Boundary-aware Temporal Sentence Grounding with Adaptive Proposal Refinement

Jianxiang Dong and Zhaozheng Yin

Stony Brook University, New York, USA
jianxiang.dong@stonybrook.edu, zyin@cs.stonybrook.edu

Abstract. Temporal sentence grounding (TSG) in videos aims to localize the temporal interval from an untrimmed video that is relevant to a given query sentence. In this paper, we introduce an effective proposal-based approach to solve the TSG problem. A Boundary-aware Feature Enhancement (BAFE) module is proposed to enhance the proposal feature with its boundary information, by imposing a new temporal difference loss. Meanwhile, we introduce a Boundary-aware Feature Aggregation (BAFA) module to aggregate boundary features and propose a Proposal-level Contrastive Learning (PCL) method to learn query-related content features by maximizing the mutual information between the query and proposals. Furthermore, we introduce a Proposal Interaction (PI) module with Adaptive Proposal Selection (APS) strategies to effectively refine proposal representations and make the final localization. Extensive experiments on Charades-STA, ActivityNet-Captions and TACoS datasets show the effectiveness of our solution. Our code is available at https://github.com/DJX1995/BAN-APR.

1 Introduction

Temporal sentence grounding (TSG) in videos has been an important but challenging problem in Vision-Language understanding area [1], which requires techniques from both Computer Vision and Natural Language Processing. It aims to localize a temporal video moment[1] in an untrimmed video that semantically matches a given query sentence, as shown in Figure 1(a). In addition, TSG can be widely used in many applications such as question answering in videos, video event captioning and human computer interaction [2–4].

Most TSG methods can be categorized into two groups: (1) proposal-based approaches and (2) proposal-free methods. The former one borrows techniques from object detection. A set of proposals with start and end timestamps are proposed, then visual-language interaction is performed to generate proposal representations. Afterwards, it computes the similarity between each proposal and the given query sentence and selects the proposal with the highest matching score as the localized video moment. The latter one also performs complex visual-language interaction but it directly predicts the start and end timestamps based on the feature representation.

[1] We define a moment to be an interval in the video with the start and end timestamps.

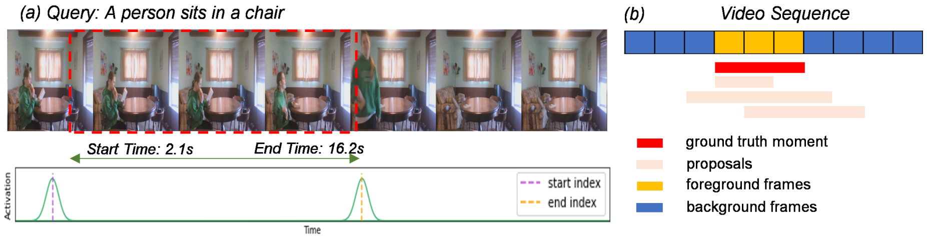


Fig. 1. (a) An illustration of the temporal sentence grounding in videos. *Bottom:* We expect the learned boundary-aware feature representation to have high activations of temporal difference around the start and end timestamps. (b) Proposal-based methods aggregate video clips within each proposal to generate the proposal feature. However, these proposals may include video clips (background frames) outside the ground truth moment (foreground frames), which will affect the quality of the learned proposal representations.

Despite the success of previous methods, there are three challenges remaining unsolved in TSG:

- *Feature discrimination* around temporal boundaries. Most of previous methods usually apply the pooling operation on cross-modal representations, e.g., average pooling or RoI (Region of Interest) pooling [5, 6] to generate proposal features, which neglects the discriminative temporal boundary information. Without this information, it is difficult to distinguish overlapping or nearby proposals correctly because the pooling results of overlapped or neighbouring regions are very similar.
- *Proposal construction* with precise feature representations. Another challenging but crucial problem in TSG task is how to extract the query-related visual information when constructing proposal representations. For example, in Figure 1(b), some proposals include the whole ground truth moment and some have an overlap with the ground truth moment. When aggregating proposal representations based on video clips within proposals' start and end timestamps, we introduce a lot of noises from the unrelated background frames (frames outside the ground truth moment), which may affect the quality of the generated proposal representations.
- *Proposal interaction.* The proposal-free methods do not construct proposal representations and also neglect the informative proposal relationship. Proposal-based approaches usually have a large number of proposals and use a complex proposal interaction module (e.g., a 2D CNN [7] or graph neural network [4]) to model the proposal relationship, which requires a lot of computations during training and inference.

To address the above challenges, we develop a novel Boundary-aware Network with Adaptive Proposal Refinement (BANet-APR) for temporal sentence grounding:

- Firstly, we design a Boundary-aware Feature Enhancement (BAFE) module to extract the start and the end boundary information. A temporal difference loss is applied onto the boundary-aware feature to ensure that the feature representation has a high activation of temporal difference around the start and end boundary positions (The concept is shown in Figure 1(a) and the validation is shown in Figure 6 in the experiment section).
- Secondly, we introduce a Boundary-aware Feature Aggregation (BAFA) module to generate discriminative and informative proposal representations where the boundary feature and semantic content feature are both considered when constructing proposal representations. Meanwhile, we design a Proposal-level Contrastive Learning (PCL) method to implicitly enforce the semantic content feature to be query-related. Unlike previous methods [8, 9] which either adopt frame-based contrastive learning or predefine an IoU threshold to construct positive and negative sets, we enforce proposals that include the whole ground truth moment to be close to the query while proposals that have no overlap with the ground truth moment to be away from the query. The boundary-aware feature and the semantic content feature are complement to each other when generating proposal representations to learn more discriminative (from boundary-aware feature) and informative (from semantic content feature) proposal features for the TSG task.
- Finally, we propose a refinement module which makes coarse predictions, selects k proposals via Adaptive Proposal Selection (APS), performs Proposal Interaction (PI) among the selected proposals and then localizes the moment. This refinement module only models the interaction between the selected k confident and representative proposals instead of all proposals as used in many previous works [10, 4, 7], which is more effective and efficient.

