Probability Distribution Based Frame-supervised Language-driven Action Localization

Shuo Yang
shuoyang@bit.edu.cn
Beijing Key Laboratory of Intelligent Information Technology, Beijing Institute of Technology
Guangdong Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University

Zirui Shang
ziruishang@bit.edu.cn
Beijing Key Laboratory of Intelligent Information Technology, Beijing Institute of Technology

Xinxiao Wu*
wuxinxiao@bit.edu.cn
Beijing Key Laboratory of Intelligent Information Technology, Beijing Institute of Technology
Guangdong Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University

ABSTRACT

Frame-supervised language-driven action localization aims to localize action boundaries in untrimmed videos corresponding to the input natural language query, with only a single frame annotation within the target action in training. This task is challenging due to the absence of complete and accurate annotation of action boundaries, hindering visual-language alignment and action boundary prediction. To address this challenge, we propose a novel method that introduces distribution functions to model both the probability of action frame and that of boundary frame. Specifically, we assign each video frame the probability of being the action frame based on the estimated shape parameters of the distribution function, serving as a foreground pseudo-label that guides cross-modal feature learning. Moreover, we model the probabilities of start frame and end frame of the target action using different distribution functions, and then estimate the probability of each action candidate being a positive candidate based on its start and end boundaries, which facilitates predicting action boundaries by exploring more positive terms in training. Experiments on two benchmark datasets demonstrate that our method outperforms existing methods, achieving a gain of more than 10% of $R1@\mu \geq 0.5$ on the challenging TACoS dataset. These results emphasize the significance of generating pseudo labels with appropriate probabilities via distribution functions to address the challenge of frame-supervised language-driven action localization. [1]

CCS CONCEPTS

• **Information systems → Novelty in information retrieval**; **Video search**.

KEYWORDS

language-driven action localization, video moment retrieval, distribution, frame-supervised.

ACM Reference Format:
Shuo Yang, Zirui Shang, and Xinxiao Wu. 2023. Probability Distribution Based Frame-supervised Language-driven Action Localization. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3612512

1 INTRODUCTION

Language-driven action localization has drawn increasing attention in recent years, which aims to locate the action interval in an untrimmed video that is semantically relevant to a language query. This task, also known as video moment retrieval [7, 34, 39, 46] or temporal sentence grounding [16, 18, 44], is a fundamental problem in video understanding and multi-modal information retrieval, which involves not only cross-modal alignment but also action boundary localization. It has been widely applied in various scenarios, such as content-based video search and automatic video editing.

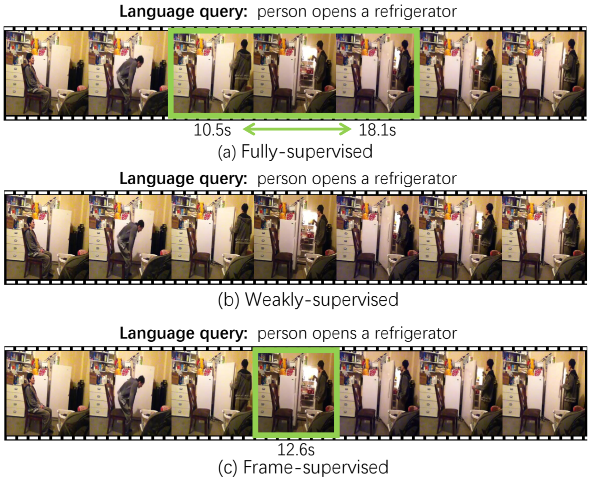
Previous methods [1, 6, 11, 31, 37, 43] have achieved remarkable success in the fully-supervised setting that requires annotating both the start and end timestamps of the target action corresponding to a given language query, as shown in Figure 1 (a). However, the frame annotation paradigm is time-consuming because annotators need to review videos multiple times to accurately identify action boundaries. Consequently, recent methods [13, 25, 33, 38, 50] explore the weakly-supervised setting where only language-video pair annotations are provided, resulting in less annotation burden but lower performance, as shown in Figure 1 (b). The frame-supervised setting, first proposed in [4, 42], uses single-frame annotation within the target action, achieving a good balance between annotation cost and performance, as shown in Figure 1 (c). However, incomplete annotations still hinder visual-language alignment and action boundary prediction.

In this paper, we propose a new method to model the probabilities of action frames and boundary frames by introducing distribution functions. By exploiting the temporal consistency between video frames and properties of probability distribution functions, we extend annotated frames to other frames with different probabilities. This enables the generation of frame-wise pseudo-labels of action frames, which is useful for learning video-language alignment. We

[*] corresponding author
[1] Codes could be found at github

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

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612512


Language query: person opens a refrigerator
10.5s
18.1s
Language query: person opens a refrigerator
Language query: person opens a refrigerator
12.6s

(a) Fully-supervised

(b) Weakly-supervised

(c) Frame-supervised

Figure 1: Illustration of different settings of language-driven action localization. Given an input language query and an input video, (a) the fully-supervised setting provides the starting and ending boundaries of the target action, (b) the weakly-supervised setting provides no additional labels, and (c) the frame-supervised setting gives a frame annotation within the target action.

also model the probabilities of a video frame as the start and end of the target action via different distribution functions. And by multiplying the start and end probabilities, we can obtain the probability of an action candidate being a positive target action segment. By doing so, all action candidates can be treated as positive action candidates with varying probabilities, enabling the exploration of all possible positive action candidates with appropriate probabilities and thus improving the accuracy of boundary estimation.

Specifically, we assign each video frame a probability of being the action frame, which is highest at the annotated frame and gradually decreases to a minimum near the boundary. To model this probability, we estimate the parameters of a specific distribution based on the visual similarity and temporal distance between the video frames and the annotated frame. In particular, we use the various asymmetrical probability curves of Beta distribution to handle the situations in which only few annotated frames are located at the center of the action segments. Using the resulting probability as a soft label for each action frame, we can optimize a binary cross-entropy loss that forces the visual feature to be similar to the language query feature with appropriate loss weights.

Furthermore, we introduce another distribution function to model the probability of each video frame being the start or end boundary of the target action. In our method, the highest probabilities are assigned to the boundaries of an action candidate, or so-called proposal, that is closest to the language query in the feature embedding space, while the annotated frame is less likely to be an action boundary, as illustrated in Figure 4. We combine the probabilities of the starting and ending boundaries to give each action candidate a probability of being the target action segment. This helps us identify more positive action candidates with reasonable probabilities, thereby facilitating the localization of action boundaries.

