This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

D3G: Exploring Gaussian Prior for Temporal Sentence Grounding with Glance Annotation

Hanjun Li[1], Xiujun Shu[1], Sunan He[2], Ruizhi Qiao[1], Wei Wen[1], Taian Guo[1], Bei Gan[1], Xing Sun[1]*
[1]Youtu Lab, Tencent [2]Hong Kong University of Science and Technology
{hanjunli, xiujunshu, ruizhiqiao, jawnrwen, taianguo, stylegan, winfredsun}@tencent.com
sunan.he@connect.ust.hk

Abstract

*Temporal sentence grounding (TSG) aims to locate a specific moment from an untrimmed video with a given natural language query. Recently, weakly supervised methods still have a large performance gap compared to fully supervised ones, while the latter requires laborious timestamp annotations. In this study, we aim to reduce the annotation cost yet keep competitive performance for TSG task compared to fully supervised ones. To achieve this goal, we investigate a recently proposed glance-supervised temporal sentence grounding task, which requires only single frame annotation (referred to as glance annotation) for each query. Under this setup, we propose a **D**ynamic **G**aussian prior based **G**rounding framework with **G**lance annotation (D3G), which consists of a Semantic Alignment Group Contrastive Learning module (SA-GCL) and a Dynamic Gaussian prior Adjustment module (DGA). Specifically, SA-GCL samples reliable positive moments from a 2D temporal map via jointly leveraging Gaussian prior and semantic consistency, which contributes to aligning the positive sentence-moment pairs in the joint embedding space. Moreover, to alleviate the annotation bias resulting from glance annotation and model complex queries consisting of multiple events, we propose the DGA module, which adjusts the distribution dynamically to approximate the ground truth of target moments. Extensive experiments on three challenging benchmarks verify the effectiveness of the proposed D3G. It outperforms the state-of-the-art weakly supervised methods by a large margin and narrows the performance gap compared to fully supervised methods. Code is available at https://github.com/solicucu/D3G.*

1. Introduction

Temporal sentence grounding is a fundamental problem in computer vision and receives an increasing attention in recent years. Given the query sentence and an untrimmed video, the goal of TSG is to localize the start and end timestamps of specific moment that semantically corresponds to the query. In recent years, full supervised temporal sentence grounding (FS-TSG) has achieved tremendous achievements [9, 1, 41, 43, 34, 29, 33, 42]. However, obtaining accurate timestamps for each sentence is labor-intensive and subjective, which prevents it from scaling to large-scale video-sentence pairs and practical applications.

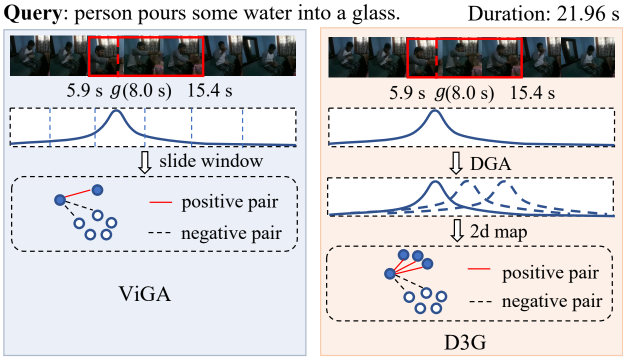


Figure 1. Illustration of glance annotation g (red dashed line) and simple comparison between ViGA and D3G. The red rectangle indicates the boundary of target moment.

Weakly supervised temporal sentence grounding (WS-TSG), which requires only the video and query pairs, receives an increasing attention recently. Although great advances [19, 32, 12, 45, 43, 44] have been achieved in recent years, there still remains a huge performance gap between WS-TSG and FS-TSG. WS-TSG suffers from severe localization issues due to the large discrepancy between video-level annotations and clip-level task.

Recently, Cui *et al*. [6] propose a new annotating paradigm called glance annotation for TSG, requiring the timestamp of only random single frame within the temporal boundary of the target moment. It is noted that such annotation only increases trivial annotating cost compared to WS-TSG. Figure 1 illustrates the details of glance annotation. With glance annotation, Cui *et al*. propose the

*Corresponding author

ViGA based on contrastive learning. ViGA first cuts the input video into clips of fixed length, which are assigned with Gaussian weights generated according to the glance annotation, and contrasts clips with queries. There are two obvious disadvantages in this way. First, moments of interest usually have various durations. Therefore, these clips cannot cover a wide range of target moments, which inevitably aligns the sentence with incomplete moment and obtains sub-optimal performance. Second, ViGA utilizes a fixed scale Gaussian distribution centered at the glance frame to describe the span of each annotated moment. However, the glance annotations are not guaranteed at the center of target moments, which results in annotation bias as shown in Figure 2. Besides, since some complex query sentences consist of multiple events, a single Gaussian distribution is hard to cover all events at the same time as shown in Figure 3. To address the aforementioned defects and fully unleash the potential of Gaussian prior knowledge with the low-cost glance annotation, we propose a **D**ynamic **G**aussian prior based **G**rounding framework with **G**lance annotation (D3G) as shown in Figure 4.

We first generate a wide range of candidate moments following 2D-TAN [43]. Afterwards, we propose a Semantic Alignment Group Contrastive Learning module (SA-GCL) to align the positive sentence-moment pairs in the joint embedding space. Specifically, for each query sentence, we sample a group of positive moments according to calibrated Gausssian prior and minimize the distances between these moments and the query sentence. In this way, it tends to gradually mine the moments which have increasing overlap with the ground truth. Moreover, we propose a Dynamic Gaussian prior Adjustment module (DGA), which further alleviates annotation bias and approximates the span of complex moments consisting of multiple events. Specifically, we adopt multiple Gaussian distributions to describe the weight distributions of moments. Therefore, the weight distributions for various moments can be flexibly adjusted and gradually approach to the ground truth. Our contributions are summarized as follows:

- We propose a Dynamic Gaussian prior based Grounding framework with Glance annotation (D3G), which facilitates the development of temporal sentence grounding with lower annotated cost.
- We propose a Semantic Alignment Group Contrastive Learning module to align the features of the positive sentence-moment pairs and a Dynamic Gaussian prior Adjustment module to ease the annotation bias and model the distributions of complex moments.
- Extensive experiments demonstrate that D3G obtains consistent and significant gains compared to method under the same annotating paradigm and outperforms weakly supervised methods by a large margin.

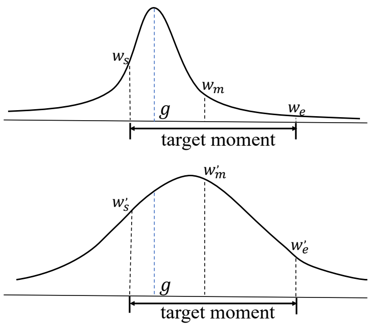


Figure 2. Illustration of annotation bias and Gaussian prior after dynamic adjustment. Top: the target moment is assigned with a low weight w_m due to the bias of glance annotation according to ViGA, which we call annotation bias. Bottom: a reasonable Gaussian distribution is obtained via DGA described in Section 3.3.