Our main contributions are fourfold: 1) We design a Boundary-aware Feature Enhancement (BAFE) module to extract boundary information and design a temporal difference loss that is directly applied on the feature representation to learn more discriminative proposal features; 2) We propose a novel Boundary-aware Feature Aggregation (BAFA) with Proposal-level Contrastive Learning (PCL) method to learn more discriminative and informative proposal features; 3) We propose a Proposal Interaction (PI) module with Adaptive Proposal Selection (APS) strategies to effectively refine proposal representations and make the final localization prediction; and 4) Compared to the latest state-of-the-art, the experiments on three datasets (Charades-STA [11], ActivityNet Captions [12] and TACoS [13]) show the effectiveness of our proposed BANet-APR.

2 Related Work

Temporal sentence grounding (TSG) is a new task introduced in the computer vision community recently [11, 14, 15]. Formally, it aims to retrieve a video moment with start and end timestamps from an untrimmed video using a query sentence. Current existing methods can be roughly grouped into two categories, namely propose-based and proposal-free methods [14, 16, 17, 10, 18–22]. Proposal-based methods first pre-define a set of video moment proposals and solve the problem by choosing the best matched proposal [7, 4, 11]. CTRL [11] adopts a sliding window to generate moment proposals and rank the proposals based on their similarity scores to the query sentence. 2D-TAN [7] and RaNet [4] enumerate all possible proposals and introduce complex proposal interaction methods using convolution and graph neural network, respectively. Proposal-free methods directly regress the start and end timestamps or classify which frames are the start and end boundary frames of the matched video moment [23, 1, 24, 25].

Extracting informative boundary feature is very important and has been explored by some previous works [26, 4, 27, 28]. RaNet [4] adopts two branches, e.g., a start branch and an end branch, to capture the start and end boundary information to learn more discriminative proposal representations. In addition to boundary branches, De-VLTrans-MSA [26] attaches a MLP as a classifier on top of the feature and defines a downstream task to predict the probability of a frame being the start or end stage to help boundary feature learning. CBP [27] designs an anchor submodule and a boundary submodule which take the feature vector at every temporal position as input and predict the probability of a proposal ending at that position. However, the classifier for the downstream tasks is easy to overfit to the statistics of the dataset and we may train a good classifier instead of good feature representations. Unlike previous methods which either have no boundary-related constraints on the feature, or use a classifier to predict the probability of a frame being the start and end boundary, our paper directly imposes a novel temporal difference loss on the boundary feature to enforce it to be boundary-aware. The loss is calculated only based on boundary regions instead of all temporal positions (so non-boundary regions with high temporal difference will not get punished). It allows the model itself to determine the high-salient temporal locations, which reduces the side effect of overfitting.

Proposal-based methods are intuitive and follow similar spirits of anchor-based approaches in object detection, but it suffers from the redundant computation cost. 2D-TAN [7] proposes a sparse sampling strategy to sparsely sample long proposals and densely sample short proposals, which is adopted by many other works. However, the proposal number is still too large for modeling dense proposal relationships. APGN [29] proposes a two-step model which first regresses a small number of (start, end) tuples as proposals and then ranks the proposed proposals. LPNet [30] proposes a novel model with a fixed set of learnable proposals where proposals' start and end positions serve as model parameters and are updated at every iteration. However, there is no constraint on the generated proposals, which may lead to repeated and useless proposals. Therefore, we propose to adaptively select unrepeated and representative proposals for further refinement.

Contrastive learning [31–34] methods are commonly used in self-supervised learning (SSL) research to learn high quality representations in an unsupervised manner. In image representation learning, it brings the representation of different transformations of the same image closer and push the representation of transformations from different images apart [35]. Contrastive learning is a

MI (mutual information)-based approach and, in practice, contrastive learning models are usually trained by maximizing an estimation of MI between different transformations of data [36]. In the TSG task, IVG-DCL [8] adopts the contrastive learning concept in their model for the TSG task in which they apply contrastive loss on each element in the video sequence (clip-level) and on each video among all videos within a batch (video-level). In our PCL, we perform proposal-level contrastive learning which directly helps the discriminative proposal representation learning. Moreover, unlike previous methods [8, 9] which construct positive and negative set based on IOUs, we do not manually set an IoU threshold for positive and negative splitting, which can reduce the side effect of confusing proposals and enforce the proposal encoder focusing more on the query-related information.

3 Methodology

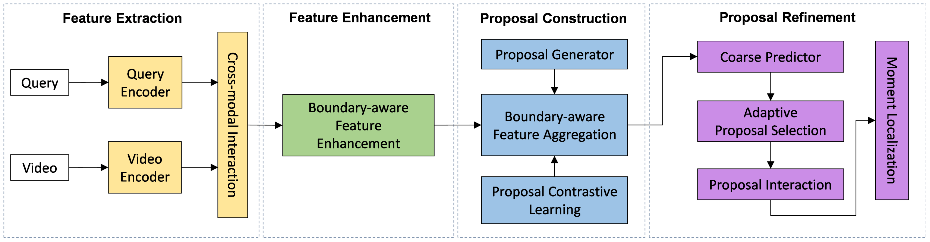


Fig. 2. An overview of our proposed model architecture for TSG.