The main contributions of this paper are as follows:

- We propose a novel method that uses distribution functions, such as the Beta distribution, to generate a probability for each video frame being the action frame, serving as a pseudo-label to enhance the cross-modal feature learning.
- We propose to use different distribution functions to model the probabilities of the start frame and end frame of the target action, so as to explore more positive action candidates during training, thus facilitating the localization of action boundaries.
- Experiments on two benchmark datasets demonstrate that our method outperforms existing methods, especially achieving a gain of more than 10% of $R1@\mu \geq 0.5$ on the challenging TACoS dataset.

2 RELATED WORK

Current language-driven action localization settings can be roughly divided into three types: fully-supervised, weakly-supervised and frame-supervised.

Fully-supervised language-driven action localization requires the annotation of start and end timestamps for each query during training. Existing methods of this setting can be broadly categorized into two groups: proposal-based and proposal-free. In the proposal-based methods, candidate proposals are first generated using sliding windows, proposal generation, or anchor-based methods, and are then ranked based on queries. For instance, CTRL [6], MCN [1], MARN [19], HVSARN [20] and TSTNet [40] generate proposals of varying lengths through sliding windows. 2D-TAN [48], MGPN [32], HLN [5] and VDI [22] generate proposals by using a two-dimensional feature map that model the relationships between segments of varying durations. The proposal-free methods directly predict the start and end boundaries of the target action on sequences of fine-grained video clips. According to the format of moment boundaries, proposal-free methods are categorized into span-based and regression-based methods. VSLNet [47], SLP [14] and D-TSG [17] directly predict the probability of each video snippet or frame being the start and end positions of the target action. TVP [49] and MGSL-Net [15] calculates a time pair and compares it with ground truth for model optimization.

Weakly-supervised language-driven action localization only requires the annotation of pairs of video and query instead of the annotation of start and end times, thus reducing the high annotation cost. Existing weakly-supervised language-driven action localization methods can be broadly classified into two categories: multi-instance learning and reconstruction-based methods. For instance, TGA [25] regards the video and its corresponding query descriptions as positive pairs, while considering the video with other queries and the query with other videos as negative pairs. This method learns video-level visual-text alignment by maximizing the matching scores of positive samples while minimizing the scores of negative samples. SAN [38] introduces a multi-scale Siamese module that progressively narrows the semantic gap between the visual and textual modalities. RTBPN [50] uses a language-aware filter to generate an enhanced video stream and a suppressed video stream, which are used to generate positive proposals and negative proposals for sufficient confrontment, separately.

Frame-supervision is a setting that aims to strike a balance between annotation cost and performance, which has been applied to various computer vision tasks. Bearman et al. [2] introduce the concept of frame supervision for semantic segmentation. Mettes et al. [24] extend the use of frame supervision to spatio-temporal action localization in videos. In recent years, Ma et al. [23] propose a SF-Net model for video temporal action localization by using single-frame supervision. Li et al. [12] develop a temporal action segmentation model that requires only timestamp annotations. More recently, some studies investigated the implementation of single-frame annotation for language-driven action localization. Cui et al. [4] originally introduce the concept of frame supervision for language-driven action localization, which uses the Gaussian distribution to model the probability distribution of foreground frames. Meanwhile, Xu et al. [42] employ a combination of Language Activation Sequence (LAS) and given frame supervision to enhance the model’s ability in language-driven action localization.

3 OUR METHOD

3.1 Problem Definition

Given an untrimmed video and a language query, the task of frame-supervised language-driven action localization aims to localize the target action boundaries (τ_s, τ_e) with an additional frame annotation t_p in the training stage, where $\tau_s \leq t_p \leq \tau_e$, and τ_s and τ_e represent the start and end frames of the action corresponding to the language query, respectively. Note that in the inference stage, the frame annotation is not available.

3.2 Baseline Model

Due to the lack of boundary annotation, we propose a baseline model of frame-supervised language-driven action localization, which follows an multiple-instance learning (MIL) strategy and consists of three components: a video encoder, a language encoder, and a cross-modal interaction module, as shown in Figure 2.

Video Encoder. We first split the given video into a sequence of non-overlap clips with a fixed length (*e.g.*, 16 frames) and extract visual features of each clip by a pre-trained 3D-CNN [3, 35]. Then we uniformly sample T features and project them into d-dimensional representations using a fully-connected (FC) layer. Finally, we encode temporal relationships using a standard Transformer block [36] that consists of multi-head self-attention, layer normalization, residual connection, and feed-forward network. The process of video encoding is represented by

$$\boldsymbol{V'} = Transformer(FC(\text{Uniform-Sample}(\text{3D-CNN}(V)))). \tag{1}$$

Language Encoder. Given an input language query, we first initialize its word features using the GloVe embedding [28] and then project their dimension to d by a fully-connected (FC) layer, followed by a three-layer bi-directional Gated Recurrent Unit (GRU) to learn the relationships of words:

$$\boldsymbol{Q'} = \text{Bi-GRU}(FC(GloVe(Q))). \tag{2}$$

Cross-modal Interaction Module. After encoding the video and language query, we adopt two cross-attention modules for the cross-modal interaction, each by regarding one modality as the query and the other as key and value, followed by a layer normalization, a residual connection and a feed-forward network:

$$\begin{aligned} \boldsymbol{F}_v &= FFN(LN(MSA(FC_Q(\boldsymbol{V'}), FC_K(\boldsymbol{Q'}), FC_V(\boldsymbol{Q'})) + \boldsymbol{V'})), \\ \boldsymbol{F}_q &= FFN(LN(MSA(FC_Q(\boldsymbol{Q'}), FC_K(\boldsymbol{V'}), FC_V(\boldsymbol{V'})) + \boldsymbol{Q'})), \end{aligned} \tag{3}$$

where MSA is the multi-head self-attention module[36]; LN denotes layer normalization; $FC_j(\cdot)$ denotes fully connected layer, $j \in \{Q, K, V\}$; $FFN(\cdot)$ is a feed-forward network.

Training loss. For the learned video features $\boldsymbol{F}_v$, we generate action candidates $P = \{P_1, P_2, \cdots, P_N\}$, also known as proposals, by sliding windows, where N is the total number of action candidates, and $P_i = \text{max-pooling}([\boldsymbol{F}_{v,s_i}, \cdots, \boldsymbol{F}_{v,e_i}]) \in \mathbb{R}^d$ is the i-th action candidate and s_i and e_i are its start and end frame index, respectively; And we also compute the sentence features by mean-pooling on the learned language features: $\boldsymbol{F}_s = \text{mean-pooling}(\boldsymbol{F}_q) \in \mathbb{R}^d$.