2. Related Work

Full Supervised Temporal Sentence Grounding. The FS-TSG methods can be categorized into two groups. Two-stage methods [1, 9, 11, 13, 14, 35] first propose candidate segments in a video through sliding window or proposal generation. A cross-modal matching network is then employed to find the best matching clip. However, these *propose-and-match* paradigms are time-consuming due to the numerous candidates. To reduce the redundant computation, some researchers proposed single-stage methods [2, 3, 40, 41, 30, 20, 43, 21, 39, 37]. 2D-TAN [43] constructs 2D feature map to model the temporal relations of video segment. Recently, Wang *et al.* [33] propose a Mutual Matching Network based on 2D-TAN, and further improve the performance via exploiting both intra- and inter-video negative samples. Although fully supervised methods achieve satisfying performance, they are highly dependent on accurate timestamp annotations. It is highly time-consuming and laborious to obtain these annotations for large-scale video-sentence pairs.

Weakly Supervised Temporal Sentence Grounding. Specifically, WS-TSG methods can be grouped into reconstruction-based methods [8, 18, 26, 4] and multi-instance learning (MIL) methods [19, 10, 5, 38, 12, 27]. SCN [18] employs a semantic completion network to recover the masked words in the query sentence with the generated proposals, which provides feedback for facilitating final predictions. To further exploit the negative samples in MIL-based methods, CNM [44] and CPL [45] propose to generate proposals with Gaussian functions and introduce intra-video contrastive learning. WS-TSG methods indeed advance with low annotation cost, however, there still remains a large performance gap compared to FS-TSG methods due to the discrepancy between video-level annotations and clip-level task.

Glance Supervised Temporal Sentence Grounding. Re-

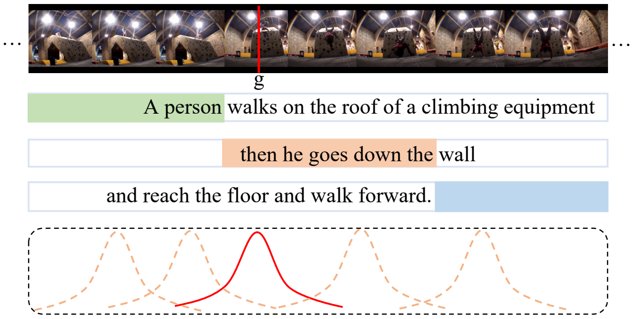

...
g
A person walks on the roof of a climbing equipment
then he goes down the wall
and reach the floor and walk forward.
...

Figure 3. Illustration of complex query consists of multiple events. Note that g indicates the position of glance annotation. The according Gaussian distribution (red curves) is hard to cover the whole target moments. We utilize DGA module to mine multiple latent Gaussian distributions (dashed line) to model such query.

cently, ViGA [6] proposes glance supervised TSG (GS-TSG) task with a new annotating paradigm. ViGA utilizes a Gaussian function to model the relevance of different clips with target moment and contrasts the clips with the queries. Though ViGA achieves promising performance, it still suffers from two limitations as mentioned in Introduction. Concurrently, Xu *et al.* [36] propose the similar task called PS-VTG, and generate pseudo segment-level labels based on language activation sequences. To better explore the Gaussian prior for TSG task with glance annotation, we propose a simple yet effective D3G, which achieves competitive performance compared with both WS-TSG and FS-TSG methods. Concurrent with our work, Ju *et al.* [15] propose a robust partial-full union framework (PFU) and achieve excellent performance with glance annotation or short-clip labels.

3. Proposed Method

3.1. Overview

Given an untrimmed video V and query sentence S, the temporal sentence grounding task aims to determine the start timestamp t_s and end timestamp t_e, where the moment $V_{t_s:t_e}$ best semantically corresponds to the query. As for FS-TSG, the exact timestamps (t_s,t_e) of corresponding moment is provided given a query description. In contrast, Cui *et al.* [6] propose a new low-cost annotating paradigm called glance annotation, which requires only single timestamp g, satisfying $g \in [t_s, t_e]$. Following the setting of [6], we propose a **D**ynamic **G**aussian prior based **G**rounding framework with **G**lance annotation (D3G) to fully unleash the potential of glance annotations.

Our D3G adopts the network architecture similar to [43, 33]. Given an untrimmed video, we firstly encode the video into feature vectors with pre-trained 2D or 3D convolutional network [25, 28] and segment the video features into N video clips. Specifically, we apply average pooling to each clip to obtain clip-level features $V = \{f_1^v, f_2^v, ..., f_N^v\} \in \mathbb{R}^{N \times D_v}$. These clip features are then passed through an FC layer to reduce their dimension, denoted as $F^{1d} \in \mathbb{R}^{N \times d_v}$. Afterwards, we encode them as 2D temporal feature map $\hat{F} \in \mathbb{R}^{N \times N \times d_v}$ following 2D-TAN [43] with the max pooling. As for language encoder, we choose DistilBERT [23] to obtain sentence-level feature $\hat{f}^s \in \mathbb{R}^{d_s}$ following [33]. Finally, to estimate the matching scores of candidate moments and the query, we utilize a linear projection layer to project the textual and visual features into same dimension d, respectively. The final representation of sentence is $f^s \in \mathbb{R}^d$ and the features of all moments are $F \in \mathbb{R}^{N \times N \times d}$. The final matching scores are given by the cosine similarity between f^s and elements of F.

3.2. Semantic Alignment Group Contrastive Learning

In this section, we aim to mine the moment which most semantically corresponds to the query and maximize the similarity between them. To achieve this goal, we have two crucial steps. First, we generate abundant candidate moments following 2D-TAN and assign them with reliable Gaussian prior weights generated with the guidance of glance annotation. Second, we propose a semantic alignment group contrastive learning to align a group of positive moments with corresponding query sentence.

To be specific, given the encoded video features $F^{1d} \in \mathbb{R}^{N \times d_v}$ and glance annotation g, we also utilize a Gaussian function parameterized with (μ, σ) to model the relations between frames and target moment, where the μ is determined by the glance g. We first scale the sequence indices $I \in \{1, 2, ..., N\}$ into domain $[-1, 1]$ by a linear transformation as follows:

$$h(i) = 2 \cdot \frac{i-1}{N-1} - 1. \tag{1}$$

Given the index i, we can obtain corresponding Gaussian weight via Eq. (2).

$$G(i, \mu, \sigma) = Norm(\frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(h(i) - h(\mu))^2}{2\sigma^2})), \tag{2}$$

where $\mu \in I$ and σ is a hyperparameter, and Norm($\cdot$) is a function used to scale values into range [0, 1].

