 50.01 | 38.47 | 27.84 |
| | TripNet | – | 23.95 | 19.17 | – | – | – |
| Two-stage | MCN | 14.42 | - | 5.58 | 37.35 | – | 10.33 |
| | CTRL | 24.32 | 18.32 | 13.30 | 48.73 | 36.69 | 25.42 |
| | ACRN | 24.22 | 19.52 | 14.62 | 47.42 | 34.97 | 24.88 |
| | ACL-K | 31.64 | 24.17 | 20.01 | 57.85 | 42.15 | 30.66 |
| | SLTA | 23.13 | 17.07 | 11.92 | 46.52 | 32.90 | 20.86 |
| One-stage | ABLR | 34.70 | 19.50 | 9.40 | – | – | – |
| | FIAN | 39.55 | 33.87 | 28.58 | 56.14 | 47.76 | 39.16 |
| | 2D-TAN | 47.59 | 37.29 | 25.32 | 70.31 | 57.81 | 45.04 |
| | **Ours** | **60.33** | **49.04** | **35.59** | **81.73** | **70.13** | **57.69** |

## 6. Ablation study

### 6.1. Influence of the attention module

In this section, we present the ablation studies to understand the influence of the attention module. Period predicts module and the effect of the hyperparameters setting. We re-train our approach with the following settings:

- [•]**STCM-Attention:**Instead of generating the $F^{S-new}$ through the attention module first and then generating the $T_P$ through the period predict module, We only keep the attention module and remove the period predict module. We directly get the $T_P$ from the new sentence feature $F^{S-new}$ through the fully connected layer.
- [•]**STCM-Time-Predict:** We only keep the period predict module and remove the attention module without extracting the typical words. We directly input the sentence feature $F^S$ into the time predict module
- [•]**STCM-Net:** Our full of STCM-Net model.

Table 8 shows the performance comparisons of our STCM-Net and these ablations on the TACoS dataset. As shown in the table, when we only keep the Attention module, the performance is significantly better than the origin 2DTAN network in each evaluation

**Table 7**
Comparison of state-of-the-art methods on ActivityNet Captions.

| Framework | Method | Rank1 @IoU0.3 | Rank1 @IoU0.5 | Rank1 @IoU0.7 | Rank5 @IoU0.3 | Rank5 @IoU0.5 | Rank5 @IoU0.7 |
|---|---|---|---|---|---|---|---|
| RL | TripNet | 48.42 | 32.19 | 13.93 | – | – | – |
| Two-stage | MCN | 39.35 | 21.36 | 6.43 | 68.12 | 52.23 | 29.70 |
| | CTRL | 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 |
| One-stage | ABLR | 55.67 | 36.79 | – | – | – | – |
| | FIAN | **64.10** | **47.90** | **29.81** | **87.59** | 77.64 | 59.66 |
| | 2D-TAN | 59.45 | 44.51 | 26.54 | 85.53 | 77.13 | 61.96 |
| | **Ours** | 61.18 | 46.23 | 29.04 | 86.81 | **78.43** | **63.46** |

**Table 8**
Ablation study on the TACoS dataset.

| Method | Rank1 @IoU0.1 | Rank1 @IoU0.3 | Rank1 @IoU0.5 | Rank5 @IoU0.1 | Rank5 @IoU0.3 | Rank5 @IoU0.5 |
|---|---|---|---|---|---|---|
| Origin-2DTAN | 47.59 | 37.29 | 25.32 | 70.31 | 57.81 | 45.04 |
| STCM-Attention | 60.21 | 47.46 | 34.19 | 79.58 | 69.46 | 57.91 |
| STCM-Time-Predict | 56.01 | 43.99 | 30.59 | 75.55 | 65.86 | 51.81 |
| STCM-Net | 60.33 | 49.04 | 35.59 | 81.73 | 70.13 | 57.69 |

metrics. It also can be seen that the effect of the attention module is much more significant than that of the period predict module.

Table 9 and Table 10 show the performance on the Activitynet and Charades-STA dataset. When we only keep the attention module, it still has an improvement compared to the original 2DTAN model, which shows that the attention module has a positive effect on performance improvement.