Figure 2 depicts the overall framework of our BANet-APR, which consists of feature extraction, feature enhancement, proposal construction and proposal refinement: 1) Given a video and a query sentence, we use two encoders to generate the visual feature and the language feature, which are then fed into a cross-modal interaction module; 2) The output cross-modal feature will pass through a Boundary-aware Feature Enhancement module to strengthen the boundary information; 3) Then, we aggregate the enhanced feature and generate proposal representations by a proposal construction module; and 4) Finally, we design a Proposal Refinement module to adaptively select confident and informative proposals and refine their feature representations for the final moment localization.

3.1 Feature Extraction

Given an untrimmed video $\mathcal{V}$ and a sentence query $\mathcal{Q}$, we denote the video as $\mathcal{V} = \{v_t\}_{t=1}^{\hat{T}}$, where v_t is the t-th frame and $\hat{T}$ is the total number of frames. Similarly, each query sentence is represented by $\mathcal{Q} = \{w_i\}_{i=1}^{N}$, where w_i is the i-th word and N is the total number of words.

Feature encoding. We use a pre-trained video feature extractor, such as C3D [37] or I3D [38], to obtain video features $\widetilde{\boldsymbol{V}} = \{\widetilde{\boldsymbol{v}}_i\}_{i=1}^{T} \in \mathbb{R}^{T\times d_v}$, where $\widetilde{\boldsymbol{v}}_i$ is the i-th video feature, d_v refers to the video feature dimension, $T = \hat{T}/s$ is the total number of extracted features from the video and s is the number of frames of a video clip in C3D or I3D models. Afterwards, we feed the video feature into a stacked Bi-LSTM [39] network as the visual encoder to further aggregate its contextual information over the temporal domain as:

$$\boldsymbol{V} = \text{VisualEncoder}(\widetilde{\boldsymbol{V}}), \tag{1}$$

where $\boldsymbol{V} = \{\boldsymbol{v}_i\}_{i=1}^{T} \in \mathbb{R}^{T\times d}$ is the final encoded visual feature and d is the dimension of the visual feature.

Similarly, we use the GloVe [40] to encode each word in the sentence. The encoded feature of the input sentence is denoted as $\widetilde{\boldsymbol{S}} = \{\widetilde{\boldsymbol{s}}_i\}_{i=1}^{N} \in \mathbb{R}^{N\times d_w}$, where d_w is the GloVe embedding dimension. Similarly, we use a stacked Bi-LSTM network as the lauguage encoder to further aggregate its sequential context:

$$\boldsymbol{Q} = \text{LanguageEncoder}(\widetilde{\boldsymbol{S}}), \tag{2}$$

where $\boldsymbol{Q} = \{\boldsymbol{q}_i\}_{i=1}^{N} \in \mathbb{R}^{N\times d}$ is the final encoded query feature for N words in the query, each of which has the same feature dimension of the visual feature.

Cross-modal Interaction After getting the visual feature and the language feature, we adopt the Context-Query Attention Layer (CQA) [41] to model the interaction between the visual and the language feature. It takes the visual and the language feature as input and outputs the fused feature that provides rich cross-modal information for localization. The output will then be fed into another BiLSTM($\cdot$) to capture the sequential relationship.

$$\mathbf{F} = \text{BiLSTM}(\text{CQA}(\boldsymbol{V}, \boldsymbol{Q})) \in \mathbb{R}^{T\times d}, \tag{3}$$

where $\mathbf{F} \in \mathbb{R}^{T\times d}$ is the cross-modal representation.

3.2 Feature Enhancement

The generated cross-modal representation $\mathbf{F}$ contains both the visual and language information for the TSG task. However, it does not have the precise query-related boundary information, which is crucial for accurate moment localization. Therefore, we design a Boundary-aware Feature Enhancement (BAFE) module which consists of two branches to 1) strengthen the start and end boundary information and 2) enforce the proposal construction to pay attention on the query-related elements in cross-modal features as well. The process is shown in Figure 3. Given the cross-modal representation $\mathbf{F}$, we first pass it through two independent BiLSTMs to generate the boundary-aware feature ($\widetilde{\mathbf{F}}_b$) and the semantic content feature ($\widetilde{\mathbf{F}}_c$):

$$\widetilde{\mathbf{F}}_b = \text{BiLSTM}(\mathbf{F}), \tag{4}$$

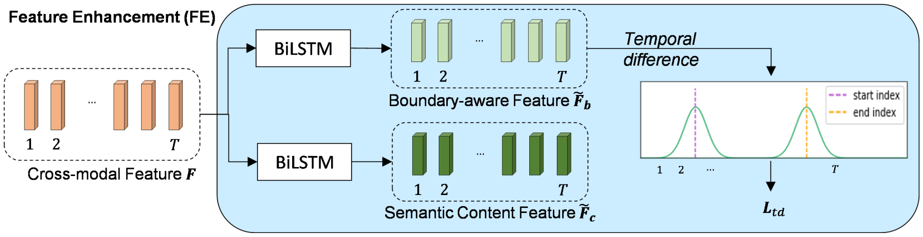


Fig. 3. Feature Enhancement. We have two branches to generate the boundary-aware feature and semantic content feature.

$$\widetilde{\mathbf{F}}_c = \mathrm{BiLSTM}(\mathbf{F}), \tag{5}$$

where $\widetilde{\mathbf{F}}_{j\in\{b,c\}} = \{\widetilde{\boldsymbol{f}}_j^i\}_{i=1}^T \in \mathbb{R}^{T\times d}$ are the enhanced cross-modal feature.