Different from ViGA [6], we utilize the characteristic of 2D-TAN to generate a wide range of candidate moments with various durations. Given the video features $F^{1d} \in \mathbb{R}^{N \times d_v}$, we encode them into 2D feature map $F \in \mathbb{R}^{N \times N \times d}$ as shown in Figure 4, where F_{ij} denotes the feature of moment that starts at position i and ends at position j. Note that the moment is valid only when $i \leq j$. We then propose a triplet-sample strategy to generate more reasonable weights for candidate moments instead of only sampling the weight at middle point as in [6]. Specifically,

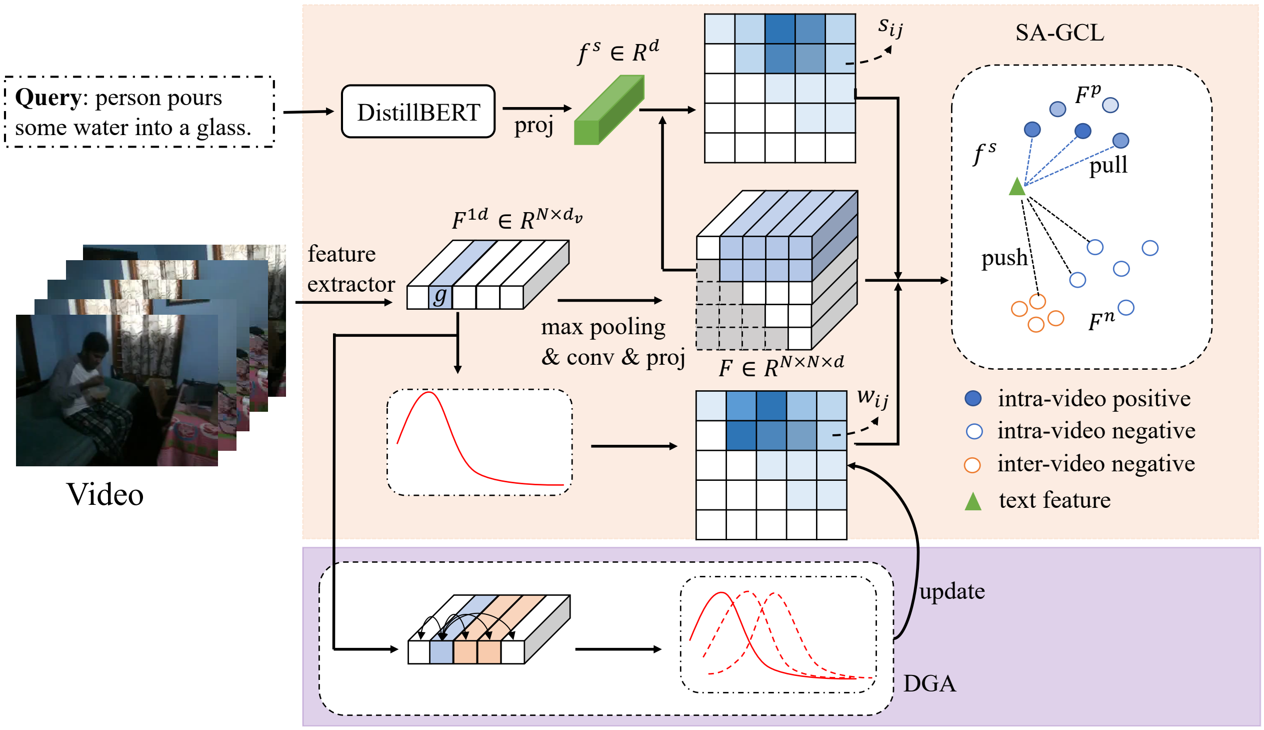


Figure 4. The overview of proposed D3G, which consists of Semantic Alignment Group Contrastive Learning (SA-GCL) and Dynamic Gaussain prior Adjustment (DGA). Note that g indicates the position of glance annotation and the grids with dashed line in F are invalid candidate moments. "proj" denotes the linear projection layer. "intra/inter" indicate the positive or negative moments sampled from same/different videos.

for each moment with start position i and end position j, we compute its Gaussian prior weight as follows:

$$w_{ij} = \frac{1}{3} \cdot (G(i, g, \sigma) + G(j, g, \sigma) + G(\lfloor \frac{i+j}{2} \rfloor, g, \sigma)), \quad (3)$$

where g is glance annotation for current target moment. In this way, those moments containing target moment but having longer durations will be penalized with lower weights.

To remedy the annotation bias, we additionally introduce semantic consistency prior to calibrate the Gaussian prior weight w_{ij} for each candidate moment. Given the query features $f^s \in \mathbb{R}^d$ and the features $F \in \mathbb{R}^{N \times N \times d}$ of candidate moments, we compute their semantic consistency scores via Eq. (4).

$$s_{ij} = \frac{f^s \cdot F_{ij}}{\| f^s \| \| F_{ij} \|}, \quad (4)$$

where $\| \cdot \|$ is l_2-norm. Afterwards, we rectify the Gaussian weight w_{ij} with semantic consistency score s_{ij} via multiplication to obtain new prior weight $p_{ij} = w_{ij} \cdot s_{ij}$.

The objective of Temporal Sentence Grounding is to learn a cross-modal embedding space, where the query sentence feature should be well aligned with the feature of corresponding moment and far way from those of irrelevant video moments. Motivated by [31, 17], we propose a Semantic Alignment Group Contrastive Learning module (SA-GCL) to gradually mine candidate moments most semantically aligned with given query sentence. To be specific, we first sample top-k candidate moments from F as positive keys for query f^s according to the new prior p_{ij}, denoted as $F^p = \{F_{ij} | 1 \le i \le j \le N\} \in \mathbb{R}^{k \times d}$. Simultaneously, we sample Gaussian weights of corresponding moments denoted as $W^p = \{w_{ij} | 1 \le i \le j \le N\} \in \mathbb{R}^k$. We then gather other candidate moments which do not contain the glance g from intra-video and all candidate moments from other videos within same batch as negative keys, denoted as $F^n = \{F_{ij} | 1 \le i \le j \le N\} \in \mathbb{R}^{N_n \times d}$, where N_n denotes the number of negative moments. The objective of SA-GCL can be described as follows:

$$\begin{aligned} L_{align} &= -\frac{1}{k} \sum_{z=0}^{k} W_z^p \log \frac{exp(f^s \cdot F_z^p / \tau)}{SUM}, \\ SUM &= \sum_{z=0}^{k} exp(f^s \cdot F_z^p / \tau) + \sum_{z=0}^{N_n} exp(f^s \cdot F_z^n / \tau), \end{aligned} \quad (5)$$

where τ is the temperature scaling factor. SA-GCL aims to maximize the similarity between the query f^s and a group of corresponding positive moments F^p under the joint embedding space while pushing away negative pairs. Note that different positive moments are assigned with corresponding prior weight W_z^p. In this way, SA-GCL effectively avoids being dominated by inaccurate moments with less similarity and tends to mine the candidate moments having large overlap with the target moment.

3.3. Dynamic Gaussian prior Adjustment

To further ease the annotation bias and characterize complex target moments, we propose a novel Dynamic Gaussian prior Adjustment module (DGA). Specifically, we utilize multiple Gaussian functions with different centers to model the local distributions of target moment and aggregate them to approximate the distribution of target moment.