### 6.2. Influence of the time span predict module

In this module, we design an STCM-Time-Predict model that removes the attention module and directly inputs the sentence feature $F^S$ into the period predict module to verify the role of the period prediction module. And the results on the three datasets are shown in Table 8,9, and Table 10, respectively. Table 8 shows that only keeping the period prediction module on the TACoS dataset also has a significant improvement. The performance on the remaining two datasets proves that the period prediction module has a positive effect.

It also can be seen from the results that the attention module plays a more critical role in the entire mining time concepts module.

### 6.3. Influence of time-related typical words

In this part, we further verify the role of time-related typical words in the sentences. Specifically, we counted three typical time-related words in the three public data sets, removed the top ten keywords in each data set, and then trained and tested them on the 2dtan and our models. The results are shown in Table 11. The results show that after the keywords are removed, the original model's performance and our model have declined, indicating that the keywords have their effects. And our model is more sensitive to the typical words. Indicate our positive significance in the mining and use of related keywords.And after removing the keywords, our model still performs better than the original model in most cases.

### 6.4. Influence of the hyper parameters

**Influence of The Conversion Factor $\lambda$**

The value of the $\lambda$ represents the confidence in $T_P$. During the conversion, the smaller the $\lambda$, the more segments that are too different from the prediction duration will be excluded, the efficiency and accuracy of the localization will be improved. At the same time, if our predicted prediction $T_P$ is wrong, it will eliminate many correct answers. We set $\lambda$ for 0.1 and 0.6 as representative.

Fig. 12 shows that the influence of the different $\lambda$ is undeniable. For the ActivityNet Captions, because of a large amount of data and the more complex and diverse forms of time concept, it is more difficult to extract the time concept, and confidence of the predicted duration $T_P$ is low. So a larger conversion factor can prevent the erroneous exclusion of the candidate segments and improve the accuracy. For the other two datasets, a small conversion factor can effectively eliminate error candidates because of the small amount of data and high accuracy of $T_P$.

**Influence of The Joint Loss Weight $\eta$**

In calculating the joint loss, the value of the coefficient $\eta$ determines the weight of the two modules in the optimization process. The larger value of $\eta$ will lead to a decrease in localization accuracy. For this reason, this paper adopts smaller $\eta$ values, and we test the cases of $\eta = 0.01$ and $\eta = 0.1$, and the results are as follows.

Table 12 shows that when the joint loss coefficient is increased from 0.01 to 0.1, the overall accuracy rate is slightly reduced in the three datasets. This indicates that the time concept mining module as the auxiliary optimization localization module, the larger loss coefficient will decline the positioning accuracy rate.

## 7. Related works

In recent years, more and more people have begun to pay attention to the tasks of temporal language localization. It is generally believed that since (Gao J et al. 2017) proposed the CTRL [9] model in 2017, the temporal language localization task was considered as formally proposed. [19] explored an efficient way to fusion the text information with the visual information. And by 2020 (Zhang S, Peng H, et al., 2020) proposed the 2D-TAN [6] model to optimize the one-dimensional time representation task. (Liu, Daizong and Qu, et al., 2020) [20] proposed a Deep Rectification-Modulation Network to optimize the deficiencies of single-step attentional frameworks. [21] used the distances between the frame as dense supervisions to improve the video grounding accuracy. Many researchers have made significant progress in feature extraction and fusion and the final temporal language localization.

In computer vision, many tasks are very similar to this task and have particular relevance. For example, the temporal action localization in the video, where [5] explores the extraction of video features by a three-dimensional convolutional network, [22–26] and other representative works explored different network structures for efficiently detects the temporal actions in the video. But there are two differences between the temporal action localization task and the temporal language localization task. On the one hand, the traditional temporal action localization task can only locate a specific action. Our goal is to locate the action described in the sentence and include other objects, such as the description of the scene, animals, and the environment. On the other hand, the temporal action localization task needs to mark all the possible actions