We impose a temporal difference loss on the boundary-aware feature $\widetilde{\mathbf{F}}_b$ by calculating its temporal difference:

$$\Delta\widetilde{\boldsymbol{f}}_b = \{\Delta\widetilde{f}_b^i\}_{i=1}^T, \text{ where } \Delta\widetilde{f}_b^i = \|\widetilde{\boldsymbol{f}}_b^i - \widetilde{\boldsymbol{f}}_b^{i-1}\|_2 + \|\widetilde{\boldsymbol{f}}_b^i - \widetilde{\boldsymbol{f}}_b^{i+1}\|_2 \tag{6}$$

$$\boldsymbol{d} = Softmax(\Delta\widetilde{\boldsymbol{f}}_b) \tag{7}$$

where $\boldsymbol{d} = \{d_i\}_{i=1}^T$, d_i is a scalar value representing the temporal feature difference at position i, and $\|\cdot\|_2$ refers to l_2-norm.

For each video-query pair with the start and end ground truth timestamps $\hat{t}_s$ and $\hat{t}_e$, we calculate the ground truth temporal difference regarding boundaries:

$$\tilde{d}_i = \frac{1}{\sqrt{2\pi\sigma^2}}\left(e^{-\frac{(i-\hat{t}_s)^2}{2\sigma^2}} + e^{-\frac{(i-\hat{t}_e)^2}{2\sigma^2}}\right), \tag{8}$$

where i is the position index in the temporal domain, σ is the standard deviation of a Gaussian distribution (In this paper, $\sigma = \alpha(\hat{t}_e - \hat{t}_s)$ where α is a hyper-parameter), $\hat{t}_s$ and $\hat{t}_e$ are the ground truth start and end positions. We expect the temporal difference of $\widetilde{\boldsymbol{F}}_b$ to have high activation values at the ground truth start and end boundaries as $\tilde{d}_i$. Thus, we define a temporal difference loss L_{td} as below.

$$L_{td} = -\sum_{i=1}^{T} \hat{d}_i \log(d_i) \tag{9}$$

where $\hat{d}_i = \tilde{d}_i / \sum_{i=1}^{T} \tilde{d}_i$ is the normalized ground truth temporal difference. By minimizing L_{td}, $\widetilde{\boldsymbol{F}}_b$ is forced to be a boundary-aware feature.

3.3 Proposal Construction

The general idea of proposal construction is to leverage both the discriminative boundary-aware feature and the informative semantic content feature to construct high quality proposal representations.

Proposal generator. As illustrated in Figure 4(a), in the proposal generator, the vertical and horizontal axes in the proposal generator denote the start and end indices of a proposal, respectively. Each block indicates a (start, end) pair and blocks in blue color are proposals with the valid (start, end) pairs, where the start index is smaller than the end index. The proposal generator will enumerate all possible valid pairs. Then, we densely sample short length proposals and sparsely sample longer length proposals so that proposals with large overlaps will not be selected [7].

Boundary-aware Feature Aggregation (BAFA). As shown in Figure 4(b), given a proposal with the start and end timestamps (s, e) from the proposal generator, a BAFA module is designed to aggregate the boundary information from the boundary-aware feature, $\widetilde{\mathbf{F}}_b$, and the content information from the semantic content feature, $\widetilde{\mathbf{F}}_c$, to construct the proposal feature representation.

$$\mathbf{R} = \{\boldsymbol{r}_m\}_{m=1}^{M}, \text{ where } \boldsymbol{r}_m = MLP([\bar{\boldsymbol{f}}_c^m; \widetilde{\boldsymbol{f}}_b^{s_m}; \widetilde{\boldsymbol{f}}_b^{e_m}]) \in \mathbb{R}^{1\times d}, \tag{10}$$

where $\mathbf{R} \in \mathbb{R}^{M\times d}$ are proposal features, M is the number of sampled proposals, $\widetilde{\boldsymbol{f}}_b^{s_m}$ and $\widetilde{\boldsymbol{f}}_b^{e_m}$ are elements at the start and end timestamps in the boundary-aware feature and $\bar{\boldsymbol{f}}_c^m = maxpool(\widetilde{\boldsymbol{f}}_c^{s_m} : \widetilde{\boldsymbol{f}}_c^{e_m})$ is maxpooling the features in the semantic content feature sequence from the start to the end position of the mth proposal.

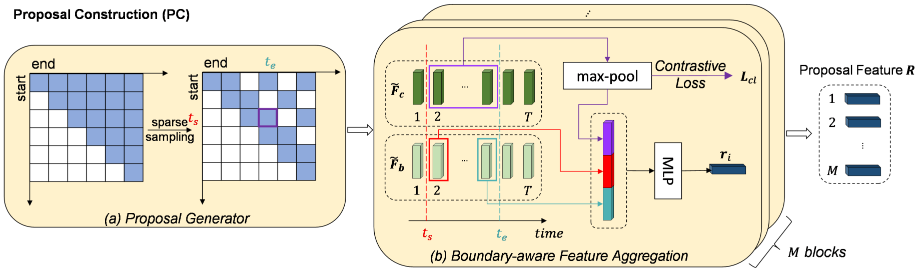


Fig. 4. (a)Proposal Constructor. We enumerate all possible proposals and sparsely sample long proposals. (b)Boundary-aware Feature Aggregation. We aggregate video clips in the boundary-aware feature and semantic content feature and pass it through a MLP layer to generate the final proposal feature.

