# STRONG: Spatio-Temporal Reinforcement Learning for Cross-Modal Video Moment Localization

Da Cao
Hunan University
caoda0721@gmail.com

Yawen Zeng
Hunan University
yawenzeng11@gmail.com

Meng Liu
Shandong Jianzhu University
mengliu.sdu@gmail.com

Xiangnan He
University of Science and Technology of China
xiangnanhe@gmail.com

Meng Wang
Hefei University of Technology
eric.mengwang@gmail.com

Zheng Qin*
Hunan University
zqin@hnu.edu.cn

## ABSTRACT

In this article, we tackle the cross-modal video moment localization issue, namely, localizing the most relevant video moment in an untrimmed video given a sentence as the query. The majority of existing methods focus on generating video moment candidates with the help of multi-scale sliding window segmentation. They hence inevitably suffer from numerous candidates, which result in the less effective retrieval process. In addition, the spatial scene tracking is crucial for realizing the video moment localization process, but it is rarely considered in traditional techniques. To this end, we innovatively contribute a spatial-temporal reinforcement learning framework. Specifically, we first exploit a temporal-level reinforcement learning to dynamically adjust the boundary of localized video moment instead of the traditional window segmentation strategy, which is able to accelerate the localization process. Thereafter, a spatial-level reinforcement learning is proposed to track the scene on consecutive image frames, therefore filtering out less relevant information. Lastly, an alternative optimization strategy is proposed to jointly optimize the temporal- and spatial-level reinforcement learning. Thereinto, the two tasks of temporal boundary localization and spatial scene tracking are mutually reinforced. By experimenting on two real-world datasets, we demonstrate the effectiveness and rationality of our proposed solution.

## CCS CONCEPTS

• **Information systems → Multimedia and multimodal retrieval**; • **Theory of computation → Reinforcement learning**.

## KEYWORDS

Video Moment Localization, Cross-Modal Retrieval, Reinforcement Learning, Alternative Optimization

**ACM Reference Format:**
Da Cao, Yawen Zeng, Meng Liu, Xiangnan He, Meng Wang, and Zheng Qin. 2020. STRONG: Spatio-Temporal Reinforcement Learning for Cross-Modal Video Moment Localization. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413840

## 1 INTRODUCTION

In recent years, there has been great progress in the field of video understanding [34]. For example, the video retrieval [35] aims to retrieve videos from a set of collections to match the given language query. However, in some scenarios, such as video preview, a more related video moment instead of a whole video is required. As revealed in Figure 1, one may have interest in the moment, "A woman walks to a refrigerator and takes out some vegetables", which is only a clip of the whole video. As such, localizing video moment of interest within a video given a natural language as the query is one step further with more challenging, as compared to simply retrieving an entire video.

In fact, great progress has been made on the video moment localization (or retrieval) in the research community. Specifically, the task of video moment localization is initially proposed in [1, 9], which jointly models language query and video moment to output alignment scores and boundary regression results for candidate video moments. Further efforts have been made by employing attention mechanism [15, 18, 19] or considering the interaction between the visual and textual content [36, 42]. Pioneer methods adopt the strategy of sliding window over the entire video to form and rank all possible video moment-sentence pairs, which are generally time-consuming due to the numerous candidates. Toward this end, reinforcement learning-based methods [10, 32] formulate the video moment localization task as a sequential decision making process. Unfortunately, these work localizes the boundary of a desired video moment (i.e., start and end points) without considering the spatial information. As revealed in Figure 1, only a small fraction of the whole image frames (i.e., the woman and the refrigerator) are tightly related to the query sentence, while other parts are less relevant and redundant. Therefore, if a video moment localization algorithm localizes a video moment without considering the spatial information, noises may be involved by the less relevant part of the frames and further hinder the performance.

Despite its value and significance, the video moment localization has not been well addressed due to the following challenges: 1)

[*] Corresponding author.

*MM '20, October 12–16, 2020, Seattle, WA, USA*

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413840

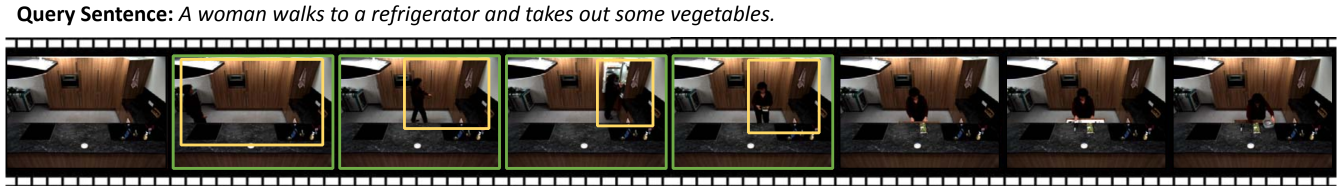

Query Sentence: A woman walks to a refrigerator and takes out some vegetables.

**Figure 1: An example of the video moment localization where the green boxes represent the temporal boundary localization and the yellow boxes represent the spatial scene tracking.**

Unlike the conventional video moment retrieval task that employs a static sliding window strategy to generate and rank all possible video moment-sentence pairs, the video moment localization is a dynamic adjustment process, indicated by the variable-length video moment. In light of this, how to dynamically adjust the boundary of a localized video moment is a non-trivial task. 2) As the information carrier of videos, image frames contain abundant information, including relevant and irrelevant semantics. The simplest way to discover the information embedded in an image is to regard it as a whole. However, this will lead to serious information redundancy because there exist some irrelevant semantics with respect to the query sentence. Therefore, how to discover valuable information and discard useless information embedded in image frames is of great interest. 3) The temporal and spatial clues hidden in the video are equally important to the video moment localization task. Along this line, both temporal boundary localization and spatial scene tracking are crucial for achieving the video moment localization task. Hence, how to jointly optimize these two subtasks in a unified framework is worth to explore.

To address aforementioned challenges, we contribute a joint reinforcement learning framework, namely STRONG (short for "**S**patio-**T**emporal **R**einf**O**rcement learni**NG**"), to investigate the video moment localization task comprehensively. The general framework of STRONG is illustrated in Figure 2. First of all, we employ a temporal-level reinforcement learning to learn the boundary of a desired video moment. Thereafter, a spatial-level reinforcement learning is proposed to track the scene on consecutive image frames. The purpose of this reinforcement learning is to filter out irrelevant information with respect to the given language query. Ultimately, the temporal- and spatial-level reinforcement learning are jointly optimized with an alternative optimization strategy. Along this manner, the tasks of temporal boundary localization and spatial scene tracking are mutually reinforced. By conducting experiments on two real-world datasets, we have illustrated the superiority of our proposed framework on both overall performance comparison and micro-scope studies.