In the training stage, we use an intra-video loss and an inter-video loss to learn the video-language alignment. The intra-video loss treats the action candidates containing the annotated frame as positive candidates P^+ and others as negatives P^-. It enforces the similarities between the language query and positive action candidates larger than the similarities between the language query and negative candidates by the InfoNCE [27] loss, given by

$$\mathbb{L}_{intra} = -\frac{1}{M} \sum_{p_i \in P^+} \log \frac{\exp(S(p_i, \boldsymbol{F}_s)/\tau)}{\exp(S(p_i, \boldsymbol{F}_s)/\tau) + \sum\limits_{p_j \in P^-} \exp(S(p_j, \boldsymbol{F}_s)/\tau)}, \tag{4}$$

where $S(\cdot, \cdot)$ is the cosine similarity function; M is the number of positive action candidates; τ is a temperature parameter and set to 0.07 as ViGA [4].

The inter-video loss is also an InfoNCE loss and calculated in a mini-batch, where the positive candidates P^+ in the paired video-sentences are positive terms, and all action candidates of unpaired video-sentences are negative terms. The inter-video enforces the similarities between positive terms larger than the similarities between negative terms in a mini-batch, given by

$$\begin{aligned} \mathbb{L}_{inter} &= -\tfrac{1}{M \times B} \sum_{b=0}^{B} \sum_{p_{b,i} \in P_b^+} \log \frac{\exp(S(p_{b,i}, \boldsymbol{F}_b^s)/\tau)}{\exp(S(p_{b,i}, \boldsymbol{F}_b^s)/\tau) + \mathcal{N}}, \\ \mathcal{N} &= \sum_{j \neq b} \Big(\exp(S(p_{b,i}, \boldsymbol{F}_j^s)/\tau) + \exp(S(p_j, \boldsymbol{F}_b^s)/\tau) \Big), \end{aligned} \tag{5}$$

where $p_{b,i} \in P_b^+$ is the i-th positive candidates of b-th video in a mini-batch; M is the size of P_b^+; B is batch size; $\mathcal{N}$ denotes the negative terms of none paired video-sentence in the mini-batch.

3.3 Pseudo-label of Action Frame

The baseline model uses frame annotations to distinguish positive and negative action candidates, which ignores the temporal coherence of videos. Indeed, the annotated frames play a pivotal role in the frame-supervised language-driven action localization. A common intuition is that the frames adjacent to the annotated frame are more likely to be action frames, while the frames far from the annotated frame are less likely to be action frames. However, how these possibilities change remains an open problem.

In this study, we propose to use distribution functions to model the probability changes, taking into account temporal distance and visual similarity between video frames. The resulting probabilities

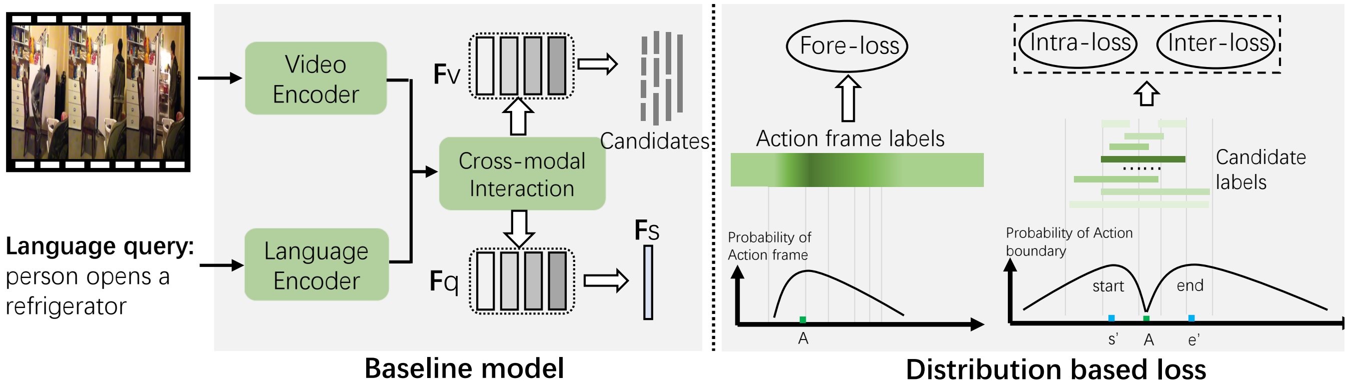
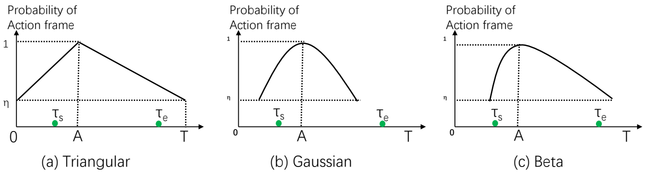


Figure 2: Overview of the proposed method. In the distribution based loss, A denotes the annotated frame; s' and e' are the boundaries of the candidate with maximum similarity to the language query; the curves represent the probabilities by the estimated distributions, by which the pseudo labels of action frames and action candidates are generated (dark color means high probability).



(a) Triangular (b) Gaussian (c) Beta

Figure 3: Probabilities of target action frames by different distributions: (a) Triangular Distribution, (b) Gaussian Distribution, and (c) Beta Distribution. τ_s and τ_e denote the ground-truth boundary of target action; A is the annotated frame; T is the video length; η is the minimum probability.

can be viewed as pseudo-labels for action frames, which extends the annotated frame to its neighbors with different probabilities. This extension provides valuable guidance for learning cross-modal alignment, thereby improving the accuracy of action localization. By leveraging distribution functions in this way, we aim to improve the performance of action localization in videos. Here we explore three distributions to model the probability of frames being the action frames.

Triangular distribution. We start with a simple distribution, the Triangular distribution, which models the probability changes based only on the temporal distance between frames. In this distribution, the probability decreases linearly with the increasing distance from the annotated frame. As shown in Figure 3(a), we assume that the minimum probability of vidoe frame is η, *i.e.*, $\eta = \frac{1}{T}$, where T is the length of video, and the maximum probability of the annotated frame A is 1, the probability of the frame x being the action frame is calculated by

$$P_t^f(x) = \begin{cases} \frac{1}{D}(\frac{x(1-\eta)}{A} + \eta), & 0 \le x \le A \\ \frac{1}{D}(\frac{(T-x)(1-\eta)}{(T-A)} + \eta), & A < x \le T \end{cases} \tag{6}$$

where the factor $\frac{1}{D}$ is used as a normalization factor to ensure that the sum of all probabilities is equal to 1. However, in cases where we want the maximum probability to be 1, we can discard this normalization factor by setting $D = 1$ without affecting the shape of the distribution.