**Table 9**
Ablation study on the ActivityNet Captions dataset.

| Method | Rank1 @IoU0.3 | Rank1 @IoU0.5 | Rank1 @IoU0.7 | Rank5 @IoU0.3 | Rank5 @IoU0.5 | Rank5 @IoU0.7 |
|---|---|---|---|---|---|---|
| Origin-2DTAN | 59.45 | 44.51 | 26.54 | 85.53 | 77.13 | 61.96 |
| STCM-Attention | 60.08 | 44.72 | 28.17 | 86.27 | 78.10 | 63.77 |
| STCM-Time-Predict | 59.63 | 44.95 | 26.39 | 86.13 | 77.50 | 60.67 |
| STCM-Net | 61.18 | 46.23 | 29.04 | 86.81 | 78.43 | 63.46 |

**Table 10**
Ablation study on the Charades-STA dataset.

| Method | Rank1 @IoU0.5 | Rank1 @IoU0.7 | Rank5 @IoU0.5 | Rank5 @IoU0.7 |
|---|---|---|---|---|
| Origin-2DTAN | 42.80 | 23.25 | 80.54 | 54.14 |
| STCM-Attention | 43.68 | 24.70 | 82.64 | 59.74 |
| STCM-Time-Predict | 43.29 | 23.95 | 81.64 | 58.65 |
| STCM-Net | 44.30 | 25.99 | 85.08 | 61.34 |

**Table 11**
Comparison of the effect of time-related typical words on three datasets.

| TACoS | R1@0.1 | R1@0.3 | R1@0.5 | R5@0.1 | R5@0.3 | R5@0.5 |
|---|---|---|---|---|---|---|
| Ours Model | 60.33 | 49.04 | 35.59 | 81.73 | 70.13 | 57.69 |
| Prune typical words | 56.84 | 45.71 | 33.52 | 78.73 | 67.34 | 57.24 |
| 2D-TAN | 47.59 | 37.29 | 25.32 | 70.31 | 57.81 | 45.04 |
| Prune typical words | 42.61 | 31.07 | 21.25 | 66.98 | 52.96 | 42.09 |
| **ActivityNet** | **R1@0.3** | **R1@0.5** | **R1@0.7** | **R5@0.3** | **R5@0.5** | **R5@0.7** |
| Ours Model | 61.18 | 46.23 | 29.04 | 86.81 | 78.43 | 63.46 |
| Prune typical words | 59.29 | 45.27 | 26.23 | 84.39 | 76.69 | 62.38 |
| 2D-TAN | 59.45 | 44.51 | 26.54 | 85.53 | 77.13 | 61.96 |
| Prune typical words | 60.23 | 44.28 | 26.67 | 85.39 | 76.71 | 60.87 |
| **Charades-STA** | **R1@0.3** | **R1@0.5** | **R1@0.7** | **R5@0.3** | **R5@0.5** | **R5@0.7** |
| Ours Model | 59.33 | 44.30 | 25.99 | 94.06 | 85.08 | 61.34 |
| Prune typical words | 56.05 | 42.15 | 23.66 | 92.26 | 82.53 | 59.27 |
| 2D-TAN | – | 42.80 | 23.25 | – | 80.54 | 54.14 |
| Prune typical words | – | 39.95 | 23.15 | – | 79.11 | 53.20 |