The main contributions of this work are three-fold:

- To the best of our knowledge, this is the first work that attempts to solve the cross-modal video moment localization problem by jointly considering temporal boundary localization and spatial scene tracking.
- We develop a novel solution to improve the performance of video moment localization, which employs an alternative optimization strategy to jointly optimize the temporal- and spatial-level reinforcement learning.
- Extensive experiments are conducted on two datasets, which demonstrate the rationality and effectiveness of our method. Meanwhile, we have released the dataset and implementation to facilitate the research community[1].

## 2 RELATED WORK

### 2.1 Video Moment Retrieval

Video moment retrieval has attracted great attention in the research community, aiming to retrieve relevant video clips from a set of pre-segmented video moments given a textual sentence as the query. Finding a right video moment that matches the query sentence is inherently a difficult and significant task, which is initially proposed by [1, 9]. Benefiting from the merit of neural attention network [3–5], the methods of attentive cross-modal retrieval network [18], cross-modal temporal moment localization [19], and spatial and language-temporal attention model [15] are introduced. Thereafter, by jointly learning the embedding space of visual and textual content, the dependency between vision and language are well learnt to boost the language-video moment retrieval performance [6, 36, 41, 42]. Besides, the techniques of graph structure [40], recurrent neural network [7], and weakly supervised learning [13, 22] are employed to solve the video moment retrieval problem.

Among the aforemention work, none of them explore the temporal boundary localization and the spatial scene tracking simultaneously, which are crucial clues to understand the video moment retrieval process. Temporal boundary localization and spatial scene tracking provide different but correlating and complementary information, and would be thoroughly investigated in our solution.

### 2.2 Reinforcement Learning for Localization

Reinforcement learning is an advanced technique of machine learning, concerning how software agents ought to take actions in an environment so as to maximize the notion of cumulative reward. By utilizing the merit of interactively learning from the environment, the reinforcement learning has been widely applied to text generation [14, 21], person re-identification [20, 25], adaptive video streaming [29, 43], and item recommendation [8, 33, 38].

In fact, the reinforcement learning has been investigated in the localization for visual content. Caicedo et al. [2] presented an active detection model to localize objects in scenes, where a localization agent is trained using deep reinforcement learning. By jointly considering visual and textual content, Shah et al. [30] proposed to localize a person by using deep reinforcement learning. For video content, Yun et al. [39] proposed a novel visual tracking method, which is controlled by sequentially pursuing actions learned by deep reinforcement learning. Thereafter, Yeung et al. [37] introduced a fully end-to-end approach for action detection in videos that learns

[1] https://github.com/yawenzeng/STRONG

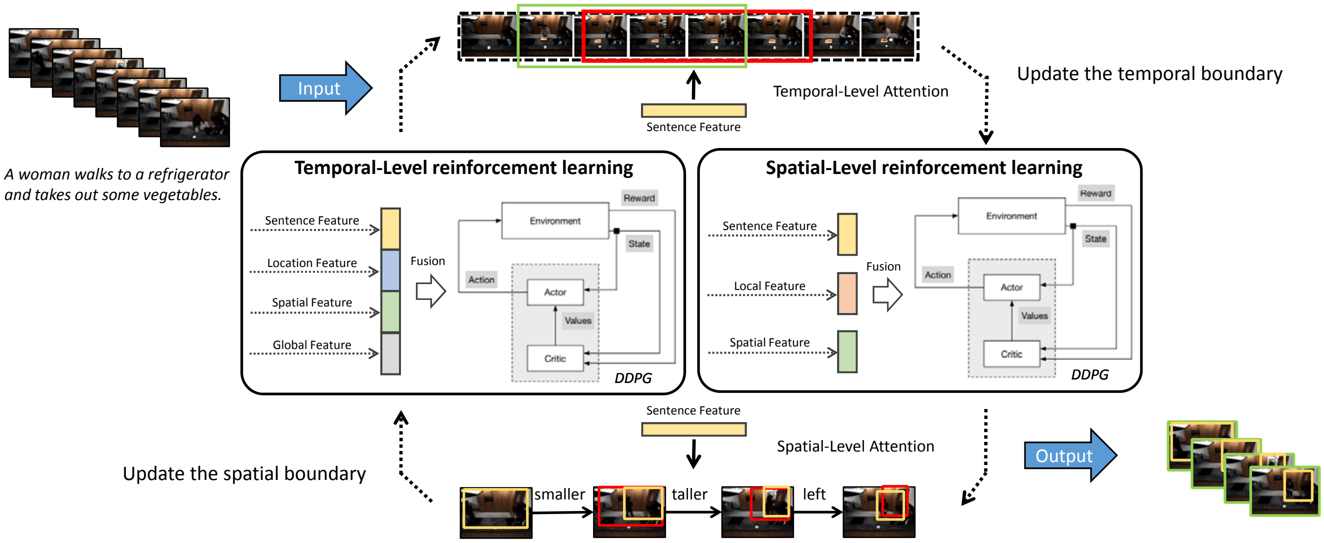


**Figure 2: The graphical representation of our proposed STRONG framework. The input is an untrimmed video and its query sentence, while the output is the video with both localized temporal boundary and localized spatial scene. The green boxes represent localized image frames, the yellow boxes indicate currently localized spacial scene, and the red boxes denote localized image frames or spacial scene on previous time step.**

to directly predict the temporal bounds of actions. In addition, the reinforcement learning has also been investigated for video moment localization. He et al. [10] formulated the problem of temporally grounding natural language descriptions in videos as a sequential decision making process, and it fits naturally into the reinforcement learning paradigm. Wang et al. [32] proposed a recurrent neural network-based reinforcement learning model for language driven temporal activity localization.

Inspired by these pioneering efforts, the intention of our STRONG framework is to take full advantage of the merits of both temporal boundary localization and spatial scene tracking by utilizing the temporal- and the spatial-level reinforcement learning, respectively.

# 3 METHOD

As illustrated in Figure 2, our proposed framework consists of two components: the temporal- and spatial-level reinforcement learning. These two parts are mutually reinforced with our proposed alternative optimization scheme. Specifically, we first formulate the video moment localization in Section 3.1. Thereafter, we introduce the two key ingredients of our proposed model, namely, the temporal-level reinforcement learning in Section 3.2 and the spatial-level reinforcement learning in Section 3.3. Ultimately, the alternative optimization method and training details are illustrated in Section 3.4.