Given the video features $F^{1d} \in \mathbb{R}^{N \times d_v}$ and annotation glance g, we compute the relevance of other position i with position g via Eq. (6).

$$r_{gi} = \frac{F_g^{1d} \cdot F_i^{1d}}{\parallel F_g^{1d} \parallel \parallel F_i^{1d} \parallel}. \tag{6}$$

$$\bar{r}_{gi} = (1-\alpha)\bar{r}_{gi} + \alpha r_{gi}. \tag{7}$$

To make the relevance scores more stable, we update $\bar{r}_{gi}$ with momentum factor α as shown in Eq. (7), where $\bar{r}_{gi} = r_{gi}$ at first training epoch. According to the relevance $\{\bar{r}_{gi}\}$, we can mine latent local centers for target moment. Specifically, we utilize a specific threshold T_r to filter the candidate positions and obtain a mask $M_g \in \{0,1\}^N$ for glance g as follows:

$$M_g^i = \begin{cases} 1, & if\ \bar{r}_{gi} \geq T_r \\ 0, & otherwise \end{cases} \tag{8}$$

With the mask of latent local centers, we then adjust the Gaussian prior dynamically via Eq. (9).

$$\hat{G}(i,g,\sigma) = \frac{1}{C}\sum_{z=1}^{N} M_g^z \cdot \bar{r}_{gz} \cdot G(i,z,\sigma), \tag{9}$$

where C is the summation of mask M_g. Afterwards, we replace the $G(i,g,\sigma)$ in Eq. (3) with $\hat{G}(i,g,\sigma)$, and naturally obtain dynamic Gaussian prior weight during training. Compared to ViGA, our dynamic Gaussian prior is more flexible and able to adjust the center of Gaussian distribution adaptively. Therefore, DGA further alleviates the annotation bias and provides more reliable prior weights. Besides, multiple Gaussian distributions are well suited for modeling complex target moments consisting of multiple events as shown in Figure 3. DGA tends to widen the region of high Gaussian weight via self-mining neighboring frames based on the feature of glance g and gradually generates the Gaussian prior weight well aligned with target moment. In this way, SA-GCL will be provided with positive moments of high quality, which eventually promotes the cross-modal semantic alignment learning and accurate localization of target moments.

Discussion. To clearly distinguish the differences between D3G and few similar works, we give some explanations here. As for MMN, D3G shares the same process of generating candidate moments following 2D-TAN, which is not the key contribution of our method. MMN utilizes normal one-to-one contrastive learning, which is no longer suitable to glance annotation. However, D3G instead adopts a suitable sample strategy and corresponding adapted group contrastive learning, which is key component to unleash the potential of glance annotations. As for CPL, we also know that it utilizes multiple Gaussian distributions to describe positive moments. However, it actually selects one most matched positive moment guided by the loss of masked language reconstruction for contrastive learning, while D3G utilizes multiple Gaussian functions to adaptively model complex queries consisting of multiple events and samples a group of positive moments for contrastive learning.

4. Experiments

In order to validate the effectiveness of the proposed D3G, we conduct extensive experiments on three publicly available datasets: Charades-STA [9], TACoS [9] and ActivityNet Captions [16].

4.1. Datasets

Charades-STA is built on dataset Charades [24] for temporal sentence grounding. It contains 12,408 and 3,720 moment-sentence pairs for training and testing.

TACoS consists of 127 videos selected from the MPII Cooking Composite Activities video corpus [22]. We follow the standard split from [9], which contains 10,146, 4,589 and 4,083 moment-sentence pairs for training, validation and testing, respectively. We report the evaluation results on the test set for fair comparison.

ActivityNet Captions is originally designed for video captioning and recently introduced into temporal sentence grounding. It contains 37,417, 17,505 and 17,031 moment-sentence pairs for training, validation and testing, respectively. We report the evaluation results following [43, 33].

Specially, we adopt the glance annotation released by [6] for training set, where the temporal boundary is replaced with the timestamp g uniformly sampled within the original temporal boundary.

4.2. Evaluation Metric and Implementation Details

Evaluation Metric. Following previous works [9, 43], we evaluate our model with metric 'R@n,IoU=m', which means the percentage of at least one of the top-n results having Intersection over Union (IoU) larger than m. Specifically, we report the results with $m \in \{0.5, 0.7\}$ for

Charades-STA, $m \in \{0.3, 0.5, 0.7\}$ for TACoS and ActivityNet Captions, and $n \in \{1, 5\}$ for all datasets.

Implementation Details. In this work, our main framework is extended from MMN [33] and most of experiment settings keep the same. For fair comparison, following [33], we adopt off-the-shelf video features for all datasets (VGG feature for Charades and C3D feature for TACoS and ActivityNet Captions). Specifically, the dimension of joint feature space d is set to 256 and τ is set to 0.1. In SA-GCL, we set the k as 10, 20 and 20 for Charades, TACoS and ActivityNet Captions, respectively. The σ in Eq. (2) is set to 0.3, 0.2 and 0.6 for Charades, TACoS and ActivityNet Captions. In DGA, T_r and α is set as 0.9 and 0.7, respectively.

4.3. Comparisons with the State-Of-The-Art

In order to provide comprehensive analysis, we compare the proposed D3G with both fully/weakly/glance supervised methods. As shown in Table 1, Table 2 and Table 3, D3G achieves highly competitive results on three datasets under glance supervision, and achieves comparable performance compared with fully supervised methods. Note that we highlight the best value for each setting respectively. Based on the experimental results, we can draw the following conclusions:

(1) Glance annotation provides more potential to achieve better performance for temporal sentence grounding with lower annotation cost. Although it is not entirely fair to directly compare D3G with other weakly supervised methods due to introducing extra supervision, D3G significantly exceeds most of weakly supervised methods by a large margin with trivial increment of annotation cost. Since PS-VTG and PFU adopt more robust I3D feature, they obviously outperform D3G on Charades-STA. However, D3G instead is superior to PS-VTG on more challenging TACoS with same features. Besides, weak supervised methods are often not tested on TACoS, where the videos are very long and contain a large number of target moments. However, D3G obtains promising performance and outperforms ViGA by a large margin on TACoS as shown in Table 2.

(2) D3G effectively exploits the information provided by glance annotation and mines more moments of high quality for training compared with ViGA. Due to the limitations of fixed scale Gaussian function and fixed sliding window, ViGA fails to mine accurate candidate moments to learn a well-aligned joint embedding space. Instead, D3G generates a wide range of candidate moments and samples a group of reliable candidate moments for group contrastive learning. Compared to ViGA, D3G achieves obvious gains 5.08% and 3.5% at R@1 IoU=0.5 and R@1 IoU=0.7 on Charades-STA, respectively. Specially, significant improvements are obtained at R@5 on three datasets.

(3) D3G substantially narrows the performance gap between weakly/glance supervised methods and fully supervised methods. Specifically, D3G already surpasses previous method (*e.g.*, CTRL) on both TACoS and ActivityNet Captions. Undeniably, there are still non-negligible margins compared to the state-of-the-art fully supervised methods (*e.g.*, MMN). Note that D3G is very concise and not embedded with auxiliary module (*e.g.*, MLM used in [45]). D3G still can be enhanced with some complementary modules.

Method	R@1		R@5	
	IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
MAN [41]	41.21	20.54	83.21	51.85
2D-TAN [43]	39.70	23.31	80.32	51.26
SSCS [7]	43.15	25.54	**84.26**	54.17
MMN [33]	**47.31**	**27.28**	83.74	**58.41**
CRM [12]	34.76	16.37	-	-
CNM [44]	35.43	15.45	-	-
LCNet [38]	**39.19**	**18.87**	**80.56**	**45.24**
CPL† [45]	32.27	14.22	78.34	43.45
PS-VTG‡ [36]	39.22	20.17	-	-
PFU‡ [15]	**54.66**	**28.34**	-	-
VIGA* [6]	36.56	16.10	48.90	25.86
D3G	41.64	19.60	**79.25**	**49.30**

Table 1. Performance comparison on Charades-STA under different supervision settings.Top:full supervision, Middle: weak supervision, Bottom:glance supervision. †we reproduce the results with official code and VGG features for fair comparison. *we reproduce the results with official code for results at R@5.‡ indicates the method utilizes I3D features.