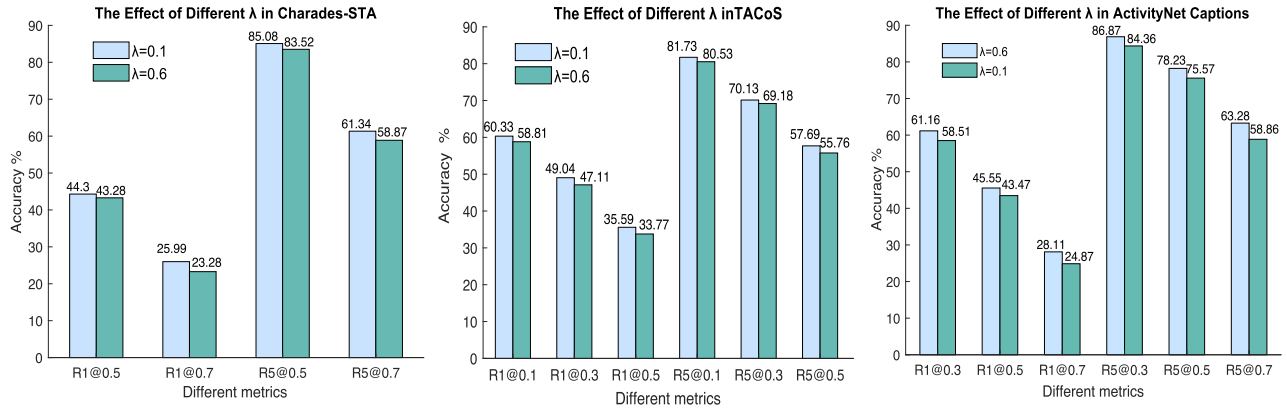



**Fig. 12.** Effect of Conversion Factor $\lambda$ on Accuracy.

in advance, but The actual scene is complex and changeable; there is no way to label all activities. Our network does not need to label all possible activities.

In addition, The task of image localization through natural language is to locate the local position in the image described by language, and its mainstream pipeline is very similar to our task. For example, in [27], also used a two-stage method was used, which is very similar to the pipeline of temporal language localization tasks. First, each candidate area is generated and then scored. In [28–30], some models have been proposed to extract and optimize the tar-

**Table 12**
Comparison of the effect of $\eta$ on three datasets.

| TACoS | R1@0.1 | R1@0.3 | R1@0.5 | R5@0.1 | R5@0.3 | R5@0.5 |
|---|---|---|---|---|---|---|
| $\eta$**=0.01** | 60.33 | 49.04 | 35.59 | 81.73 | 70.13 | 57.69 |
| $\eta$=0.1 | 59.56 | 47.11 | 33.77 | 82.53 | 69.68 | 56.76 |
| **ActivityNet** | **R1@0.3** | **R1@0.5** | **R1@0.7** | **R5@0.3** | **R5@0.5** | **R5@0.7** |
| $\eta$**=0.01** | 61.18 | 46.23 | 29.04 | 86.81 | 78.43 | 63.46 |
| $\eta$=0.1 | 61.16 | 45.55 | 28.11 | 86.87 | 78.23 | 63.28 |
| **Charades-STA** | **R1@0.3** | **R1@0.5** | **R1@0.7** | **R5@0.3** | **R5@0.5** | **R5@0.7** |
| $\eta$**=0.01** | 59.33 | 44.30 | 25.99 | 94.06 | 85.08 | 61.34 |
| $\eta$=0.1 | 58.76 | 45.13 | 25.22 | 93.58 | 85.30 | 60.19 |

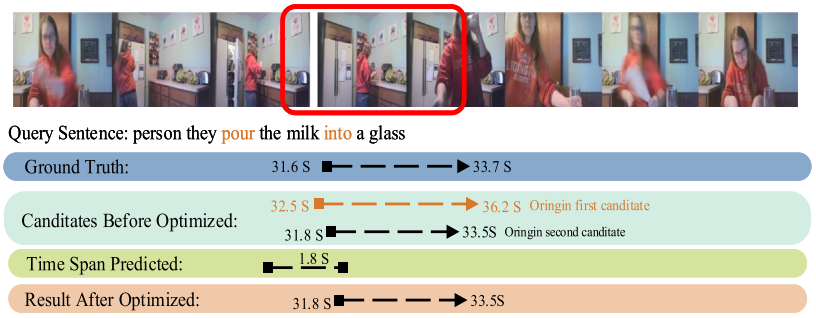

**Fig. A1.** Some examples of the influence about the time span prediction on the final positioning.

get area according to the reconstruction error rate and probability. In [31–34], some models achieved higher precision in the target area and global semantic information acquisition.

Compared to image localization by language, video temporal language localization needs to process more complex context information in video and sentence. However, the two tasks still have some essential ideas to learn from each other. The image localization by language and temporal action localization can be regarded as the simplified version of our task.