## 3.1 Problem Formulation

Given a long untrimmed video $\mathcal{V} = \{v_1, v_2, \cdots, v_n\}$, where $v_i (i = 1, 2, \cdots, n)$ is the $i$-th image frame, as well as a query sentence $s$, the goal is to identify the boundary of desired video moment, namely $\mathbf{l} = [l_{start}, l_{end}]$. In particular, we formulate the model as a spatio-temporal reinforcement learning agent that interacts with the whole video and its consecutive image frames. More concretely, in the temporal-level reinforcement learning, the agent receives the whole video and the query sentence as inputs, and then it takes a sequence of decisions to output the boundary of desired video moment. In the spatial-level reinforcement learning, the agent receives sequence of image frames and the query sentence as inputs, observes a proportion of the frames, and outputs the spatial boundary of consecutive image frames.

## 3.2 Temporal-Level Reinforcement Learning

The implementation of our temporal boundary localization is realized with the help of Deep Deterministic Policy Gradient (DDPG) [17]. The DDPG is an actor-critic method [23], which combines the merits of policy-based and value-based reinforcement learning algorithms. Meanwhile, it extends ideas from Deep Q-Network (DQN) [24] to optimize the policy with respect to the Q-values. The advantages of DDPG are three-fold: deep neural networks for function approximation, the utilization of experience replay, and the implementation of dual target networks.

**State, Action, and Reward.** To obtain richer environmental information and make video and text semantically complement, state $\mathbf{s}_e$ is defined as the combination of the query sentence feature $\mathbf{f}_e$, the location feature $\mathbf{l}$ with its corresponding spatial video feature $\mathbf{f}_o$, and the global video feature $\mathbf{f}_g$. It is formally defined as:

$$\mathbf{s}_e^t = [\mathbf{f}_e, \mathbf{l}^t, \mathbf{f}_o^t, \mathbf{f}_g], \tag{1}$$

where $t$ indicates the time step. Specifically, skip-thought [16] is employed to obtain a 4,800-D vector for $\mathbf{f}_e$. The location feature is defined as $\mathbf{l}^t = [l_{left}^t, l_{right}^t]$. $\mathbf{f}_o^t$ is generated using a spatial pyramid pooling [11] on the localized video moment, which will be illustrated in Eqn. (13) and Eqn. (14). It ensures that the spatial-level reinforcement learning is able to process spatial scene information with different picture sizes. For $\mathbf{f}_g$, ResNet [12] is employed on frames of the video and a temporal-level attention is utilized to get a 2, 048-D vector as follows:

$$e_t^k = softmax(\mathbf{q}_1^T ReLU(\mathbf{W}_1\mathbf{f}_q + \mathbf{W}_2\mathbf{f}_g^k + \mathbf{b}_1)), \tag{2}$$

$$\mathbf{f}_g = \sum_{k=l_{left}^t}^{l_{right}^t} e_t^k \mathbf{f}_g^k, \tag{3}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are weight matrices, $\mathbf{b}_1$ is a biased vector, and $\mathbf{q}_1$ is a weight vector. The temporal-level attention is employed to integrate image frames by endowing dynamic weights for them.

The action space $\mathcal{A}_e$ consists of 7 predefined actions, namely, moving start/end point forward, moving start/end point backward, shifting both start and end points forward/backward, and a stop action. The initial position is set as $\mathbf{l}^0 = [0.25 * n, 0.75 * n]$, where $n$ is the length of image frames in a video. The step size is set as $n/2z_e$, where $z_e$ is a hyper-parameter which determines the number of searching steps to traverse the whole video. In this way, the searching speed is controlled by varying the step size.

The purpose of video moment localization is to locate the boundary as accurately as possible, we hence use $IoU$ for evaluation. $IoU$ is short for "intersection over union" which measures the overlap between the localized boundary and the ground truth. It is formulated as:

$$IoU^t = \frac{min(l_{end}, l_{end}^t) - max(l_{start}, l_{start}^t)}{max(l_{end}, l_{end}^t) - min(l_{start}, l_{start}^t)}. \tag{4}$$

If $IoU^t$ is larger than $IoU^{t-1}$, the current step will be considered as a positive feedback and the reward $r_e$ is positive, otherwise the reward is zero or negative. Meanwhile, to prevent the agent from taking too much actions, we further add a penalty $-\phi * t$, where $\phi$ is the penalty factor ranging from 0 to 1. In summary, the reward is formulated as:

$$r_e^t = \begin{cases} +1 - \phi * t, & IoU^t > IoU^{t-1} \\ -\phi * t, & IoU^t = IoU^{t-1} \\ -1 - \phi * t, & IoU^t < IoU^{t-1} \end{cases}. \tag{5}$$

**Critic.** Critic is used to approximate value function, and evaluate whether the action $a^t$ in the current state $s^t$ is valid or not. The maximum expected reward will be achieved when the optimal action value function $Q(s, a)$ is regarded as approximating the optimal policy $\pi$. The action-value function for policy $\pi$ is formulated as $Q^\pi(s, a) = \mathbb{E}[R|s^t = s, a^t = a]$. Furthermore, this function can be recursively rewritten as $Q^\pi(s, a) = \mathbb{E}_s'[r(s, a) + \gamma\mathbb{E}_{a'\sim\pi}[Q^\pi(s', a')]]$, where $t$ is the time step, $\gamma$ is the discount factor of the $Q$-value. Critic learns the action-value function $Q$ corresponding to the optimal policy by minimizing the loss:

$$y = r_e^t + \gamma maxQ^*(s', a'|w^*), \tag{6}$$

$$L(w) = \mathbb{E}_{s,a,r,s'\sim M}[(Q(s, a|w) - y)^2], \tag{7}$$

where $Q^*$ is a target $Q$ function whose parameters are periodically updated with the most recent $w^*$, which helps to improve the stabilized learning. Meanwhile, $(s, a, r, s')$ will be sampled from the experience replay memory $M$. For a better approximation, the Critic network consists of four fully connected networks. Action $a$ will be added on the second layer, and the final output is a real value.

**Actor.** Actor is a parameterization policy, which performs the action $a = \pi(s;\theta)$ of moving location $\mathbf{l}^t$. The intuitive idea is to directly adjust the parameters $\theta$ of the policy in order to maximize the objective $J(\theta) = \mathbb{E}_{s\sim p^\pi}[Q^\pi(s, a)]$ by taking steps in the direction of $\nabla_\theta J(\theta)$. To obtain a "good" actor, the evaluation criteria at this time should be able to update the parameter $\theta$ toward the direction in which the Q-value is increased. Therefore, under the deterministic policy $\mu$, the action space $\mathcal{A}_e$ needs to be continuous.

$$\nabla_\theta J(\theta) = \mathbb{E}_{s\sim M}[\nabla_\theta \mu_\theta(a|s)\nabla_a Q^\mu(s, a)|_{a=\mu_{\theta(s)}}]. \tag{8}$$

To better output an action, the Actor network consists of three fully connected networks($FC^a$). The function of $softmax$ is employed to map the output to the action space and the probability of the action is obtained.