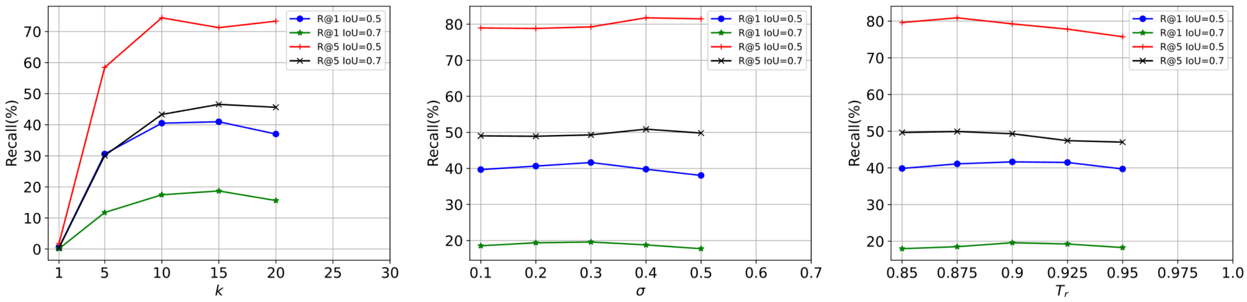
Method	R@1			R@5		
	IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7
CTRL [9]	18.32	13.30	-	36.69	25.42	-
2D-TAN [43]	37.29	25.32	-	57.81	24.04	-
SSCS [7]	41.33	29.56	-	60.65	48.01	-
MMN [33]	38.57	27.24	-	65.31	50.69	-
MAT [42]	**48.79**	**37.57**	-	**67.63**	**57.91**	-
VIGA* [6]	20.82	9.52	3.10	27.92	15.35	6.10
PS-VTG [36]	23.64	10.00	3.35	-	-	-
D3G	**27.27**	**12.67**	**4.70**	**54.61**	**31.34**	**12.35**

Table 2. Performance comparison on TACoS under different supervision settings.Top:full supervision, Bottom:glance supervision. *we reproduce the results with official code for results at R@5.

4.4. Ablation Study

To validate the effectiveness of different components of the proposed D3G and investigate the impacts of hyperparameters, we perform ablation studies on Charades-STA.

Effectiveness of SA-GCL and DGA. Since L_{align} is the only loss of D3G, to validate the effectiveness of SA-GCL, we need to simplify the SA-GCL module as a baseline. Specifically, we only sample the top-1 positive moment to



(a) Effect of top-k.

(b) Effect of σ.

(c) Effect of T_r.

Figure 5. Effect of different hyper-parameters on Charades-STA dataset.

Method	R@1			R@5		
	IoU=0.3	IoU=0.5	IoU=0.7	IoU=0.3	IoU=0.5	IoU=0.7
CTRL [9]	47.43	29.01	10.34	75.32	59.17	37.54
2D-TAN [43]	59.46	44.51	26.54	85.53	77.13	61.96
LGI [20]	58.52	41.51	23.07	-	-	-
SSCS [7]	61.35	46.67	27.56	86.89	78.37	63.78
MMN [33]	**65.05**	**48.59**	29.26	**87.25**	**79.50**	**64.76**
MAT [42]	-	48.02	**31.78**	-	78.02	63.18
CRM [12]	55.26	32.19	-	-	-	-
CNM [44]	**55.68**	**33.33**	-	-	-	-
LCNet [38]	48.49	26.33	-	**82.51**	**62.66**	-
CPL [45]	53.67	31.24	-	63.05	43.14	-
VIGA* [6]	**59.78**	35.39	16.25	72.19	53.19	32.69
PS-VTG [36]	59.71	**39.59**	**21.98**	-	-	-
PFU [15]	59.63	36.35	16.61	-	-	-
D3G	58.25	36.68	18.54	**87.84**	**74.21**	**52.47**

Table 3. Performance comparison on ActivityNet Captions under different supervision settings.Top:full supervision, Middle: weak supervision, Bottom:glance supervision. *we reproduce the results with official code for results at R@5.

Module		R@1		R@5	
SA-GCL	DGA	IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
✓†		5.08	0.81	14.78	3.36
✓†	✓	13.92	3.31	33.55	11.77
✓		40.51	16.10	74.41	43.31
✓	✓	41.64	19.60	79.25	49.30

Table 4. Effectiveness of SA-GCL and DGA in D3G on Charades-STA. ✓† denotes an simplified implementation of SA-GCL.

Types		R@1		R@5	
GW	SC	IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
✓		38.09	16.10	66.53	36.51
	✓	25.67	9.57	65.43	38.52
✓	✓	40.51	16.10	74.41	43.31

Table 5. Impact of different strategies used to sample positive moments for SA-GCL on Charades-STA. GW: Gaussian weight, SC: semantic consistency.

compute the normal contrastive loss (degraded to simplified MMN) as shown in the first row of Table 4. However, the top-1 moment tends to be the shortest moment and has small overlap with target moment, which is decided by the intrinsic characteristic of 2D-TAN. Therefore, the performance of baseline is undoubtedly very poor, which demonstrates that the main improvement of D3G is not brought by the backbone of MMN. This phenomenon then encourages us to sample a group of positive moments in SA-GCL. With full SA-GCL, the model obtains notable performance gains. Moreover, we introduce the DGA to alleviate annotation bias and model some complex target moments consisting of multiple events. After equipped with DGA, D3G and simplified D3G achieve obvious performance improvement.

Impact of Sampling Strategy. In SA-GCL, sampling a group of reliable positive moments is of great importance. We investigate the impacts of two priors: Gaussian weight and semantic consistency, respectively. As shown in the first row of Table 5, we sample top-k positive moments according to the Gaussian prior weight. An alternative scheme is that we sample top-k positive moments according to the semantic consistency scores between candidate moments and query sentence. However, both of them obtain sub-optimal performance. This is because Gaussian prior weight is not always reliable due to the annotation bias and semantic consistence scores are highly dependent on the stability of features. Therefore, we finally fuse these two priors to obtain relatively reliable prior. As shown in the third row of Table 5, obvious performance gains are obtained after both of them are utilized, which demonstrates that these two priors indeed complement each others.