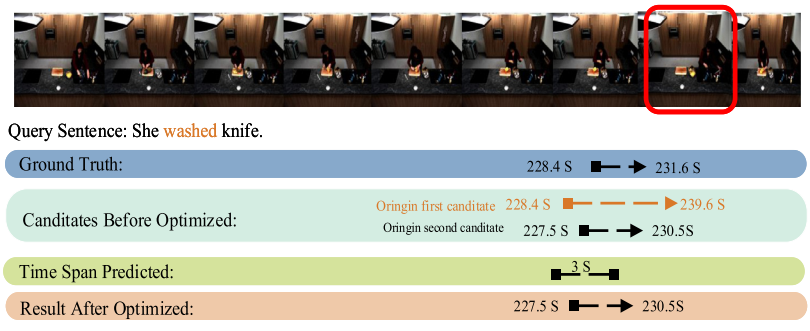
## 8. Conclusion

In this paper, we explore some specific aspect semantics of the query sentence in-depth, especially the exploration of the time-




(e)


(f)




(g)

**Fig. 13** (*continued*)

related information in a sentence. A simple neural network is also designed to mine the time concept semantics in the sentence, which can make up for the insufficient mining of sentence semantics in previous work. In this network, we combine the visual information to guide the extraction and use the results to optimize the localization. The results from the public datasets show that time concept mining in the sentence has a positive effect on improving localization accuracy. We also achieved significant improvement on the three public datasets compared with the SOTA model. In the future, we plan to explore more sentence semantics and the bidirectional guide method to achieve fundamental unlabeled video understanding and video retrieval.

## CRediT authorship contribution statement

**Zixi Jia:** Conceptualization, Methodology. **Minglin Dong:** Software, Data curation, Writing - original draft. **Jingyu Ru:** Conceptualization, Methodology. **Lele Xue:** Supervision. **Sikai Yang:** Software, Data curation. **Chunbo Li:** Visualization, Writing - review & editing.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This work was funded by the National Natural Science Foundation of China under Grants (61872073, 61973093, 61901098, 61971118 and 61973063); the National Key Robot Project Grant No.\2017YFB1300900.

## Appendix A

We select some illustrative examples of the Charades-STA and TACoS datasets to show the effect of the semantic mining of sentences and the impact of time concept mining as in Figure A1. The ground truth is the actual time of the target segment, and the candidates before optimization refer to the video segment obtained through the segment localization module. The original first candidate is the segment with the highest score. The predicted time span refers to the target segment's possible duration obtained through the mining time concept module. The results show that optimization can exclude some segments with higher scores from the original segment localization module but are pretty different from the actual situation and select the candidate closer to the ground truth.

## References


[1] R. Ge, J. Gao, K. Chen, R. Nevatia, Mac: Mining activity concepts for language-based temporal localization, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV) IEEE, 2019, pp. 245–253.

[2] M. Liu, X. Wang, L. Nie, X. He, B. Chen, T.-S. Chua, Attentive moment retrieval in videos, in: The 41st international ACM SIGIR conference on research & development in information retrieval, 2018, pp. 15–24.

[3] J. Dong, X. Li, C. Xu, X. Yang, G. Yang, X. Wang, M. Wang, Dual encoding for video retrieval by text, IEEE Trans. Pattern Anal. Mach. Intell..

[4] X. Qu, P. Tang, Z. Zou, Y. Cheng, J. Dong, P. Zhou, Z. Xu, Fine-grained iterative attention network for temporal language localization in videos, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 4280–4288.

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497.

[6] S. Zhang, H. Peng, J. Fu, J. Luo, Learning 2d temporal adjacent networks for moment localization with natural language, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 12870–12877..

[7] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[8] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[9] J. Gao, C. Sun, Z. Yang, R. Nevatia, Tall: Temporal activity localization via language query, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 5267–5275.

[10] D. Zhang, X. Dai, X. Wang, Y.-F. Wang, L.S. Davis, Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1247–1257.

[11] Z. Zhang, Z. Lin, Z. Zhao, Z. Xiao, Cross-modal interaction networks for query-based moment retrieval in videos, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 655–664.

[12] Y. Yuan, T. Mei, W. Zhu, To find where you talk: Temporal sentence localization in video with attention based location regression, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 9159–9166..

[13] Y. Yuan, L. Ma, J. Wang, W. Liu, W. Zhu, Semantic conditioned dynamic modulation for temporal sentence grounding in videos, arXiv preprint arXiv:1910.14303..

[14] L. Anne Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, B. Russell, Localizing moments in video with natural language, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 5803–5812.