**Target Network Update.** The target network is updated in each step of the source network. Meanwhile, the update is relatively small, balanced with $\tau$:

$$target_{\{w^*,\theta^*\}} = (1 - \tau) * target_{\{w^*,\theta^*\}} + \tau * source_{\{w,\theta\}}. \tag{9}$$

This formula is consistent with the soft update in DDPG, which is inspired of DQN.

## 3.3 Spatial-Level Reinforcement Learning

The intuitive idea of locating scene in each image frame is utilizing multiple agents to perform the actions independently. However, this will lead to serious time-consumption and instable convergence. To this end, we borrow the idea of visual tracking [39] to sequentially pursue actions. Specifically, the tracking box of current image frame is based on the position of previous image frame's tracking box and then sequentially update them.

**State, Action, and Reward.** Different from the temporal-level reinforcement learning, state $\mathbf{s}_s$ is the combination of the query sentence feature $\mathbf{f}_e$, the local video feature $\mathbf{f}_s$, and the spatial video feature $\mathbf{f}_o$. It is formally defined as:

$$\mathbf{s}_s^k = [\mathbf{f}_e, \mathbf{f}_s, \mathbf{f}_o^k], \tag{10}$$

where $k$ indicates the time step. In particular, $\mathbf{f}_e$ and $\mathbf{f}_o^k$ utilize the same techniques in Section 3.2 to obtain their representations. For $\mathbf{f}_s$, ResNet is employed on frames of the temporal video moment and a mean pooling is further utilized to get a 2, 048-D vector.

The action space $\mathcal{A}_s$ consists of 9 predefined actions, namely, moving left/right/up/down, zoom bigger/smaller/taller/fatter, and a stop action. A tracking box is represented as $\mathbf{b} = [x, y, m, n]$, where $[x, y]$ is the center point of the box, $[m, n]$ is the box size. The initial position is $\mathbf{b}^0 = [0.25 * h, 0.25 * w, 0.75 * h, 0.75 * w]$, where $h$ and $w$ are the height and width of the image frame. The step size is set as $h/2z_s$ in the vertical direction and $w/2z_s$ in the horizontal direction, where $z_s$ is a hyper-parameter which determines the number of searching steps to traverse the whole image frame.

The purpose of spatial-level reinforcement learning is to enhance the performance of temporal-level reinforcement learning by focusing on the scene. To validate the effectiveness of our proposed spatial-level reinforcement learning, we utilize the image feature updated by the spatial-level reinforcement learning to replace its counterpart in the temporal-level reinforcement learning. Specifically, we utilize $Q_w$ and $Q_o$ to denote the $Q$-values obtained from the temporal-level reinforcement learning by employing the

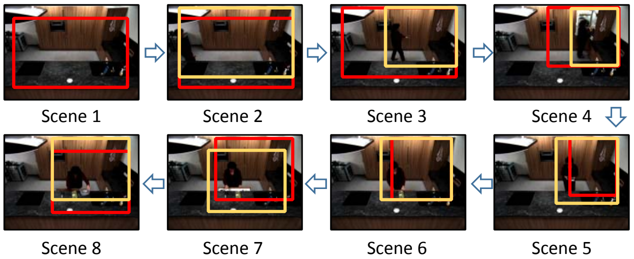


**Figure 3: The illustration of spatial scene tracking. The red box represents the previous time step, while the yellow box represents the current time step. Meanwhile, the movement of yellow box is based on the red box.**

image frame with and without the update by the spatial-level reinforcement learning, respectively. Thereafter, the reward is formulated as:

$$r_s = \begin{cases} +1, & Q_w > Q_o \\ 0, & Q_w = Q_o \\ -1, & otherwise \end{cases}. \tag{11}$$

Through this way, we are able to fine-tune the parameters in our spatial-level reinforcement learning.

**Update of Tracking Box.** The structure of spatial-level reinforcement learning is similar to that of temporal-level reinforcement learning, but the action exploration method is different. For each image frame on the current temporal boundary $\mathbf{l}^t = [l^t_{left}, l^t_{right}]$, it is necessary to use tracking box to learn the object information. The movement of tracking box on each image frame is based on the tracking box of previous image frame:

$$\mathbf{b}_{v_k} = update(\mathbf{b}_{v_{k-1}}, Actor(\mathbf{s}^k_s)), \tag{12}$$

where $\mathbf{b}_{v_k}$ represents the tracking box on image frame $v_k$. Figure 3 visualizes the changing of tracking box. It is worth emphasizing that the image frames as the environment are also changing. But in a short period, the change is relatively small. So the environment is considered to be stable to ensure the stability of convergence.

All image frames of localized video moment are utilized to update the spatial-level reinforcement learning. To deal with the different sizes of tracking boxes on different image frames, we utilize spatial pyramid pooling [11] (SPP) to divide an image into boxes with the number of [1, 4, 16] and then fix it to 2048-D vector $\mathbf{f}^k_o$. Furthermore, to improve the representation of the spatial local feature $\mathbf{f}_o$, a spatial-level attention is employed as follows:

$$e^k_s = softmax(\mathbf{q}^T_2 ReLU(\mathbf{W}_3\mathbf{f}_q + \mathbf{W}_4\mathbf{f}^k_o + \mathbf{b}_2)), \tag{13}$$

$$\mathbf{f}_o = \sum_{k=l^t_{left}}^{l^t_{right}} e^k_s \mathbf{f}^k_o, \tag{14}$$

where $\mathbf{W}_3$ and $\mathbf{W}_4$ are weight matrices, $\mathbf{b}_2$ is a biased vector, and $\mathbf{q}_2$ is a weight vector.

## 3.4 Alternative Optimization

The optimization of STRONG follows an alternative optimization scheme. The temporal- and spatial-level reinforcement learning are jointly optimized and mutual reinforced under this paradigm. Specifically, during the optimization of temporal-level reinforcement learning, the spatial local feature $\mathbf{f}_o$ is obtained from the spatial-level reinforcement learning, which filters out less relevant information and is able to boost the boundary localization performance. If the temporal-level reinforcement learning is optimized independently, $\mathbf{f}_o$ degenerates to $\mathbf{f}_s$, which ignores the crucial spatial information. In addition, as illustrated in Eqn. (11), the optimization of spatial-level reinforcement learning relies on the feedback of temporal-level reinforcement learning. Since the ground truth of the boundary of spatial scene in each image frame is lacked, we resort to the Q-value in the temporal-level reinforcement learning. If the Q-value increases after an action is performed, the reward is positive, otherwise zero or negative.