Gaussian distribution. The Gaussian distribution can also model the probability changes of action frames. This more sophisticated distribution assumes that the probability changes follow a bell curve, with the highest probability occurring at the annotated frame and gradually decreasing as the distance from the annotated frame increases, as illustrated in Figure 3(b). The mean μ of the Gaussian distribution is set to the annotated frame A, and we estimate its variance σ^2 by considering both the temporal distance and visual similarity between frames. The temporal distance between the frame t and the annotated frame A is $D(t, A) = \frac{|(t-A)|}{T}$. The visual similarity is calculated by $V(t, A) = 0.5 \times cosine\_similarity(\boldsymbol{F}_{v_t}, \boldsymbol{F}_{v_A}) + 0.5$, and finally, the similarity is a weighted sum of temporal distance and visual similarity:

$$SIM(t, A) = \lambda_1 \cdot V(t, A) + \lambda_2 \cdot (1 - D(t, A)), \tag{7}$$

where λ_1 and λ_2 are hyper-parameters. The standard deviation σ can be estimated by $\sigma = (C_l + C_r)/2$, where C_l and C_r denote the number of frames on the left side ($t < A$) and right side ($t > A$) of the annotated frame, respectively. These frames satisfy the condition $SIM(t, A) \ge \theta$, where θ is a hyper-parameter set to 0.9 times the maximum value of $SIM(t, A)$. With μ and σ, the probability of the video frames x being the action frame is calculated by

$$P_g^f(x) = \exp(-\frac{(x-\mu)^2}{2\sigma^2}). \tag{8}$$

Since the annotated frame is not always located at the center of the action, the probability changes of the frames on either side of the annotated frame may not be symmetrical. However, the Gaussian distribution assumes that the changes are symmetrical, which limits its accuracy in certain situations.

Beta distribution. To overcome the limitation of the Gaussian distribution, we introduce the Beta distribution. As shown in Figure 3(c), the diverse curve and asymmetric probability of the Beta distribution enables modeling different relative positions of the annotated frame within an action. The probability density function

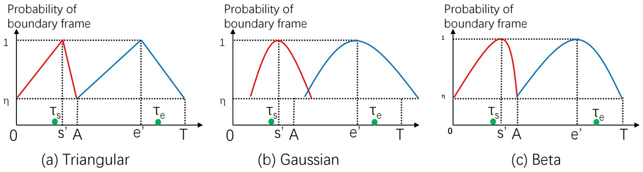


Figure 4: Probabilities of start frames (red curves) and end frames (blue curves) of the target action by the (a) Triangular Distribution, (b) Gaussian Distribution, and (c) Beta Distribution. s' and e' denote the start and end boundaries of the selected action candidate, whose similarity with language query is the largest;

of the Beta distribution is an exponential function of the variable x and its reflection $(1-x)$ as follows:

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} \quad \text{for } 0 \le x \le 1, \tag{9}$$

where α and β are the shape parameters, and $\mathrm{B}(\alpha,\beta)$ is the beta function serving as a normalization factor to ensure the total probability is 1. The mean μ and variance σ^2 are calculated by

$$\mu = \frac{\alpha}{\alpha+\beta}, \qquad \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{10}$$

As the work [41] says, it is hard to directly estimate the parameters α and β. Thus, we estimate the parameters of Beta Distribution by its mean and variance. To do so, we again calculate the similarity between the video frame and the annotated frame by Eq.(7), and set the similarity value larger than the threshold θ to 1 and others to minimum probability η, same as Triangular distribution, we set $\eta = \frac{1}{T}$, where T is the length of video:

$$SIM(x,A)' = \begin{cases} 1, & SIM(x,A) \ge \theta \\ \eta, & others \end{cases} \tag{11}$$

Then the mean $\bar{\mu}$ and variance $\bar{\sigma}^2$ are calculated by

$$\begin{aligned} \bar{\mu} &= \frac{1}{T}\sum_{x=1}^{T} SIM(x,A)' \cdot \frac{x}{T}, \\ \bar{\sigma}^2 &= \frac{1}{T}\sum_{x=1}^{T} SIM(x,A)' \cdot (\frac{x}{T}-\mu)^2. \end{aligned} \tag{12}$$

Finally, the parameters of the Beta distribution are derived as

$$\alpha = \bar{\mu}\left(\frac{\bar{\mu}(1-\bar{\mu})}{\bar{\sigma}^2}-1\right), \qquad \beta = (1-\bar{\mu})\left(\frac{\bar{\mu}(1-\bar{\mu})}{\bar{\sigma}^2}-1\right) \tag{13}$$

Therefore, the probability of the video frames x being the action frame is calculated by

$$P_b^f(x) = \frac{1}{\mathrm{B}(\alpha,\beta)}\left(\frac{x}{T}\right)^{\alpha-1}\left(1-\frac{x}{T}\right)^{\beta-1}. \tag{14}$$

In cases where we want the maximum probability to be 1, we replace the normalization factor $\mathrm{B}(\alpha,\beta)$ with a min-max normalization. In general, the Beta distribution is more complex to estimate than other distributions, which may limit its applications in some scenarios. Nonetheless, this distribution is a valuable tool for modeling the probability of action frames.

Foreground Loss. With the probabilities of video frames being the action frames, we introduce a foreground loss to enforce the embedding of the relevant video frame close to the language query, which helps to learn cross-modal alignment, given by

$$\mathbb{L}_{fore} = BCE(cosine\_similarity(\boldsymbol{F}_q, \boldsymbol{F}_v)/\tau, P_d^f), \tag{15}$$

where $\boldsymbol{F}_q$ and $\boldsymbol{F}_v$ are the language features and video features, respectively; BCE is the binary cross entropy loss; τ is the temperature parameter; P_d^f with $d \in \{t,g,b\}$ are the probabilities of action frames computed by the Triangular, Gaussian, or Beta distributions.