[15] X. Song, Y. Han, Val: visual-attention action localizer, in: Pacific Rim Conference on Multimedia, Springer, 2018, pp. 340–350..

[16] B. Jiang, X. Huang, C. Yang, J. Yuan, Cross-modal video moment retrieval with spatial and language-temporal attention, in: Proceedings of the 2019 on International Conference on Multimedia Retrieval, 2019, pp. 217–225.

[17] M. Hahn, A. Kadav, J.M. Rehg, H.P. Graf, Tripping through time: Efficient localization of activities in videos, arXiv preprint arXiv:1904.09936..

[18] W. Wang, Y. Huang, L. Wang, Language-driven temporal activity localization: a semantic matching reinforcement learning model, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 334–343.

[19] H. Xu, K. He, B.A. Plummer, L. Sigal, S. Sclaroff, K. Saenko, Multilevel language and vision integration for text-to-clip retrieval, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 9062–9069..

[20] D. Liu, X. Qu, J. Dong, P. Zhou, Reasoning step-by-step: Temporal sentence localization in videos via deep rectification-modulation network, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 1841–1851.

[21] R. Zeng, H. Xu, W. Huang, P. Chen, M. Tan, C. Gan, Dense regression network for video grounding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10287–10296.

[22] T. Lin, X. Zhao, Z. Shou, Single shot temporal action detection, in: Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 988–996.

[23] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: Single shot multibox detector, in: European conference on computer vision, Springer, 2016, pp. 21–37..

[24] B. Singh, T.K. Marks, M. Jones, O. Tuzel, M. Shao, A multi-stream bi-directional recurrent neural network for fine-grained action detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1961–1970.

[25] H. Xu, A. Das, K. Saenko, R-c3d: Region convolutional 3d network for temporal activity detection, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 5783–5792.

[26] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, arXiv preprint arXiv:1506.01497..

[27] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2016) 1137–1149.

[28] J. Mao, J. Huang, A. Toshev, O. Camburu, A.L. Yuille, K. Murphy, Generation and comprehension of unambiguous object descriptions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 11–20.

[29] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, T. Darrell, Natural language object retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4555–4564.

[30] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, B. Schiele, Grounding of textual phrases in images by reconstruction, European Conference on Computer Vision, Springer (2016) 817–834.

[31] L. Yu, P. Poirson, S. Yang, A.C. Berg, T.L. Berg, Modeling context in referring expressions, European Conference on Computer Vision, Springer (2016) 69–85.

[32] K. Chen, R. Kovvuri, J. Gao, R. Nevatia, Msrc: Multimodal spatial regression with semantic context for phrase grounding, in: Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, 2017, pp. 23–31.

[33] K. Chen, R. Kovvuri, R. Nevatia, Query-guided regression network with context policy for phrase grounding, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 824–832.

[34] H. Zhang, Y. Niu, S.-F. Chang, Grounding referring expressions in images by variational context, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4158–4166.


**Zixi Jia** received the Ph.D. degree in pattern recognition and intelligent systems from Northeastern University, Shenyang, China, in 2009, where he is currently an Associate Professor and the Vice Dean of Faculty of Robot Science and Engineering. His research interests include artificial intelligence, robotics, big data and wireless sensor networks.

**Minglin Dong** A graduate student of Faculty of Robot Science and Engineering, Northeastern University,Shenyang,China.The research intersts include computer vision, natural language processing and multimodal fusion.

**Jingyu Ru** received the Ph.D. degrees in Northeastern University, Shenyang, China, in 2019. He is currently a post doctor as a staff in Robot Science and Engineering in Northeastern University, China. His research interests include artificial intelligence, big data, wireless sensor networks and robotics.

**Sikai Yang** A graduate student of Faculty of Robot Science and Engineering, Northeastern University,Shenyang,China.The research interst is computer vision.

**Lele Xue** A graduate student of Faculty of Robot Science and Engineering, Northeastern University,Shenyang, China.The research interst is computer vision and 3D reconstruction of virtual reality.

**Chunbo Li** A graduate student of Faculty of Robot Science and Engineering, Northeastern University, Shenyang, China. The research intersts is computer vision.