In summary, the performance of spatial-level reinforcement learning is enhanced by utilizing the Q-value in temporal-level reinforcement learning, while the spatial-level reinforcement learning reinforces the performance of temporal-level reinforcement learning by updating the spatial local feature $\mathbf{f}_o$.

# 4 EXPERIMENTS

In this section, we conducted extensive experiments on two real-world datasets to answer the following four research questions:

**RQ1** How does our proposed STRONG framework perform as compared to other state-of-the-art competitors?

**RQ2** How does the spatial-level reinforcement learning and its scene tracking strategy affect the localization performance?

**RQ3** How do different predefined settings (e.g., the discount factor, the penalty factor, and the tradeoff) affect our framework?

**RQ4** How are the temporal- and spatial-level reinforcement learning mutually reinforced with the alternative optimization?

## 4.1 Experimental Settings

*4.1.1 Datasets.* We experimented with two real-world datasets, one is related to daily activities at home and the other one is related to cooking activities in lab kitchen.

**1. Charades-STA.** Based on the original Charades dataset [31], Charades-STA [9] is proposed for temporal activity localization via language query, which contains 6, 672 videos in total. Since Charades dataset only contains video-level paragraph description, Charades-STA further annotates video moment with language query. Original videos[2] and their corresponding caption annotations[3] are downloaded.

**2. TACoS.** The TACoS dataset[4] is constructed by [26] on the top of MPII-Compositive dataset [28] and contains 127 cooking videos. Each video is affiliated with two kinds of annotations. One is fine-grained activity labels with temporal location (i.e., start and end time). The other kind of annotation is natural language descriptions with temporal locations.

To narrow down the searching space, we further segmented each video on both dataset by utilizing multi-scale sliding windows with the size of [64, 128, 256, 512] frames and 80% overlap on adjacent video moments. Based upon these criteria, in the Charades-STA dataset, we obtained 9, 981 and 2, 572 video moment-query sentence pairs in the training and testing sets, respectively. Meanwhile, in

[2] https://allenai.org/plato/charades
[3] https://github.com/jiyanggao/TALL
[4] http://www.coli.uni-saarland.de/projects/smile/tacos

**Table 1: Overall performance comparison among various methods on Charades-STA and TACoS datasets (Section 4.2).**

| | Charades-STA | | | | | | TACoS | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc@0.1 | Acc@0.3 | Acc@0.5 | Acc@0.7 | Acc@0.9 | mIoU | Acc@0.1 | Acc@0.3 | Acc@0.5 | Acc@0.7 | Acc@0.9 | mIoU |
| CTRL | 81.57% | 65.86% | 37.87% | 7.32% | 0.63% | 37.71% | 79.25% | 51.24% | 34.56% | 9.62% | 1.34% | 36.35% |
| MCN | 79.08% | 66.87% | 40.36% | 6.75% | 0.22% | 38.84% | 81.24% | 48.70% | 32.08% | 10.58% | 0.97% | 36.64% |
| ACRN | 84.65% | 68.78% | 41.15% | 8.96% | 0.81% | 40.03% | 82.24% | 53.89% | 36.24% | 11.42% | 1.55% | 38.05% |
| READ | 87.87% | 71.07% | 42.42% | 14.85% | 1.56% | 43.35% | 84.75% | 58.11% | 39.24% | 14.23% | 1.63% | 41.31% |
| SM-RL | 90.86% | 71.23% | 44.98% | 16.56% | 1.61% | 45.92% | 87.24% | 65.27% | 43.27% | 15.24% | 2.02% | 43.02% |
| STRONG-A | 93.39% | 77.76% | 49.84% | 18.93% | 2.02% | 48.49% | 90.15% | 71.50% | 49.26% | 17.76% | 3.62% | 46.77% |
| **STRONG** | **94.02%** | **78.10%** | **50.14%** | **19.30%** | **2.42%** | **49.07%** | **90.85%** | **72.14%** | **49.73%** | **18.29%** | **3.91%** | **47.18%** |

**Table 2: Performance comparison of STRONG and its variants on Charades-STA and TACoS datasets (Section 4.3).**

| | Charades-STA | | | | | | TACoS | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc@0.1 | Acc@0.3 | Acc@0.5 | Acc@0.7 | Acc@0.9 | mIoU | Acc@0.1 | Acc@0.3 | Acc@0.5 | Acc@0.7 | Acc@0.9 | mIoU |
| TRL | 91.35% | 72.73% | 45.67% | 15.86% | 1.46% | 45.61% | 87.93% | 66.65% | 44.65% | 15.58% | 2.65% | 43.84% |
| TRL+local | 91.62% | 73.32% | 46.83% | 16.51% | 1.47% | 46.31% | 89.42% | 66.72% | 44.93% | 15.63% | 2.86% | 44.81% |
| TRL+dis | 91.87% | 75.22% | 48.32% | 17.05% | 1.53% | 47.01% | 89.85% | 68.54% | 45.08% | 16.01% | 3.21% | 45.32% |
| TRL+gau | 91.95% | 75.31% | 48.99% | 17.30% | 1.54% | 47.34% | 89.94% | 68.98% | 45.40% | 16.04% | 3.24% | 45.77% |
| TRL+fix | 92.80% | 76.62% | 49.68% | 18.15% | 1.82% | 48.10% | 90.43% | 70.83% | 46.90% | 16.56% | 3.43% | 46.40% |
| **STRONG** | **94.02%** | **78.10%** | **50.14%** | **19.30%** | **2.42%** | **49.07%** | **90.85%** | **72.14%** | **49.73%** | **18.29%** | **3.91%** | **47.18%** |

the TACoS dataset, 7, 463 video moment-query sentence pairs are obtained and we randomly selected 80% and 20% of them for training and testing, respectively.

*4.1.2 Evaluation Protocols.* To evaluate the performance of video moment localization, we adopted the localization accuracy proposed by [10] as the evaluation metric. Specifically, given each query sentence, we calculated the *IoU* between the localization result generated by a model and the ground truth. If the *IoU* is larger than *m*, it indicates the localization result is positive. This metric itself is on the query level, so the overall performance is the average result among all queries:

$$Acc@m = \frac{1}{N_q} \sum_{i=1}^{N_q} acc(m, s_i), \tag{15}$$

where $acc(m, s_i)$ indicates whether the *IoU* between the localization result and the ground truth is larger than $m$ given a query $s_i$ (1 for yes and 0 for no), $N_q$ is the total number of queries, and *Acc@m* is the averaged result. In addition, to be consistent with baselines, we also adopted the *mIoU* proposed by [1] as our evaluation metric. *mIoU* indicates mean *IoU*, which measures the average *IoU* over all testing samples.