Proposal-level contrastive learning. We impose a contrastive loss on proposal representations aggregated from the semantic content feature (shwon in Figure 4(b)), which serves as an additional supervision to guide the BAFA module to generate query-related proposals. Given all proposals P, we treat proposals that include the whole ground truth interval as positive sample set P_+, and the ones that have no overlap with the ground truth interval as negative sample set P_-. Then, we use the MIL-NCE loss [42] for our contrastive loss:

$$L_{cl} = -\log \frac{\sum\limits_{p \in P_+} e^{h_t(q)^\top h_v(p)}}{\sum\limits_{p \in P_+} e^{h_t(q)^\top h_v(p)} + \sum\limits_{p \in P_-} e^{h_t(q)^\top h_v(p)}}, \tag{11}$$

where $q = AvgPool(Q)$ is the global query feature representation and p (equivalent to the $\widetilde{\boldsymbol{f}}_c^m$ in the BAFA module) is the proposal feature generated via max pooling over the semantic content feature from video clips. Similar to the SimCLR approach [34], we add two learnable projectors $h_v(\cdot)$ and $h_t(\cdot)$ to project proposal feature and text feature to a compatible embedding spaces, respectively, and apply the contrastive loss on the projected features.

3.4 Proposal Refinement

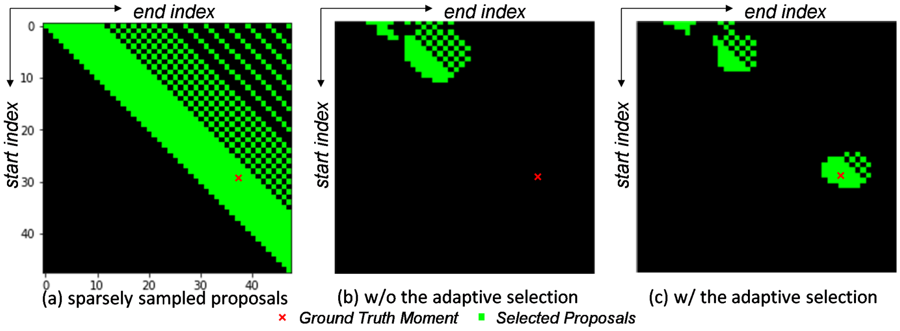
The proposal refinement module aims to effectively select a subset of proposals and refine their feature representations for the final localization.

Coarse predictor. After getting all proposals (shown in Figure 5(a)) and their corresponding proposal features, we impose a predictor to predict the matching score of each proposal regarding the query. The coarse predictor takes the proposal feature as the input and predicts the matching score of that proposal:

$$p_i = Sigmoid(MLP(\boldsymbol{r}_i)) \tag{12}$$

where $\boldsymbol{r}_i$ and p_i are the feature representation and the predicted matching score of the ith proposal, respectively.

Adaptive proposal selection. After obtaining the coarse matching scores of all proposals, we select k proposals out of them. However, instead of directly selecting proposals with the *top* k highest matching scores (shown in Figure 5(b)), we design adaptive selection strategies which consider both the global diversity and local compaction. Specifically, we first perform Non-maximum Suppression to select *top* m confident anchor proposals with small overlaps. Then, for each anchor proposal, we select *top* n confident nearby proposals that have large overlaps with the corresponding anchor proposal ($k = m \times (n+1)$). Therefore, we are able to model both the global (between anchors) and local (between an anchor and its neighbors) relationships in the next proposal interaction step. Figure 5(c) shows the effectiveness of the adaptive selection. By comparing Figure 5(b) and Figure 5(c), we can observe that using adaptive proposal selection strategies has a larger chance to include the ground truth moment in the pool of proposals for further refinement.



(a) sparsely sampled proposals

(b) w/o the adaptive selection

(c) w/ the adaptive selection

Fig. 5. Effectiveness of the adaptive proposal selection. Each cell denotes a possible proposal and the x axis and y axis represent the end and start indices of the proposal, respectively. Cells in green are selected proposals, and cells in black are not selected. The red cross represents the ground truth temporal interval.

Proposal interaction. Given a proposal with the start and end positions, we embed it with its positional information based on sine and cosine functions of different frequencies [43]. Then, we adopt an MLP to project it into a new space:

$$\hat{\boldsymbol{r}}_i = MLP([\boldsymbol{r}_i; \boldsymbol{f}_s^{pos}; \boldsymbol{f}_e^{pos}]) \tag{13}$$

where $\boldsymbol{r}_i$ is the ith proposal feature and $\boldsymbol{f}_s^{pos}$ and $\boldsymbol{f}_e^{pos}$ are the positional embeddings for the start and end positions, respectively.

Then, we adopt a Graph Neural Network (GNN) with Edge Convolution [29] to perform proposal interactions to refine proposal features where all the selected proposals are densely connected:

$$\tilde{\boldsymbol{R}} = GNN(\hat{\boldsymbol{R}}), \hat{\boldsymbol{R}} = \{\hat{\boldsymbol{r}}_i\}_{i=1}^{k} \tag{14}$$

where $\tilde{\boldsymbol{R}} = \{\tilde{\boldsymbol{r}}_i\}_{i=1}^{k}$ is the refined proposal representations.

Moment localization After getting the refined proposal feature, we impose two predictors to predict the matching score of a proposal and the boundary offsets of that proposal.

$$\hat{p}_i = Sigmoid(MLP(\tilde{\boldsymbol{r}}_i)) \tag{15}$$

$$(\Delta t_s^i, \Delta t_e^i) = MLP(\tilde{\boldsymbol{r}}_i), \tag{16}$$

where $\hat{p}_i$ is the final predicted matching score and Δt_s^i and Δt_e^i are the predicted start and end boundary offsets, respectively. The refined proposal representations enable the newly predicted matching scores in Eq.15 to be more precise than the matching scores from the coarse predictor.