3.4 Pseudo-label of Boundary Frame

In the baseline model of frame-supervised language-driven action localization, all action candidates containing the annotated frame are treated as positive, and others are negative, which may introduce significant noise. As such, we propose to use distribution functions to model the probability of each video frame being a potential starting or ending boundary of the target action moment. By modeling the temporal boundaries in this manner, each action candidate is assigned a probability of being the target action moment by multiplying the probabilities of its boundaries.

Given that the annotated frame is inclined to be within the target action, it is expected to have a minimum probability as a boundary. And for the positive action candidates P^+ that include the annotated frame, we designate the one with the maximum cosine similarity to the language query as the pseudo target action segment, whose staring and ending boundaries are denoted as s' and e' in Figure 4, and assign it a maximum probability of 1. We formulate the starting boundary probability within the interval $[0,A]$ and the ending boundary probability within the interval $[A,T]$. As discussed in Section 3.3, the starting probability P_d^s and the ending probability P_d^e are modeled using distribution functions, and details are omitted here. Consequently, the probability of an action candidate being positive (i.e., belonging to the target action) is computed by taking the multiplication of its boundary probabilities,given by

$$P_d^p(i) = P_d^p(x_1,x_2) = P_d^s(x_1) \times P_d^e(x_2), \quad d \in \{t,g,b\}, \tag{16}$$

where x_1 and x_2 are the starting and ending frame indexes of the i-th action candidate, and d denotes the index of distributions.

As we assign each action candidate a probability by Eq.(16), we update the loss function in the baseline model. Specifically, we use the $P_d^p(i)$ as the loss weight, and accordingly the intra-video and inter-video losses are re-written as

$$\mathbb{L}_{intra}' = -\frac{1}{M}\sum_{p_i \in P^+} P_d^p(i) \cdot \log \frac{\exp(S(p_i,\boldsymbol{F}_s)/\tau)}{\exp(\frac{S(p_i,\boldsymbol{F}_s)}{\tau}) + \sum_{p_j \in P^-}\exp(\frac{S(p_j,\boldsymbol{F}_s)}{\tau})}, \tag{17}$$

$$\mathbb{L}_{inter}' = -\frac{1}{M \cdot B}\sum_{b=0}^{B}\sum_{p_{b,i} \in P_b^+} P_d^p(i) \cdot \log \frac{\exp(S(p_{b,i},\boldsymbol{F}_b^s)/\tau)}{\exp(S(p_{b,i},\boldsymbol{F}_b^s)/\tau) + \mathcal{N}}. \tag{18}$$

3.5 Inference

In the inference stage, we first compute the cosine similarity between the video features and language features and then filter out the action candidates that do not contain the top-k frames based on their similarity scores. After filtering, we rank all the remaining action candidates to determine the most likely one. This method allows us to efficiently identify the action in the video corresponding to the language query by considering both frame and segment features.

4 EXPERIMENTS

4.1 Datasets

To evaluate the proposed method, we conduct experiments on two benchmark datasets, including the TACoS and the Charades-STA datasets.

The TACoS dataset is built on the MPII Cooking Compositive dataset [29], which consists of 127 videos with an average length of 4.79 minutes. There are around 148 annotated segments per video. The dataset contains 18,818 samples, including 10,146 for training, 4,589 for validation, and 4,083 for testing. This dataset is more challenging due to the long videos and short action segments.

The Charades-STA dataset is built on the Charades dataset [30] and contains 6,672 daily life videos. The average duration of the videos is 29.76 seconds. There are about 2.4 annotated segments per video, whose average duration is 8.2 seconds. The whole dataset contains 16,128 samples (i.e., pairs of query and action segment), and we follow the standard split of 12,408 and 3,720 samples for training and testing.

4.2 Evaluation Metrics

We adopt two metrics for the performance evaluation: (1) $R@n; IoU \geq \mu$, which denotes the recall of top-n predictions at various thresholds of the temporal Intersection over Union (IoU). It measures the percentage of predictions that have IoU with ground truth larger than the threshold μ; (2) mean averaged IoU (mIoU), which denotes the average IoU over all the test samples. We set $n = 1$ and $\mu \in \{0.3, 0.5, 0.7\}$.

4.3 Implementation Details

We use C3D [35] for the TACoS dataset and I3D [3] for the Charades-STA dataset to extract video features. Adam [10] is adopted for optimization with an initial learning rate of 1e-4 and half decaying on plateau. The intermediate feature dimension d is set to 512, and the head number of multi-head self-attention is set to 8. The hyper-parameters λ_1 and λ_2 in Eq.(7) are set to 0.2 and 1 for Charades-STA, and 0.6 and 0.8 for TACoS. The loss weights for all loss items are set to 1.

4.4 Ablation Studies

We perform in-depth analysis to evaluate each component of our method on the TACoS and Charades-STA datasets.

Effectiveness of different distributions. We perform an ablation study using the Triangular, Gaussian, and Beta distributions to demonstrate the effectiveness of incorporating different distributions into the baseline model. The results on the TACoS and Charades-STA datasets are shown in Table 1 and Table 2, respectively. It is obvious that each of these distributions significantly improves the performance of the baseline model on the TACoS dataset, and similar trends can be seen on the Charades-STA dataset except for the Triangular distribution, where the Triangular distribution gives small probabilities to the frame near the ground-truth due to the short videos. Specifically, on the TACoS dataset, the Triangular distribution improves the $R@1; IoU \geq 0.3$ by 17.59%, the Gaussian distribution improves the $R@1; IoU \geq 0.3$ by 18.82%, and the Beta distribution achieves the greatest improvement with the gain of 19.09% on the $R@1; IoU \geq 0.3$. These results demonstrate that the pseudo-labels with appropriate probabilities provide positive guidance for learning cross-modal alignment and boundary estimation.

Table 1: Ablation studies of different distributions on the TACoS dataset.

Methods	$R@1; IoU \geq \mu$			$mIoU$
	0.3	0.5	0.7	
Baseline	17.05	6.45	1.87	15.47
Ours (Triangular)	34.64	19.02	6.47	22.23
Ours (Gaussian)	35.87	19.47	6.95	22.85
Ours (Beta)	**36.14**	**20.17**	**7.30**	**23.09**

Table 2: Ablation studies of different distributions on the Charades-STA dataset.