*4.1.3 Baselines.* To justify the effectiveness of our method, we compared it to the following state-of-the-art baselines.

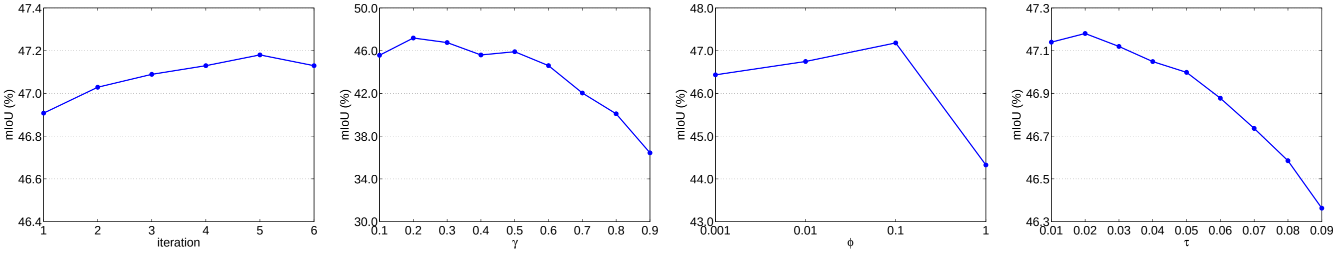
- **CTRL [9].** This is a novel cross-modal temporal regression localizer that jointly models query description and video moments, and outputs alignment scores and action boundary regression results for the moment candidates.
- **MCN [1].** This method is designed for the moment-query retrieval task, which emphasizes the local and global moment features to strengthen the expressive ability.
- **ACRN [18].** This is an attentive cross-modal retrieval network that introduces a memory attention to emphasize the visual features based on the query information and simultaneously incorporates its context moments.
- **READ [10].** This is a reinforcement learning-based method, which formulates the problem of video moment localization as a sequential decision making process. Besides, it combines the supervised learning in a multi-task learning framework.
- **SM-RL [32].** This reinforcement learning-based model selectly observes a sequence of frames and associates the given sentence with video content in a matching-based manner. In addition, semantic concepts of videos are further extracted to enhance the performance.
- **STRONG-A.** This is a variant of our STRONG method by removing both temporal- and spatial-level attention.

To further evaluate the effectiveness of our designed spatial-level reinforcement learning and its scene tracking scheme, we have designed various variants of our method.

- **TRL.** TRL is short for "**T**emporal-level **R**einforcement **L**earning", which removes the spatial-level reinforcement learning of our framework.
- **TRL+local.** TRL+local indicates TRL with local object information. Instead of the scene tracking on each image frame, it utilizes Faster R-CNN [27] to obtain the local object information for each image frame.
- **TRL+dis.** TRL+dis means TRL with discrete spatial scene tracking. Instead of consecutive scene tracking, it employs spatial-level reinforcement learning on each image frame without considering the dependence among consecutive image frames.
- **TRL+gau.** TRL+gua indicates TRL with a gaussian weighting. Instead of scene tracking scheme, it employs a gaussian weighting on consecutive image frames, which is able to focus on the center part and filter out less important information.
- **TRL+fix.** TRL+fix represents TRL with a fixed bounding box for scene tracking. Instead of consecutive scene tracking, it only adjusts the boundary of the first image frame and applies the boundary to following image frames.

*4.1.4 Implementation Details.* We implemented our framework based on the PyTorch framework[5] on a server equipped with a

[5] http://www.pytorch.org



**(a) Iteration** **(b) Discount** **(c) Penalty** **(d) Tradeoff**

**Figure 4: Performance of STRONG w.r.t. the iteration, the discount $\gamma$, the penalty $\tau$, and the tradeoff $\tau$.**

NVIDIA 2080TI-11G GPU. To initialize the embedding layer and hidden layers of neural networks, we randomly set their parameters with a Gaussian distribution (a mean of 0 and a standard deviation of 0.1). Adam optimizer is employed for all gradient-based methods where the mini-batch size and learning rate were set as 128 and 0.01, respectively. For specific hyper-parameters in our framework, the step sizes $z_e$ and $z_s$ are set as 10, the discount $\gamma$ is set as 0.2, the penalty $\phi$ is set as 0.1, and the tradeoff $\tau$ is set as 0.02. In addition, to better train STRONG, we pre-trained it with a simplified version that removes the spatial-level reinforcement learning, i.e., only employing the temporal-level reinforcement learning to perform the video moment localization process. With the pre-trained model as an initialization, we further alternatively trained the spatial and temporal-level reinforcement learning.

## 4.2 Overall Performance Comparison (RQ1)

To demonstrate the effectiveness of our proposed STRONG solution, we compared it to several state-of-the-art approaches: 1) CTRL; 2) MCN; 3) ACRN; 4) READ; 5) SM-RL; and 6) STRONG-A. CTRL, MCN, and ACRN employ sliding window strategy to generate video moment candidates which is able suitable to our video moment localization scenario, while READ, SM-RL, and STRONG-A are reinforcement learning-based algorithms which directly move the boundary to locate their desired video moment.

Experimental results are shown in Table 1. We have the following observations: 1) Our STRONG approach achieves the best performance on both Charades-STA and TACoS datasets, significantly outperforming state-of-the-art baselines. It is mainly because STRONG model considers the influence of temporal boundary localization and spatial scene tracking simultaneously. 2) Reinforcement learning-based algorithms READ, SM-RL, and STRONG-A outperform traditional regression-based algorithms CTRL, MCN, and ACRN by a great margin. Compared to the statistic supervised learning (i.e., regression), the reinforcement learning is a dynamic process which is able to adaptively adjust the interaction between the agent and the environment. That is why the performance of reinforcement learning-based methods is superior to that of regression-based approaches. 3) Although READ and SM-RL belong to the category of reinforcement learning-based methods, the performance of SM-RL is superior to that of READ. This is because except for the global video feature, SM-RL also takes semantic concepts of videos into account. Jointly considering the visual and semantic information, the performance of video moment localization is significantly improved. 4) STRONG consistently beats STRONG-A, which manifests the rationality of our designed temporal- and spatial-level attention.

## 4.3 Ablation Study (RQ2)

The overall performance comparison reveals that STRONG obtains the best results, demonstrating the effectiveness of the integrated solution. To further investigate the importance of spatial-level reinforcement learning and its scene tracking strategy, we performed some ablation studies. In particular, we compared STRONG to its various variants, namely, TRL, TRL+local, TRL+dis, and TRL+fix.