We use the truncated IoU value [7] as the ground truth matching score and adopt a binary cross entropy loss for both the coarse and final matching score prediction losses L_{mc} and L_{mf}, respectively. And we use smooth $L1$ loss for the boundary offsets loss L_b:

$$L_{mc} = -\frac{1}{M}\Sigma_{i=1}^{M}(\hat{y_i}\log p_i + (1-\hat{y_i})\log(1-p_i)) \tag{17}$$

$$L_{mf} = -\frac{1}{k}\Sigma_{i=1}^{k}(\hat{y_i}\log \hat{p}_i + (1-\hat{y_i})\log(1-\hat{p}_i)) \tag{18}$$

$$L_b = SmoothL1(\Delta\hat{t}_s - \Delta t_s) + SmoothL1(\Delta\hat{t}_e - \Delta t_e) \tag{19}$$

where $\hat{y_i}$ is the ground truth matching score, k is the number of selected proposals, M is the total number of sampled proposals and $\Delta\hat{t}_s$ and $\Delta\hat{t}_e$ are the ground truth boundary offsets.

3.5 Training and Inference

The overall training objective consists of the temporal difference loss L_{td} (Eq.9), contrastive loss L_{cl} (Eq.11), matching score prediction losses L_{mc} (Eq.17) and L_{mf} (Eq.18) and boundary offset loss L_b (Eq.19):

$$L = \alpha_{td}L_{td} + \alpha_{cl}L_{cl} + \alpha_{mc}L_{mc} + \alpha_{mf}L_{mf} + \alpha_b L_b, \tag{20}$$

where α_{td}, α_{cl}, α_{mc}, α_{mf} and α_b are loss weights to balance different loss contributions and are determined by the validation set.

During inference, we pass the video sequence and query sentence into the model and get a set of proposals together with their corresponding matching scores. We then select the proposal with the highest matching score to generate the final localization.

4 Experiments

We evaluate the proposed BANet-APR method and compare to the state-of-the-art approaches on Charades-STA, ActivityNet Captions and TACoS dataset.

4.1 Dataset and Evaluation

Charades-STA. Charades-STA is built on the Charades dataset [44] for indoor activities and is extended by [11] with temporal annotation of text descriptions. In total, it has 6,672 videos and 16,128 moment-query pairs, where 12408 pairs and 3720 pairs for training and testing, respectively.
ActivityNet-Captions. ActivityNet Captions is first introduced by [12] and there are more than 20k videos. Following the given split, we use val_1 as the validation set and val_2 as the testing set, resulting in 37417, 17505, and 17031 samples for training, validation and testing, respectively.
TACoS. TACoS [13] dataset is based on 127 indoor cooking videos in around 7 minutes on average. We follow the same splits as in [11] and have 10146, 4589 and 4083 moment-query pairs for training, validation and testing.

Evaluation Metric. For a fair comparison, we adopt the recall 1 at various thresholds of the Intersection over Union, *R1@IoU=m*, following previous works [11, 45] to measure the percentage of the predicted proposals that have IoU with the ground truth annotation larger than m, where $m \in \{0.3, 0.5, 0.7\}$.

4.2 Implementation Details

For fair comparisons with state-of-the-art, we adopt the C3D [37] feature for ActivityNet and TACoS dataset, and use the I3D [38] feature for the Charades-STA dataset videos. We utilize a pre-trained Glove 840B 300d [40] to encode query sentences. In the experiment, we set the number of hidden units in Bi-LSTM to 256 and the feature dimension d to 512. In the contrastive loss, the dimension of the projected feature space is set to 128. In the adaptive proposal selection, we set the number of anchors $m = 16$ and the number of neighbors $n = 4$. The video feature sequence length for Charades-STA, ActivityNet and TACoS is set to 48, 64 and 128, respectively. The batch size is set to 32. We train our model using the Adam optimizer with a learning rate of 1×10^{-3}.

4.3 Comparison with State-of-the-arts

In Table 1 we compare our BANet-APR with recent state-of-the-art methods [7, 4, 21, 46, 24, 23, 8, 47, 30, 29]. Our model achieves the highest score in terms of *R1@IoU*=0.7 on Charades-STA and ActivityNet. In terms of *R1@IoU*=0.5, we have the highest score on Charades-STA dataset and the second best score on both ActivityNet and TACoS dataset. Although our model only achieves the second best scores on TACoS dataset, it has a large improvement compared with all other methods (except CPN [48]), and it gives much higher performance on the other two datasets compared with CPN.

4.4 Ablation Study

Effectiveness of model components. As shown in Table 2, we perform in-depth ablation studies on the effectiveness of different components in our model, including Boundary-aware Feature Enhancement (BAFE) in Feature Enhancement, Proposal-level Contrastive Learning (PCL) in Proposal Construction and Proposal Interaction (PI) and Adaptive Proposal Selection (APS) in Proposal Refinement modules based on the Charades-STA dataset. We can observe that the removal of any component will decrease the performance. Moreover, we can also observe from the last three rows that introducing a proposal interaction can help moment localization and using adaptive selection strategies can further improve the performance.

More ablations on the boundary-aware feature enhancement. Table 3 illustrates the results of different designs of the boundary-aware feature enhancement module, from which we can observe the followings. First, simply taking the

Table 1. Comparisons with other state-of-the-art methods. Bold and underline denote the best and the second best results, respectively. For a fair comparison, we only compare with models using the I3D or C3D feature.