Methods	$R@1; IoU \geq \mu$			$mIoU$
	0.3	0.5	0.7	
Baseline	70.4	45.05	20.03	44.30
Ours (Triangular)	66.94	42.63	19.22	42.67
Ours (Gaussian)	71.10	48.15	25.65	46.75
Ours (Beta)	**71.72**	**50.13**	**26.72**	**47.35**

Effectiveness of the Beta distribution. The asymmetric nature of the Beta distribution makes it well-suited to handle the annotated frames that occur near the action boundaries. To evaluate the effectiveness of the Beta distribution, we conduct experiments where the training videos are re-annotated by placing annotated frames at various positions within the video. For example, the annotated frames are at the first ten percent of the action segment, which we denote as 0.1 in Table 3. We re-implement ViGA [4] by re-training the model using the new annotated frames via the open-source codes for comparison with our method. The comparison results are reported in Table 3.

Our results show consistent improvements in $R@1; IoU \geq 0.5$ and $R@1; IoU \geq 0.7$, regardless of whether the annotation is located in the center of the action (represented as 0.5 in Table 3) or near the action boundaries (represented as 0.1 and 0.9 in Table 3). However, we observed fewer improvements when the annotations were near the boundaries, such as at ten (0.1) and ninety (0.9) percentages of the action segment. This can be attributed to the difficulty in estimating the distribution parameters accurately in such situations. We also noticed a slight decrease (less than 1%) in performance on $R@1; IoU \geq 0.3$, which we believe may be due to the neglect of some information when the probabilities for certain video frames are lower. Nonetheless, we are encouraged by the overall effectiveness

Table 3: Ablation studies of different annotation positions P on the Charades dataset. The value format a/b in the table denotes that a is the result of the re-trained ViGA [4] and b is the result of our method.

P	$R@1; IoU \geq \mu$			$mIoU$
	0.3	0.5	0.7	
0.1	65.59/65.05	41.42/43.39	20.13/21.64	42.44/42.80
0.3	69.25/70.13	46.05/**48.60**	20.73/**25.38**	44.19/**46.46**
0.5	**71.88**/71.53	44.97/47.98	21.26/23.28	45.01/45.93
0.7	68.39/68.09	44.41/45.24	19.41/21.34	43.12/43.77
0.9	61.77/61.05	36.29/36.99	15.32/16.75	38.75/38.90

Table 4: Ablation studies of different components on the Charades-STA dataset. "r1i3" is the short of $R@1; IoU \geq 0.3$. "v-s" is the short of visual similarity. "intra" and "inte" represent the intra-video loss and the inter-video loss, respectively.

	v-s	intra	inter	r1i3	r1i5	r1i7	$mIoU$
1	x	x	x	70.4	45.05	20.03	44.30
2	✓	x	x	71.32	47.77	22.5	45.44
3	x	✓	x	70.43	47.58	24.7	46.05
4	x	x	✓	71.34	45.11	20.65	44.85
5	✓	x	✓	70.97	47.93	23.47	45.85
6	x	✓	✓	71.53	47.98	24.73	46.58
7	✓	✓	x	71.13	49.14	25.91	47.06
8	✓	✓	✓	**71.72**	**50.13**	**26.72**	**47.35**

of our method, as demonstrated by the consistent improvements in higher IoU thresholds.

Effectiveness of visual similarity. We estimate the distribution parameters by considering both visual similarity and temporal distance, as shown in Eq.(7). To evaluate the effectiveness of the visual similarity, we set $\lambda_1 = 0$ and remove it from the calculation of similarity between frames. The results are shown in Table 4, where "v-s" represents the visual similarity. Comparing the results in line 1 and 2 where the frame similarity computed by Eq.(7) is directly used as the probability, and comparing the results in line 6 and 8 where the Beta distribution is used for probability calculation, we observe improvements of more than 2.5% and 2% on $R@1; IoU \geq 0.5$, respectively, which clearly demonstrates the importance of the visual similarity.

Effectiveness of intra-video loss. The intra-video loss, defined in Eq.(17), enforces similarities between language queries and positive action candidates more than negative action candidates. It is computed in a single language-video pair and helps to rank all the action candidates, thus improving the accuracy of boundary estimation. To evaluate the effectiveness of the intra-video loss, we remove it for comparison, and the results are shown in Table 4. We observe that compared with the result in line 1 without the intra-video loss, the result in line 3 with the intra-video loss achieves significant improvements of more than 4% on $R@1; IoU \geq 0.7$, and about 2.5% on $R@1; IoU \geq 0.5$. This demonstrates that the intra-video loss effectively improves the localization accuracy, especially in high-precision scenarios. Similar trends of comparing the results in line 5 and line 8, line 2 and line 7, and line 4 and line 6, further verify the effectiveness of the intra-video loss.

Effectiveness of inter-video loss. The inter-video loss, defined in Eq.(18), helps to leverage inter-sample information to learn the diversities of actions and facilitate model training in the early stage. However, it mainly focuses on learning the differences between different action instances rather than the fine-grained details of action boundaries. To evaluate the effectiveness of the inter-video loss, we remove it for comparison, and the results are shown in Table 4. We observe that using the inter-video loss (lines 4, 5 and 8) improves the accuracy by about 1% compared with the results without the inter-video loss (lines 1, 2 and 7). However, this improvement is relatively small compared to that of the intra-video loss, suggesting that the intra-video loss is more effective in improving accuracy and localization performance, especially in terms of high precision, while the inter-video loss plays a complementary role in improving the diversity of learned action representations.

4.5 Comparison with State-of-the-art Methods

We compare the proposed method with several state-of-the-art methods at different levels of supervision, including fully-supervised methods (CTRL [6], 2D-TAN [48], VSLNet [47]), weakly-supervised methods (TGA [25], SCN [13], LoGAN [33], CRM [9]), and frame-supervised methods (ViGA [4], LAS [42]).

The comparison results on the TACoS and Charades-STA datasets are shown in Table 5 and Table 6, respectively. From the results, we have observations as follows: (1) Compared with the frame-supervised methods, *i.e.*, ViGA [4], and LAS [42], our method achieves more than 10% improvements on the $R@1; IoU \geq 0.3$ and $R@1; IoU \geq 0.5$ on the challenging TACoS dataset, and more than 5% improvements on the $R@1; IoU \geq 0.5$ and $R@1; IoU \geq 0.7$ on the Charades-STA dataset, which demonstrates the effectiveness of the proposed distribution-based method on modeling the probabilities of pseudo-labels, especially in more difficult scenarios; (2) Compared with the fully-supervised methods shown in the upper parts of Table 5 and Table 6, our method achieves comparable results, demonstrating the huge potential of the performance of frame-supervised language-driven action localization; (3) Compared with the weakly supervised methods shown in the middle part of Table 6, our method outperforms all other methods in terms of all metrics by a large margin, showing the superiority of our method in the scenario that lacks full annotations. These results suggest that our method achieves satisfying performance on the TACoS and Charades-STA datasets and is a promising direction for language-driven action localization.