The performance of STRONG and its variants is shown in Table 2. We have the following observations: 1) STRONG consistently and significantly outperforms TRL on both Charades-STA and TACoS datasets. This indicates that the component of spatial-level reinforcement learning is beneficial to the task of video moment localization. 2) STRONG exceeds TRL+local, TRL+dis, and TRL+fix on both datasets. This reveals that the scene track strategy is superior to the strategies of local scene information extraction, discrete spatial scene localization, and fixed bounding box. 3) Regarding the performance of TRL+local, TRL+dis, TRL+gau, and TRL+fix, TRL+fix beats TRL+local, TRL+dis, and TRL+gau by a large margin. TRL+local ignores the interactions among objects, TRL+dis ignores the correlations among consecutive image frames, and TRL+gau is unable to capture the key information in each image frame. Different from other variants, TRL-fix applies a fixed bounding box and the scene on consecutive frames is relatively stable. That is why the performance of TRL-fix is still acceptable.

## 4.4 Convergence and Parameter Tuning (RQ3)

To demonstrate the robustness and effectiveness of our proposed STRONG framework, we investigated the convergence of STRONG and meticulously studied the sensibility of several factors, namely, the discount factor $\gamma$, the penalty factor $\phi$, and the tradeoff $\tau$ with respect to $mIoU$. To save space, we only revealed the experimental results on TACoS, which are consistent with that of Charades-STA.

**Convergence:** We recorded the values of mIoU along with each iteration using the optimal parameter setting. Figure 4a show the experimental results with the increasing number of iterations. We can observe that STRONG is relatively stable and reaches its optimal value around the fifth iteration. This indicates the robustness of our model design.

**Impact of Discount:** The discount factor is the measurement of how far ahead the reinforcement learning should look. To prioritise rewards in the distant future, the value should be closer to 1. A discount factor closer to 0 indicates that only rewards in

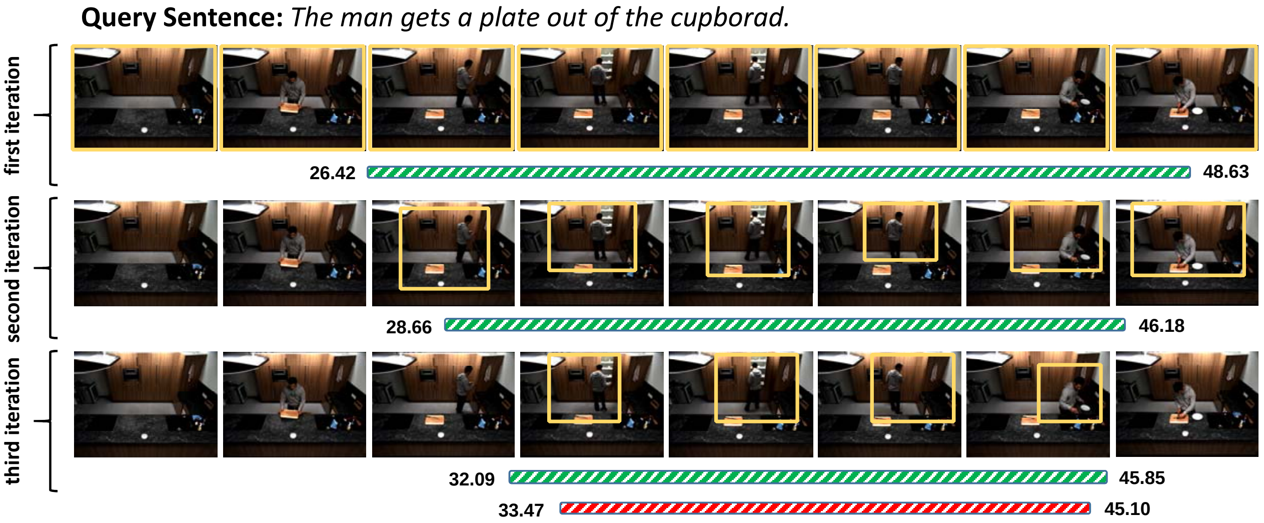


**Figure 5: The visualization of alternative optimization on dual-level reinforcement learning. The green bars represent the localized video moments, the red bar represents the ground truth, and the yellow boxes represent the localized spatial scene.**

the immediate future are being considered, implying a shallow lookahead. The parameter tuning results of $\gamma$ are revealed in Figure 4b. The value of mIoU increases first and then decreases, reaching its optimal value when $\gamma = 0.2$. It indicates that the influence of immediate future should be carefully considered.

**Impact of Penalty:** To balance the number of actions and the accuracy, the penalty factor $\phi$ is utilized to punish the number of steps. The penalty factor closer to 1 indicates less steps are needed, while more steps are needed when it is close to 0. Figure 4c shows the experimental results of mIoU with respect to different values of $\phi$. We can observe that STRONG obtains its optimal performance when $\phi = 0.1$. This shows that $\phi$ can be utilized to strike a desirable trade-off between the accuracy and the computation cost.

**Impact of Tradeoff:** The tradeoff $\tau$ is utilized to balance the impacts of the target network and the source network. If the value of $\tau$ is closer to 1, the update of the target network is highly frequent. Figure 4d exhibits the performance of STRONG w.r.t. the tradeoff parameter $\tau$. mIoU increases first and then decreases along the increasing of $\tau$, reaching its maximum value when $\tau = 0.02$. With the increasing of $\tau$, the update of the target network becomes more frequent, and the difference between the target network and the source network become smaller. Therefore, the overfitting issue becomes more serious. That is why the value of mIoU slightly decreases after it has reached its maximum value.

### 4.5 Visualization of Optimization (RQ4)

To better understand how the temporal- and spatial-level reinforcement learning are mutually reinforced under the alternative optimization paradigm, we exploited a micro-scope study. Specifically, we randomly selected a new video-query sentence pair and cast it into the STRONG framework to observe the dynamic adjustment of the boundary of localized video moment and the boundary of localized spatial scene.

Figure 5 shows the performance of STRONG on both video moment localization and spatial scene tracking in the first three iterations. We have the following observations: 1) In the first iteration, STRONG employs whole image frames to obtain the spatial video feature. Although the boundary of localized video moment is not completely consistent with the ground truth, the localization result is still acceptable. 2) In the second and third iterations, we can clearly observe that the tracking box is gradually changing on consecutive image frames and focus on the scene that is relevant to the query sentence. 3) In the second and third iterations, the temporal boundary localization and spatial scene tracking are alternatively optimized and mutually reinforced. That is why the performance of video moment localization is gradually improved.