Effect of different hyper-parameters. As shown in Figure 5, we investigate three critical hyperparameters in D3G. As verified in Table 4, sampling enough latent positive moments is beneficial to mining target moment for training. As shown in Figure 5 (a), the performance gains increase obviously as the k increases. However, it begins to decrease

Method	R@1		R@5	
	IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
ViGA	36.56	16.10	48.90	25.86
D3G	41.64	19.60	79.25	49.30
ViGA$^+$	$33.66_{(-2.90)}$	$14.65_{(-1.45)}$	$47.45_{(-1.45)}$	$25.51_{(-0.35)}$
D3G$^+$	$40.19_{(-1.45)}$	$19.62_{(+0.02)}$	$78.90_{(-0.35)}$	$49.41_{(+0.11)}$

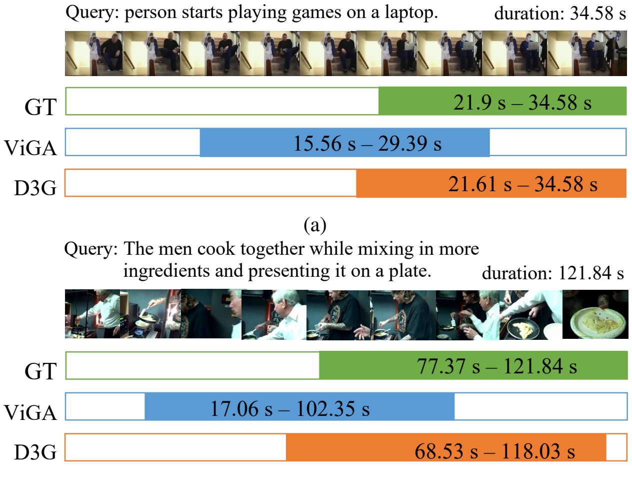
Table 6. Performance comparison on Charades-STA with extreme glance annotation. $^+$ indicates according method is trained with extreme glance annotations.

after the k reaches a specific value. We argue that selecting excessive positive moments tends to incorporate some false positive moments and therefore degrades the performance. We finally set the k to 10 for Charades-STA, which balances well the performance and computational cost. As for hyperparameter σ, it essentially decides the width of Gaussian distribution. A larger σ can well characterize the target moment of longer duration and vice versa. We vary the σ from 0.1 to 0.5, and observe that value 0.3 is relatively suitable for the Charades-STA dataset. As for hyperparameter T_r in Eq. (8), it controls the degree of dynamic Gaussian prior adjustment. We conduct experiments with relevance thresholds around 0.9. A small threshold tends to introduce interference while a large threshold fails to find the neighbor frames with consistent semantic. As shown in Figure 5 (c), the moderate threshold 0.9 relatively balances the aforementioned dilemma.

Tolerance to Extreme Glance Annotation. In order to verify the ability of addressing extreme glance annotation, we first generate extreme glance annotation, where only the positions near the start/end timestamps will be sampled as glance g. As shown in Table 6, both ViGA$^+$ and D3G$^+$ are confronted with the performance degradation at some metrics(*e.g.*, R@1 IoU=0.5). However, the performances of D3G are relatively stable compared to ViGA, which demonstrate that D3G indeed is able to alleviate annotation bias.

4.5. Qualitative Analysis

To clearly reveal the effectiveness of our method, we visualize some qualitative examples from the test split of Charades-STA dataset and ActivityNet Captions dataset. As shown in Figure 6, the proposed D3G achieves more accurate localization of target moment compared to ViGA. Specifically, ViGA cannot well align the visual content and semantic information and tend to be disturbed by irrelevant content, which may be caused by the annotation bias. Instead, D3G utilizes SA-GCL and DGA to alleviate the annotation bias, which enables D3G to well align the query with the corresponding moment. Moreover, the DGA adopts multiple Gaussian functions to model target moment, which is beneficial to representing the complete distribution of complex moments consisting of multiple events. As shown in Figure 6 (b), D3G still effectively localizes the complex moments while ViGA misses the last events "represent it on a plate". More qualitative examples will be provided in Supplementary Materials.



(a)

(b)

Figure 6. Qualitative examples of top-1 predictions. (a) and (b) is from the Charades-STA dataset and the ActivityNet Captions dataset, respectively. GT indicates the ground truth temporal boundary.

5. Conclusion

In this study, we investigate a recently proposed task, Temporal Sentence Grounding with Glance Annotation. Under this setting, we propose a **D**ynamic **G**aussian prior based **G**rounding framework with **G**lance annotation, termed D3G. Specifically, D3G consists of a Semantic Alignment Group Contrastive Learning module (SA-GCL) and a Dynamic Gaussian prior Adjustment module (DGA). SA-GCL aims to mine a wide range of positive moments and align the positive sentence-moment pairs in the joint embedding space. DGA effectively alleviates the annotation bias and models complex query consisting of multiple events via dynamically adjusting the Gaussian prior with multiple Gaussian functions, promoting the precision of localization. Extensive experiments show that D3G significantly narrows the performance gap between fully supervised methods and glance supervised methods. Without excessive interaction of visual-language, D3G provides a concise framework and a fresh insight to the challenging temporal sentence grounding under low-cost glance annotation.

Limitations. Although D3G achieves promising improvements with glance annotations, it still has some limitations. In this paper, the DAG adjusts Gaussian prior via the combination of multiple fixed scale Gaussian functions. It fails to scale down the Gaussian distribution to fit the small moments. It is expected to explore dynamic learnable Gaussian functions to model moment of arbitrary duration in future work. Besides, the sampling strategy for SA-GCL is still not enough flexible to sample accurate positive moments.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 1, 2

[2] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 162–171, 2018. 2

[3] Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. Localizing natural language in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8175–8182, 2019. 2

[4] Shaoxiang Chen and Yu-Gang Jiang. Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8425–8435, 2021. 2

[5] Zhenfang Chen, Lin Ma, Wenhan Luo, Peng Tang, and Kwan-Yee K Wong. Look closer to ground better: Weakly-supervised temporal grounding of sentence in video. *arXiv preprint arXiv:2001.09308*, 2020. 2

[6] Ran Cui, Tianwen Qian, Pai Peng, Elena Daskalaki, Jingjing Chen, Xiaowei Guo, Huyang Sun, and Yu-Gang Jiang. Video moment retrieval from text queries via single frame annotation. *arXiv preprint arXiv:2204.09409*, 2022. 1, 3, 5, 6, 7

[7] Xinpeng Ding, Nannan Wang, Shiwei Zhang, De Cheng, Xiaomeng Li, Ziyuan Huang, Mingqian Tang, and Xinbo Gao. Support-set based cross-supervision for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11573–11582, 2021. 6, 7

[8] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[9] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 1, 2, 5, 6, 7

[10] Mingfei Gao, Larry S Davis, Richard Socher, and Caiming Xiong. Wslln: Weakly supervised natural language localization networks. *arXiv preprint arXiv:1909.00239*, 2019. 2

[11] Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. Mac: Mining activity concepts for language-based temporal localization. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 245–253. IEEE, 2019. 2

[12] Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. Cross-sentence temporal and semantic relations in video activity localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7199–7208, 2021. 1, 2, 6, 7

[13] Bin Jiang, Xin Huang, Chao Yang, and Junsong Yuan. Cross-modal video moment retrieval with spatial and language-temporal attention. In *Proceedings of the 2019 on international conference on multimedia retrieval*, pages 217–225, 2019. 2

[14] Yifan Jiao, Zhetao Li, Shucheng Huang, Xiaoshan Yang, Bin Liu, and Tianzhu Zhang. Three-dimensional attention-based deep ranking model for video highlight detection. *IEEE Transactions on Multimedia*, 20(10):2693–2705, 2018. 2

[15] Chen Ju, Haicheng Wang, Jinxiang Liu, Chaofan Ma, Ya Zhang, Peisen Zhao, Jianlong Chang, and Qi Tian. Constraint and union for partially-supervised temporal sentence grounding. *arXiv preprint arXiv:2302.09850*, 2023. 3, 6, 7

[16] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 5

[17] Hanjun Li, Xingjia Pan, Ke Yan, Fan Tang, and Wei-Shi Zheng. Siod: Single instance annotated per category per image for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14197–14206, 2022. 4