4.6 Qualitative Analysis

We show several examples of action localization results of our method and the baseline model on the Charades-STA dataset in Figure 5. From the first three examples in Figure 5 (a), (b), and (c), we observe that the action boundaries predicted by our method are more accurate than the baseline model since the boundary frames participate in training with appropriate probabilities in our method. However, as shown in Figure 5 (d), both our method and the baseline model fail to locate the action boundaries because the video frames are too similar to distinguish, showing a lack of

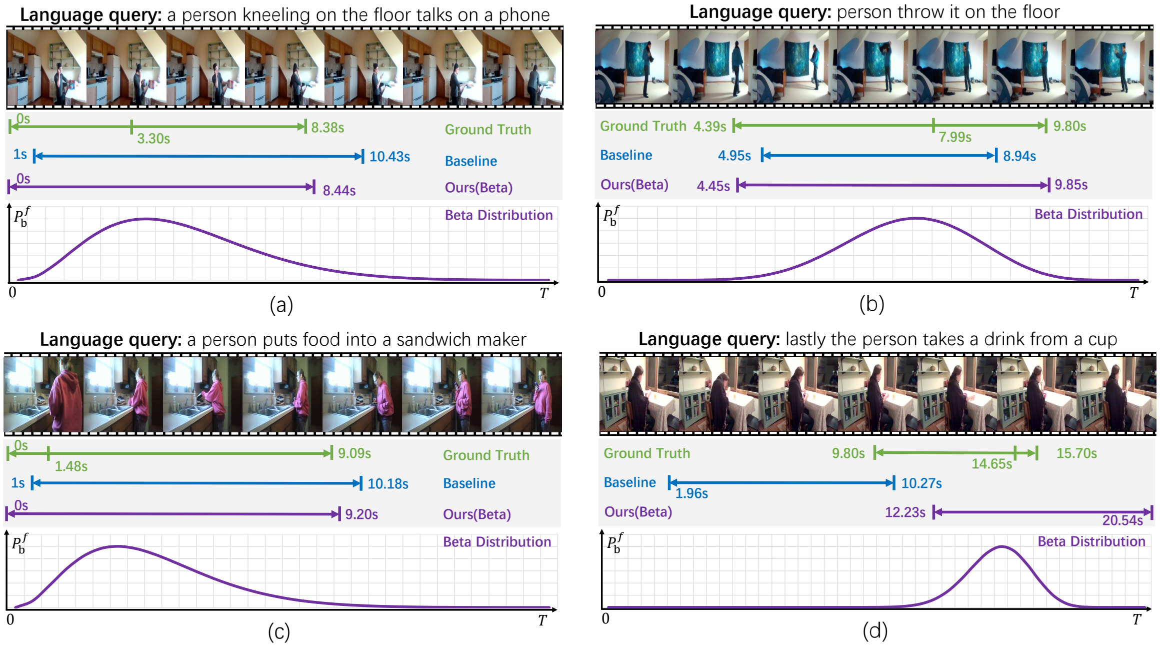


Figure 5: Examples of action localization results. "Baseline" denotes the results from the baseline model; "Ours (Beta)" denotes the results predicted by the model using the Beta distribution; "Beta Distribution" denotes the curve generated by Eq.(14).

Table 5: Comparison with the state-of-the-art methods on the TACoS dataset. Upper part: Fully-supervised methods; Lower part: Frame-supervised methods.

Methods	$R@1; IoU \geq \mu$			$mIoU$
	0.3	0.5	0.7	
CTRL [6]	18.32	13.3	-	-
TripNet [8]	23.95	19.17	-	-
ABLR [45]	19.50	9.40	-	-
DEBUG [21]	23.45	11.72	-	16.03
VSLNet [47]	29.61	24.27	20.03	24.11
2D-TAN [48]	37.29	25.32	-	-
ViGA [4]	19.62	8.85	3.22	15.47
LAS [42]	23.64	10.00	3.35	17.39
Ours (Beta)	**36.14**	**20.17**	**7.30**	**23.09**

Table 6: Comparison with the state-of-the-art methods on the Charades-STA dataset. Upper part: Fully-supervised methods; Middle part: Weakly-supervised methods; Lower part: Frame-supervised methods.

Methods	$R@1; IoU \geq \mu$			$mIoU$
	0.3	0.5	0.7	
CTRL [6]	-	23.63	8.89	-
2D-TAN [48]	-	39.70	23.31	-
LGI [26]	72.96	59.46	35.48	51.38
VSLNet [47]	70.46	54.19	35.22	50.02
TGA [25]	32.14	19.94	8.84	-
SCN [13]	42.96	23.58	9.97	-
LoGAN [33]	51.67	34.68	14.54	-
CRM [9]	53.66	34.76	16.37	-
LAS [42]	60.40	39.22	20.17	39.77
ViGA [4]	71.21	45.05	20.27	44.57
Ours (Beta)	**71.72**	**50.13**	**26.72**	**47.35**

sufficient fine-grained motion understanding. It is worth noting that in all four examples in Figure 5, the estimated Beta distributions are reasonable and provide guidance information according to the annotated frame, thereby facilitating the cross-model alignment and boundary estimation during training.

5 CONCLUSION

We have presented a novel probability distribution based method for frame-supervised language-driven action localization. By using distribution functions to model the probabilities of the action frame, as well as the starting and ending boundaries of the target action, our method is able to provide more accurate guidance in learning cross-modal alignment and boundary estimation to compensate for the lack of supervision, thus successfully improving the accuracy of action localization. Extensive experimental results on two benchmark datasets demonstrate the effectiveness of our method. We believe that the distribution-based framework will be a promising direction for further research in the field of frame-supervised language-driven action localization.

ACKNOWLEDGMENTS

This work was supported in part by the Natural Science Foundation of China (NSFC) under Grant No 62072041.

REFERENCES

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

[2] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. 2016. What's the point: Semantic segmentation with point supervision. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*. Springer, 549–565.

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

[4] Ran Cui, Tianwen Qian, Pai Peng, Elena Daskalaki, Jingjing Chen, Xiaowei Guo, Huyang Sun, and Yu-Gang Jiang. 2022. Video Moment Retrieval from Text Queries via Single Frame Annotation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1033–1043.