## 5 CONCLUSIONS

In this work, we address the video moment localization problem by employing a dual-level reinforcement learning. Specifically, a temporal-level reinforcement learning is proposed to dynamically adjust the boundary of desired video moment given a sentence as the query. Meanwhile, to enhance the localization performance, we further design a spatial-level reinforcement learning to track the scene on consecutive image frames. Lastly, the temporal- and spatial-level reinforcement learning are alternatively optimized and mutually reinforced. Extensive experiments demonstrate the effectiveness of our framework as compared to other competitors on both overall performance comparison and micro-scope studies.

## 6 ACKNOWLEDGEMENTS

This work is supported by the the National Key Research and Development Program of China (No. 2018YFB0804205 and No. 2018YFB0704000), the National Natural Science Foundation of China (No. 61802121 and No. 61772191), the Natural Science Foundation of Hunan Province (No. 2019JJ50057), the Science and Technology Key Projects of Hunan Province (No. 2015TP1004 and No. 2019GK2082), the Science and Technology Project of Changsha City (No. kq1902051), and the Fundamental Research Funds for the Central Universities.

## REFERENCES

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

[2] Juan C Caicedo and Svetlana Lazebnik. 2015. Active object localization with deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2488–2496.

[3] Da Cao, Xiangnan He, Lianhai Miao, Yahui An, Chao Yang, and Richang Hong. 2018. Attentive group recommendation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 645–654.

[4] Da Cao, Xiangnan He, Lianhai Miao, Guangyi Xiao, Hao Chen, and Jiao Xu. 2019. Social-enhanced attentive group recommendation. *IEEE Transactions on Knowledge and Data Engineering* (2019).

[5] Da Cao, Zhiwang Yu, Hanling Zhang, Jiansheng Fang, Liqiang Nie, and Qi Tian. 2019. Video-based cross-modal recipe retrieval. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 1685–1693.

[6] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. 2018. Temporally grounding natural sentence in video. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 162–171.

[7] Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. 2019. Localizing natural language in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI, 8175–8182.

[8] Jingtao Ding, Yuhan Quan, Xiangnan He, Yong Li, and Depeng Jin. 2019. Reinforced negative sampling for recommendation with exposure data. In *Proceedings of the International Joint Conference on Artificial Intelligence*. AAAI, 2230–2236.

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

[10] Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. 2019. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI, 8393–8400.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 9 (2015), 1904–1916.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 770–778.

[13] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2018. Localizing moments in video with temporal language. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 1380–1390.

[14] Qiuyuan Huang, Zhe Gan, Asli Celikyilmaz, Dapeng Wu, Jianfeng Wang, and Xiaodong He. 2019. Hierarchically structured reinforcement learning for topically coherent visual story generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. AAAI, 8465–8472.

[15] Bin Jiang, Xin Huang, Chao Yang, and Junsong Yuan. 2019. Cross-modal video moment retrieval with spatial and language-temporal attention. In *Proceedings of the ACM SIGMM International Conference on Multimedia Retrieval*. ACM, 217–225.

[16] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Proceedings of the Annual Conference on Neural Information Processing Systems*. MIT Press, 3294–3302.

[17] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* (2015).

[18] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive moment retrieval in videos. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 15–24.

[19] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018. Cross-modal moment localization in videos. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 843–851.

[20] Zimo Liu, Jingya Wang, Shaogang Gong, Huchuan Lu, and Dacheng Tao. 2019. Deep reinforcement active learning for human-in-the-loop person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 6122–6131.

[21] Yadan Luo, Zi Huang, Zheng Zhang, Ziwei Wang, Jingjing Li, and Yang Yang. 2019. Curiosity-driven reinforcement learning for diverse visual paragraph generation. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 2341–2350.

[22] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K. Roy-Chowdhury. 2019. Weakly supervised video moment retrieval from text queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 11592–11601.

[23] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *Proceedings of the International Conference on Machine Learning*. ACM, 1928–1937.

[24] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529–533.

[25] Deqiang Ouyang, Jie Shao, Yonghui Zhang, Yang Yang, and Heng Tao Shen. 2018. Video-based person re-identification via self-paced learning and deep reinforcement learning framework. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 1562–1570.

[26] 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.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems*. MIT Press, 91–99.

[28] Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. 2012. Script data for attribute-based recognition of composite activities. In *Proceedings of the European Conference on Computer Vision*. Springer, 144–157.

[29] Lucile Sassatelli, Marco Winckler, Thomas Fisichella, and Ramon Aparicio. 2019. User-adaptive editing for 360 degree video streaming with deep reinforcement learning. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 2208–2210.

[30] Ankit Shah and Tyler Vuong. 2018. Natural language person search using deep reinforcement learning. *arXiv preprint arXiv:1809.00365* (2018).

[31] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proceedings of the European Conference on Computer Vision*. Springer, 510–526.

[32] Weining Wang, Yan Huang, and Liang Wang. 2019. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 334–343.

[33] Xiang Wang, Yaokun Xu, Xiangnan He, Yixin Cao, Meng Wang, and Tat-Seng Chua. 2020. Reinforced Negative Sampling over Knowledge Graph for Recommendation. In *Proceedings of the International Conference on World Wide Web*. ACM.

[34] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. 2019. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 284–293.

[35] Gengshen Wu, Jungong Han, Yuchen Guo, Li Liu, Guiguang Ding, Qiang Ni, and Ling Shao. 2018. Unsupervised deep video hashing via balanced code for large-scale video retrieval. *IEEE Transactions on Image Processing* 28, 4 (2018), 1993–2007.

[36] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. 2019. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI, 9062–9069.

[37] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. 2016. End-to-end learning of action detection from frame glimpses in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2678–2687.

[38] Tong Yu, Yilin Shen, Ruiyi Zhang, Xiangyu Zeng, and Hongxia Jin. 2019. Vision-language recommendation via attribute augmented multimodal reinforcement learning. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 39–47.

[39] Sangdoo Yun, Jongwon Choi, Youngjoon Yoo, Kimin Yun, and Jin Young Choi. 2017. Action-decision networks for visual tracking with deep reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2711–2720.

[40] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1247–1257.

[41] Songyang Zhang, Jinsong Su, and Jiebo Luo. 2019. Exploiting temporal relationships in video moment localization with natural language. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 1230–1238.

[42] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. 2019. Cross-modal interaction networks for query-based moment retrieval in videos. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 655–664.

[43] Yin Zhao, Qi-Wei Shen, Wei Li, Tong Xu, Wei-Hua Niu, and Si-Ran Xu. 2019. Latency aware adaptive video streaming using ensemble deep reinforcement learning. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 2647–2651.