[18] Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. Weakly-supervised video moment retrieval via semantic completion network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11539–11546, 2020. 2

[19] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11592–11601, 2019. 1, 2

[20] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10810–10819, 2020. 2, 7

[21] Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. Interventional video grounding with dual contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2765–2775, 2021. 2

[22] Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. Script data for attribute-based recognition of composite activities. In *European conference on computer vision*, pages 144–157. Springer, 2012. 5

[23] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 3

[24] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016. 5

[25] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[26] Yijun Song, Jingwen Wang, Lin Ma, Zhou Yu, and Jun Yu. Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. *arXiv preprint arXiv:2003.07048*, 2020. 2

[27] Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A Plummer. 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*, pages 2083–2092, 2021. 2

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

[29] Hao Wang, Zheng-Jun Zha, Liang Li, Dong Liu, and Jiebo Luo. 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*, pages 7026–7035, 2021. 1

[30] Jingwen Wang, Lin Ma, and Wenhao Jiang. Temporally grounding language queries in videos by contextual boundary-aware prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12168–12175, 2020. 2

[31] Ximei Wang, Jinghan Gao, Mingsheng Long, and Jianmin Wang. Self-tuning for data-efficient deep learning. In *International Conference on Machine Learning*, pages 10738–10748. PMLR, 2021. 4

[32] Yuechen Wang, Wengang Zhou, and Houqiang Li. Fine-grained semantic alignment network for weakly supervised temporal language grounding. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 89–99, 2021. 1

[33] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative sample matters: A renaissance of metric learning for temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2613–2623, 2022. 1, 2, 3, 5, 6, 7

[34] Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. Boundary proposal network for two-stage natural language video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2986–2994, 2021. 1

[35] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9062–9069, 2019. 2

[36] Zhe Xu, Kun Wei, Xu Yang, and Cheng Deng. Point-supervised video temporal grounding. *IEEE Transactions on Multimedia*, 2022. 3, 6, 7

[37] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16442–16453, 2022. 2

[38] Wenfei Yang, Tianzhu Zhang, Yongdong Zhang, and Feng Wu. Local correspondence network for weakly supervised temporal sentence grounding. *IEEE Transactions on Image Processing*, 30:3252–3262, 2021. 2, 6, 7

[39] Xun Yang, Shanshan Wang, Jian Dong, Jianfeng Dong, Meng Wang, and Tat-Seng Chua. Video moment retrieval with cross-modal neural architecture search. *IEEE Transactions on Image Processing*, 31:1204–1216, 2022. 2

[40] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9159–9166, 2019. 2

[41] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1247–1257, 2019. 1, 2, 6

[42] Mingxing Zhang, Yang Yang, Xinghan Chen, Yanli Ji, Xing Xu, Jingjing Li, and Heng Tao Shen. Multi-stage aggregated transformer network for temporal language localization in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12669–12678, 2021. 1, 6, 7

[43] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12870–12877, 2020. 1, 2, 3, 5, 6, 7

[44] Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu. Weakly supervised video moment localization with contrastive negative sample mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 1, page 3, 2022. 1, 2, 6, 7

[45] Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu. Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15555–15564, 2022. 1, 2, 6, 7

Appendix

A. Effectiveness of SA-GCL and DGA

To further analyze the effectiveness of SA-GCL and DGA, we provide more detailed experimental results on ActivityNet Captions and TACoS datasets as shown in Table 7 and Table 8. Following the main manuscript, we regard the simplified implementation of SA-GCL as a baseline. After being equipped with the complete SA-GCL, our model achieves significant improvements on both ActivityNet Captions and TACoS. This phenomenon demonstrates that sampling enough positive moments for contrastive learning is of great importance. Additionally, we further incorporate the DGA module for alleviating the annotation bias and modeling complex target moments. Since the ActivityNet Captions dataset has a large number of complex query sentences consisting of multiple events, D3G obtains notable performance gains on ActivityNet Captions(*e.g.* 9.03% at R@5 IoU=0.7). However, TACoS is still challenging for D3G due to the dense distributions of target moments.

Module		R@1		R@5	
SA-GCL	DGA	IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
$\checkmark^{\dagger}$		0.83	0.28	1.78	0.58
✓		32.65	16.00	65.48	43.44
✓	✓	36.68	18.54	74.21	52.47

Table 7. Effectiveness of SA-GCL and DAG in D3G on ActivityNet Captions. $\checkmark^{\dagger}$ denotes an simplified implementation of SA-GCL.

Module		R@1		R@5	
SA-GCL	DGA	IoU=0.5	IoU=0.7	IoU=0.5	IoU=0.7
$\checkmark^{\dagger}$		2.97	0.37	5.40	1.10
✓		11.95	4.20	29.07	10.30
✓	✓	12.67	4.70	31.34	12.35

Table 8. Effectiveness of SA-GCL and DAG in D3G on TACoS. $\checkmark^{\dagger}$ denotes an simplified implementation of SA-GCL.

B. Effect of different hyper-parameters

In this section, we investigate the effect of two critical hyperparameters on ActivityNet Captions and TACoS datasets. As shown in Figure 7 and Figure 8, we report the changes in performance at four metrics. As for top-k, the performance increases dramatically as the k increases. However, the performance gradually achieves saturation after the k reaches 15. We finally select $k = 20$ for both ActivityNet Captions and TACoS. As for σ, the ActivityNet Captions dataset tends to prefer large values while small values are more suitable for the TACoS dataset. This is because the former contains a large number of long target moments while the latter contains numerous short target moments. As shown in Figure 7 and Figure 8, we eventually select $\sigma = 0.6$ and $\sigma = 0.2$ for ActivityNet Captions and TACoS for optimal performance, respectively.

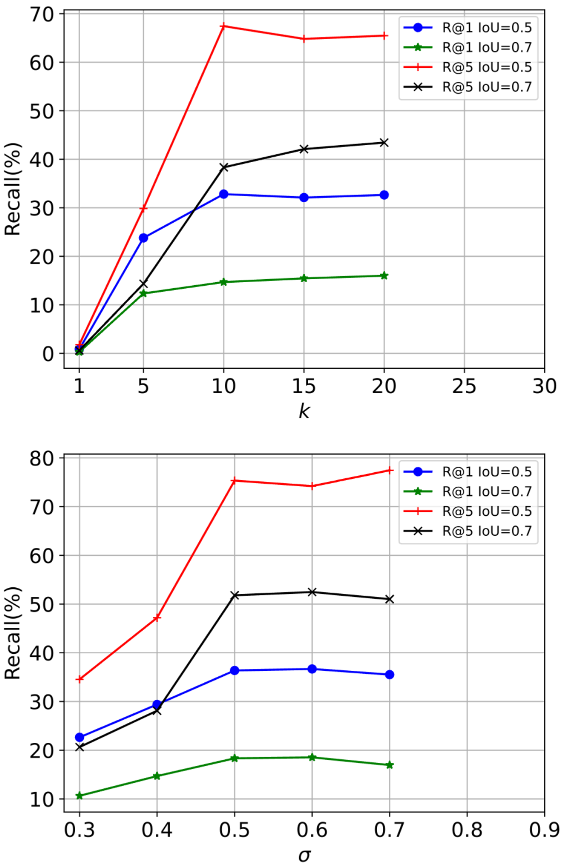


Figure 7. Effect of top-k and σ on ActivityNet Captions dataset.