[5] Xinpeng Ding, Nannan Wang, Shiwei Zhang, Ziyuan Huang, Xiaomeng Li, Mingqian Tang, Tongliang Liu, and Xinbo Gao. 2022. Exploring language hierarchy for video grounding. *IEEE Transactions on Image Processing* 31 (2022), 4693–4706.

[6] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*. 5267–5275.

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

[8] Meera Hahn, Asim Kadav, James M Rehg, and Hans Peter Graf. 2019. Tripping through time: Efficient localization of activities in videos. *arXiv preprint arXiv:1904.09936* (2019).

[9] Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. 2021. Cross-sentence temporal and semantic relations in video activity localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7199–7208.

[10] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[11] Juncheng Li, Junlin Xie, Long Qian, Linchao Zhu, Siliang Tang, Fei Wu, Yi Yang, Yueting Zhuang, and Xin Eric Wang. 2022. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3032–3041.

[12] Zhe Li, Yazan Abu Farha, and Jurgen Gall. 2021. Temporal action segmentation from timestamp supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8365–8374.

[13] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. 2020. Weakly-supervised video moment retrieval via semantic completion network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11539–11546.

[14] Daizong Liu and Wei Hu. 2022. Skimming, locating, then perusing: A human-like framework for natural language video localization. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4536–4545.

[15] Daizong Liu, Xiaoye Qu, Xing Di, Yu Cheng, Zichuan Xu, and Pan Zhou. 2022. Memory-guided semantic learning network for temporal sentence grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1665–1673.

[16] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. 2021. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11235–11244.

[17] Daizong Liu, Xiaoye Qu, and Wei Hu. 2022. Reducing the vision and language bias for temporal sentence grounding. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4092–4101.

[18] Daizong Liu, Xiaoye Qu, and Pan Zhou. 2021. Progressively Guide to Attend: An Iterative Alignment Framework for Temporal Sentence Grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 9302–9311.

[19] Daizong Liu, Xiaoye Qu, Pan Zhou, and Yang Liu. 2022. Exploring motion and appearance information for temporal sentence grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1674–1682.

[20] Daizong Liu and Pan Zhou. 2023. Jointly visual-and semantic-aware graph memory networks for temporal sentence localization in videos. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1–5.

[21] Chujie Lu, Long Chen, Chilie Tan, Xiaolin Li, and Jun Xiao. 2019. Debug: A dense bottom-up grounding approach for natural language video localization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 5144–5153.

[22] Dezhao Luo, Jiabo Huang, Shaogang Gong, Hailin Jin, and Yang Liu. 2023. Towards Generalisable Video Moment Retrieval: Visual-Dynamic Injection to Image-Text Pre-Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23045–23055.

[23] Fan Ma, Linchao Zhu, Yi Yang, Shengxin Zha, Gourab Kundu, Matt Feiszli, and Zheng Shou. 2020. Sf-net: Single-frame supervision for temporal action localization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 420–437.

[24] Pascal Mettes, Jan C Van Gemert, and Cees GM Snoek. 2016. Spot on: Action localization from pointly-supervised proposals. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*. Springer, 437–453.

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

[26] Jonghwan Mun, Minsu Cho, and Bohyung Han. 2020. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10810–10819.

[27] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[28] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing*. 1532–1543.

[29] Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. 2012. Script data for attribute-based recognition of composite activities. In *European conference on computer vision*. Springer, 144–157.

[30] 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 *European Conference on Computer Vision*. Springer, 510–526.

[31] Mattia Soldan, Mengmeng Xu, Sisi Qu, Jesper Tegner, and Bernard Ghanem. 2021. Vlg-net: Video-language graph matching network for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3224–3234.

[32] Xin Sun, Xuan Wang, Jialin Gao, Qiong Liu, and Xi Zhou. 2022. You Need to Read Again: Multi-granularity Perception Network for Moment Retrieval in Videos. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1022–1032.

[33] Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A Plummer. 2021. Logan: Latent graph co-attention network for weakly-supervised video moment retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2083–2092.

[34] Haoyu Tang, Jihua Zhu, Meng Liu, Zan Gao, and Zhiyong Cheng. 2021. Frame-wise cross-modal matching for video moment retrieval. *IEEE Transactions on Multimedia* 24 (2021), 1338–1349.

[35] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

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

[37] Hao Wang, Zheng-Jun Zha, Liang Li, Dong Liu, and Jiebo Luo. 2021. Structured multi-level interaction network for video moment localization via language query. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7026–7035.

[38] Yunxiao Wang, Meng Liu, Yinwei Wei, Zhiyong Cheng, Yinglong Wang, and Liqiang Nie. 2022. Siamese alignment network for weakly supervised video moment retrieval. *IEEE Transactions on Multimedia* (2022).

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

[40] Zeyu Xiong, Daizong Liu, Pan Zhou, and Jiahao Zhu. 2023. Tracking Objects and Activities with Attention for Temporal Sentence Grounding. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1–5.

[41] Zixuan Xu, Banghuai Li, Ye Yuan, and Anhong Dang. 2020. Beta r-cnn: Looking into pedestrian detection from another perspective. *Advances in Neural Information Processing Systems* 33 (2020), 19953–19963.

[42] Zhe Xu, Kun Wei, Xu Yang, and Cheng Deng. 2022. Point-Supervised Video Temporal Grounding. *IEEE Transactions on Multimedia* (2022).

[43] Shuo Yang and Xinxiao Wu. 2022. Entity-aware and Motion-aware Transformers for Language-driven Action Localization. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, LD Raedt, Ed*. 1552–1558.

[44] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. 2019. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. *Advances in Neural Information Processing Systems* 32 (2019).

[45] Yitian Yuan, Tao Mei, and Wenwu Zhu. 2019. 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. 9159–9166.

[46] Yawen Zeng, Da Cao, Xiaochi Wei, Meng Liu, Zhou Zhao, and Zheng Qin. 2021. Multi-modal relational graph for cross-modal video moment retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2215–2224.

[47] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2020. Span-based Localizing Network for Natural Language Video Localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 6543–6554.

[48] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 12870–12877.

[49] Yimeng Zhang, Xin Chen, Jinghan Jia, Sijia Liu, and Ke Ding. 2023. Text-visual prompting for efficient 2d temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14794–14804.

[50] Zhu Zhang, Zhijie Lin, Zhou Zhao, Jieming Zhu, and Xiuqiang He. 2020. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4098–4106.