Method	**ActivityNet Captions**		**Charades-STA**		**TACoS**	
	R1@		**R1@**		**R1@**	
	0.5	**0.7**	**0.5**	**0.7**	**0.3**	**0.5**
DRN [24]	45.45	24.36	53.09	31.75	-	23.17
LGI [23]	41.51	23.07	59.46	35.48	-	-
CPN [48]	45.10	28.10	59.77	36.67	**48.29**	**36.58**
IVG-DCL [8]	43.84	27.10	-	-	38.84	29.07
DeNet [49]	43.79	-	59.70	38.52	-	-
DRFT [45]	42.37	25.23	60.79	36.72	-	-
CBLN [47]	48.12	27.60	61.13	38.22	38.98	27.65
2D-TAN [7]	44.51	26.54	-	-	37.29	25.32
RaNet [4]	45.59	28.67	60.40	39.65	43.34	33.54
LPNet [30]	45.92	25.39	54.33	34.03	-	-
APGN [29]	**48.92**	28.64	62.58	38.86	39.34	28.34
Ours	48.12	**29.67**	**63.68**	**42.28**	48.24	33.74

max-pooling results from the semantic-content feature to construct proposal representations gives the worst result. Secondly, introducing boundary-aware feature gives better results on the higher IoU metric. Finally, the combination of them gives the best performance which verifies that they are complement to each other.

4.5 Qualitative Results

Figure 6 demonstrates how the boundary-aware feature helps the localization. The bottom row in Figure 6 displays the temporal difference values of the

Table 2. Effectiveness of different components in our model on Charades-STA dataset

Component				**R1@**	
BAFE	**PCL**	**PI**	**APS**	**0.5**	**0.7**
✗	✗	✗	✗	58.01	34.35
✓	✗	✗	✗	61.75	39.68
✓	✓	✗	✗	62.23	40.78
✓	✓	✓	✗	63.12	41.34
✓	✓	✓	✓	**63.68**	**42.28**

Table 3. Ablations of the boundary-aware feature enhancement on Charades-STA dataset. $\tilde{\mathbf{F}}_c$ is the semantic content feature and $\tilde{\mathbf{F}}_b$ is the boudary-aware feature.

Component		**R1@**	
$\tilde{\mathbf{F}}_c$	$\tilde{\mathbf{F}}_b$	**0.5**	**0.7**
✓	✗	62.35	39.07
✗	✓	61.96	40.45
✓	✓	**63.68**	**42.28**

boundary-aware feature $\tilde{\mathbf{F}}_b$ after softmax normalization. We can observe that the boundary-aware feature has high activation values around the start and end positions, aiding the temporal localization accuracy.

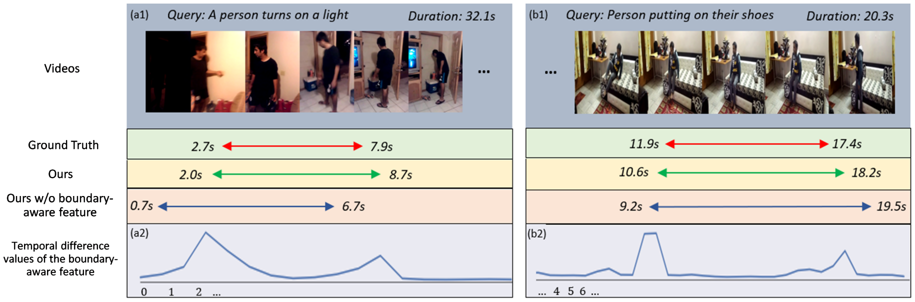


Fig. 6. Qualitative examples on the boundary-aware feature.

5 Conclusion

In this paper, we propose a novel Boundary-aware Network with Adaptive Proposal Refinement (BANet-APR) for the TSG task. Specifically, we design a Boundary-aware Feature Enhancement (BAFE) module to extract the boundary information and introduce a Boundary-aware Feature Aggregation (BAFA) module where the boundary-aware feature and the semantic content feature work together to construct discriminative and informative proposal representations. Moreover, a Proposal-level Contrastive Learning (PCL) method is proposed to enforce the semantic content feature to be query-related. Finally, we propose to adaptively select a subset of proposals and perform proposal interactions to refine their feature representations for the final localization. We conduct experiments on three benchmark datasets and show the effectiveness of our model.

Acknowledgement. Jianxiang Dong and Zhaozheng Yin have been supported by National Science Foundation via National Robotics Initiative grant CMMI-1954548 and Human Technology Frontier grant ECCS-2025929.

References

1. Zhang, H., Sun, A., Jing, W., Zhou, J.T.: Span-based localizing network for natural language video localization. arXiv preprint arXiv:2004.13931 (2020)
2. Zhu, F., Zhu, Y., Chang, X., Liang, X.: Vision-language navigation with self-supervised auxiliary reasoning tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 10012–10022
3. Huang, D., Chen, P., Zeng, R., Du, Q., Tan, M., Gan, C.: Location-aware graph convolutional networks for video question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 34. (2020) 11021–11028
4. Gao, J., Sun, X., Xu, M., Zhou, X., Ghanem, B.: Relation-aware video reading comprehension for temporal language grounding. arXiv preprint arXiv:2110.05717 (2021)
5. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. (2015) 1440–1448
6. Xu, H., Das, A., Saenko, K.: R-c3d: Region convolutional 3d network for temporal activity detection. In: Proceedings of the IEEE international conference on computer vision. (2017) 5783–5792
7. Zhang, S., Peng, H., Fu, J., Luo, J.: Learning 2d temporal adjacent networks for moment localization with natural language. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 34. (2020) 12870–12877
8. Nan, G., Qiao, R., Xiao, Y., Liu, J., Leng, S., Zhang, H., Lu, W.: Interventional video grounding with dual contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 2765–2775
9. Wang, Z., Wang, L., Wu, T., Li, T., Wu, G.: Negative sample matters: A renaissance of metric learning for temporal grounding. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 36. (2022) 2613–2623
10. Zhang, D., Dai, X., Wang, X., Wang, Y.F., Davis, L.S.: Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 1247–1257
11. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Proceedings of the IEEE international conference on computer vision. (2017) 5267–5275
12. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: Proceedings of the IEEE international conference on computer vision. (2017) 706–715
13. Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., Pinkal, M.: Grounding action descriptions in videos. Transactions of the Association for Computational Linguistics **1** (2013) 25–36
14. Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2017) 5803–5812
15. Dong, J., Li, X., Xu, C., Yang, X., Yang, G., Wang, X., Wang, M.: Dual encoding for video retrieval by text. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
16. Ge, R., Gao, J., Chen, K., Nevatia, R.: Mac: Mining activity concepts for language-based temporal localization. In: IEEE Winter Conference on Applications of Computer Vision (WACV). (2019) 245–253