C. Qualitative Analysis

In this section, we provide more qualitative examples from the test split of the Charades-STA dataset, ActivityNet Captions dataset, and TACoS dataset. For each video, we select two queries for analysis. As shown in Figure 9 (a), D3G locates the target moment accurately while ViGA ignores the reason at the front of the target moment, given Query 1. However, D3G is inferior to ViGA in some cases such as Query 2. As for complex queries in ActivityNet Captions, D3G still localizes a moment with a large overlap with the target moment. Since sentence-level features may lose some information about specific events, D3G cannot perceive accurate boundaries for some complex queries, such as Figure 9 (b) Query 2. It is expected to explore event-

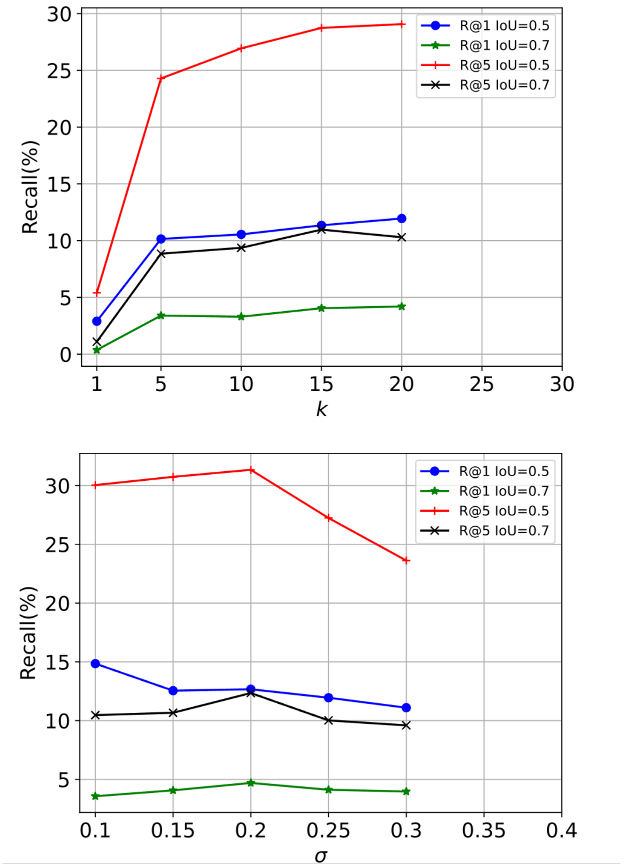
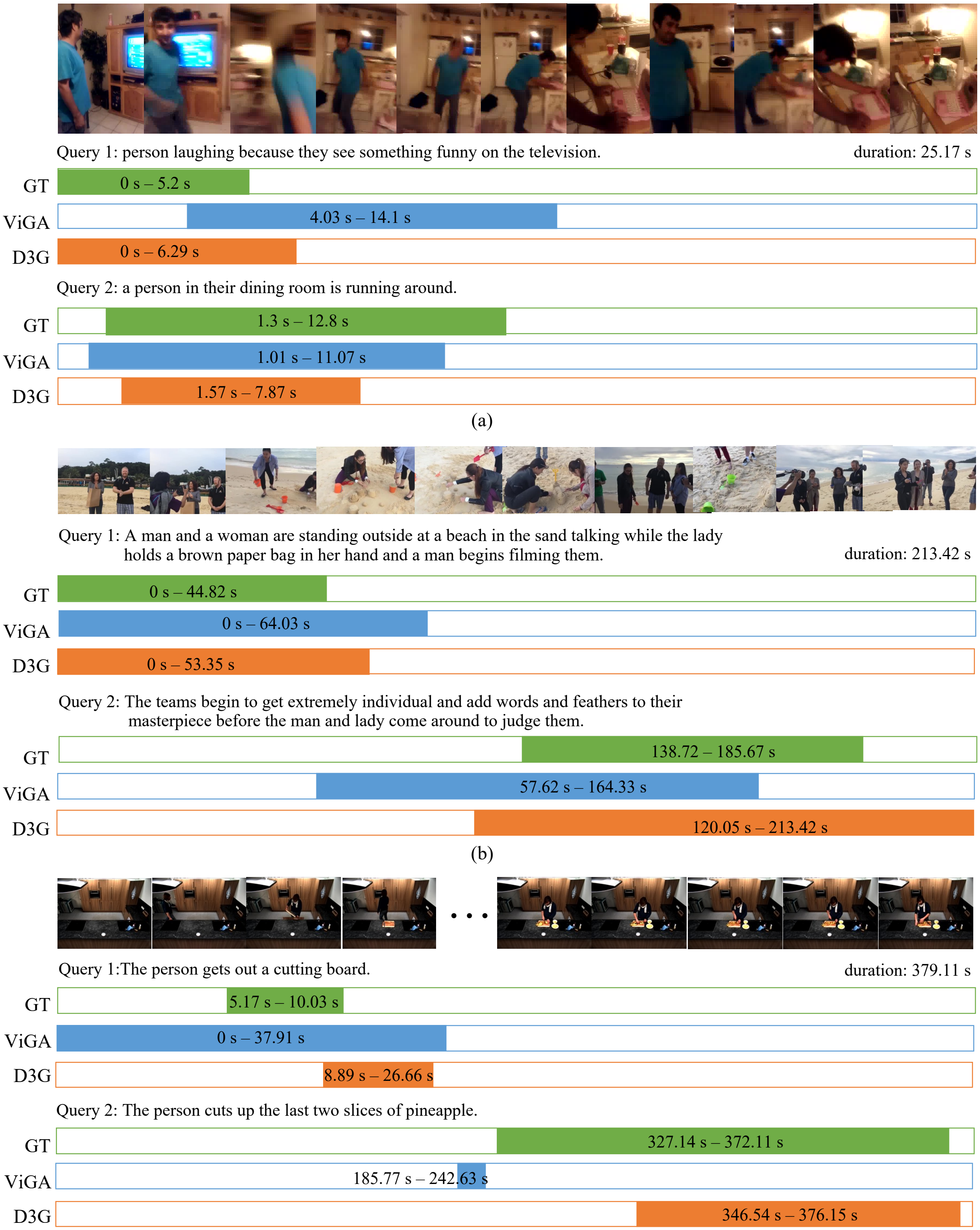


Figure 8. Effect of top-k and σ on TACoS dataset.

level features for queries consisting of multiple events in the future. TACoS is the most challenging dataset, where the videos have long durations and contain a large number of moment-sentence pairs. As shown in Figure 9 (c), we observe that D3G fails to locate a simple query of short duration from the long video, given Query 1. However, D3G accurately locates the target moment of long duration given Query 2. Note that D3G well attends to the number "the last two" of the query while ViGA fails to attend to such information and locates irrelevant moments. As observed in Figure 9, D3G is superior to ViGA, which is consistent with the experimental results in the main manuscript. However, D3G still has some limitations and needs to be improved in the future.



(a)

(b)

(c)

Figure 9. Qualitative examples of top-1 predictions. (a), (b) and (c) is from the Charades-STA dataset, the ActivityNet Captions and the TACoS dataset, respectively. GT indicates the ground truth temporal boundary.