17. Liu, M., Wang, X., Nie, L., He, X., Chen, B., Chua, T.S.: Attentive moment retrieval in videos. In: Proceedings of the 41nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). (2018) 15–24
18. Chen, J., Chen, X., Ma, L., Jie, Z., Chua, T.S.: Temporally grounding natural sentence in video. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). (2018) 162–171
19. Zhang, Z., Lin, Z., Zhao, Z., Xiao, Z.: Cross-modal interaction networks for query-based moment retrieval in videos. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR). (2019) 655–664
20. Liu, M., Wang, X., Nie, L., Tian, Q., Chen, B., Chua, T.S.: Cross-modal moment localization in videos. In: Proceedings of the 26th ACM international conference on Multimedia. (2018) 843–851
21. Yuan, Y., Ma, L., Wang, J., Liu, W., Zhu, W.: Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In: Advances in Neural Information Processing Systems (NIPS). (2019) 534–544
22. Xu, H., He, K., Plummer, B.A., Sigal, L., Sclaroff, S., Saenko, K.: Multilevel language and vision integration for text-to-clip retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 33. (2019) 9062–9069
23. Mun, J., Cho, M., Han, B.: Local-global video-text interactions for temporal grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 10810–10819
24. Zeng, R., Xu, H., Huang, W., Chen, P., Tan, M., Gan, C.: Dense regression network for video grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2020) 10287–10296
25. Cao, M., Chen, L., Shou, M.Z., Zhang, C., Zou, Y.: On pursuit of designing multi-modal transformer for video grounding. arXiv preprint arXiv:2109.06085 (2021)
26. Zhang, M., Yang, Y., Chen, X., Ji, Y., Xu, X., Li, J., Shen, H.T.: Multi-stage aggregated transformer network for temporal language localization in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 12669–12678
27. Wang, J., Ma, L., Jiang, W.: Temporally grounding language queries in videos by contextual boundary-aware prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 34. (2020) 12168–12175
28. Xu, M., Pérez-Rúa, J.M., Escorcia, V., Martinez, B., Zhu, X., Zhang, L., Ghanem, B., Xiang, T.: Boundary-sensitive pre-training for temporal localization in videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 7220–7230
29. Liu, D., Qu, X., Dong, J., Zhou, P.: Adaptive proposal generation network for temporal sentence localization in videos. arXiv preprint arXiv:2109.06398 (2021)
30. Xiao, S., Chen, L., Shao, J., Zhuang, Y., Xiao, J.: Natural language video localization with learnable moment proposals. arXiv preprint arXiv:2109.10678 (2021)
31. Van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv e-prints (2018) arXiv–1807
32. Misra, I., Maaten, L.v.d.: Self-supervised learning of pretext-invariant representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 6707–6717
33. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 3733–3742

34. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning, PMLR (2020) 1597–1607
35. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems **33** (2020) 21271–21284
36. Gupta, T., Vahdat, A., Chechik, G., Yang, X., Kautz, J., Hoiem, D.: Contrastive learning for weakly supervised phrase grounding. In: European Conference on Computer Vision, Springer (2020) 752–768
37. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2015) 4489–4497
38. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017) 6299–6308
39. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE transactions on Signal Processing **45** (1997) 2673–2681
40. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). (2014) 1532–1543
41. Yu, A.W., Dohan, D., Luong, M.T., Zhao, R., Chen, K., Norouzi, M., Le, Q.V.: Qanet: Combining local convolution with global self-attention for reading comprehension. arXiv preprint arXiv:1804.09541 (2018)
42. Miech, A., Alayrac, J.B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-end learning of visual representations from uncurated instructional videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 9879–9889
43. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
44. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: European Conference on Computer Vision, Springer (2016) 510–526
45. Chen, Y.W., Tsai, Y.H., Yang, M.H.: End-to-end multi-modal video temporal grounding. Advances in Neural Information Processing Systems **34** (2021)
46. Rodriguez, C., Marrese-Taylor, E., Saleh, F.S., Li, H., Gould, S.: Proposal-free temporal moment localization of a natural-language query in video using guided attention. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. (2020) 2464–2473
47. Liu, D., Qu, X., Dong, J., Zhou, P., Cheng, Y., Wei, W., Xu, Z., Xie, Y.: Context-aware biaffine localizing network for temporal sentence grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 11235–11244
48. Zhao, Y., Zhao, Z., Zhang, Z., Lin, Z.: Cascaded prediction network via segment tree for temporal video grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 4197–4206
49. Zhou, H., Zhang, C., Luo, Y., Chen, Y., Hu, C.: Embracing uncertainty: Decoupling and de-bias for robust temporal grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 8445–8454