Robust Temporal Sentence Video Grounding with Global Proposal Ranking

Anonymous ACL submission

Abstract

Most existing solutions to temporal sentence video grounding (TSGV) rely heavily on local classifiers to discern start and end boundaries, often compromise internal consistency and overlook boundary uncertainty. This paper introduces a novel global ranking approach that directly scores all candidate proposals using a unique loss function, thereby enhancing robustness through the integrated decoding of local and global predictions. We further incorporate pretrained language models into our framework - a largely underexplored facet in TSGV. Our methodology is evaluated across three distinct settings: distribution-consistent, distribution-changing, and composition generalization datasets, outperforming existing baselines across the board. Notably, it exhibits superior performance in out-of-distribution and composition generalization tasks. To the best of our knowledge, we are the first to combine global proposal ranking and pretrained language models for robust TSVG.

1 Introduction

Temporal Sentence Grounding in Videos (TSGV) first introduced by Gao et al. (2017), is an essential bridge between textual and visual understanding, promising significant advancements in video comprehension and interaction. TSVG aims to locate specific moments within untrimmed videos using a language query. An example is illustrated in Figure 1, where a given query "person takes a photo from the table" is used to identify the corresponding moment (12.5 to 17.7 seconds) within a 26.96-second untrimmed video.

Existing approaches to this problem primarily fall into two categories: proposal-free and proposal-based methods. Proposal-free methods (Chen et al., 2018; Ghosh et al., 2019; Zeng et al., 2020; Zhang et al., 2020a; Cao et al., 2020a,b; Li et al., 2021; Liu et al., 2021; Zhou et al., 2021; Nan et al., 2021; Xu et al., 2021) focus on determining the start and end points of the target moment, making them simpler to train but more prone to biases due to annotation uncertainties. Proposal-based methods (Gao et al., 2017; Anne Hendricks et al., 2017; Ge et al., 2019; Zhang et al., 2019, 2020b, 2021b; Zheng et al., 2022; Li et al., 2023a), on the other hand, generate candidate proposals through the aggregation of video frames and alignment with the query sentences, taking into account the interaction of text and the entire proposal. While effective, these methods heavily depend on the quality of proposal generators and the efficiency of the ranking mechanism.

Figure 1: An example of temporal sentence grounding in videos.

To combine the advantages of both categories, hybrid methods (Wang et al., 2020, 2021a; Xiao et al., 2021; Huang et al., 2022) have been introduced that combine both the segment-level and frame-level information for improved video comprehension. Nevertheless, most of these techniques generate proposals using sampled moments, failing to consider all potential moments.

The utilization of pretrained language models (PLMs) such as BERT (Devlin et al., 2019) and Roberta (Liu et al., 2019) within the TSVG realm is another underexplored area. Despite some efforts (Wang et al., 2021b; Zheng et al., 2023; Shimomoto et al., 2023) to incorporate PLMs, the results have varied, indicating the need for a more harmonious integration approach.

In this work, we put forward a novel approach, GPRank, designed to tackle these existing challenges. Our strategy employs a global ranking loss function, enabling a comprehensive consideration of all candidate proposals. This method integrates a global perspective into the ranking of all candidate moments, ensuring a more precise and comprehensive ranking process. Moreover, we design a backbone-specific integration strategy to facilitate better interaction between pretrained text features and video features. By carefully orchestrating this integration process, our method aims to fully harness the potential of PLMs, thereby effectively aligning textual queries with their corresponding video moments.

We undertake extensive evaluations across three distinct scenarios: 1) distribution-consistent benchmarks, including ActivityNet-Captions (Krishna et al., 2017), Charades-STA (Gao et al., 2017), and TACoS (Regneri et al., 2013); 2) distribution-changing datasets, containing ActivityNet-CD and Charades-CD (Yuan et al., 2021); and 3) composition generalization datasets, including ActivityNet-CG and Charades-CG (Li et al., 2022b). Experimental outcomes demonstrate that our approach outperforms multiple strong baselines across all settings, with distinct effectiveness in out-of-distribution and composition generalization tasks.

Our contributions in this paper are two-fold: we introduce a new global span ranking method, and we present a novel way of integrating pretrained language models with video features for TSVG. The source code of our approach will be made publicly available.

2 Global Proposal Ranking

Temporal sentence video grounding identifies a specific moment in an untrimmed video, denoted as $\boldsymbol{V}$ with K frames, using a natural language sentence $\boldsymbol{Q}$ with L words. The moment is defined by start (S) and end (E) points with $S, E \in [1, K]$. Instead of learning a direct mapping function $f : f(\boldsymbol{V}, \boldsymbol{Q}) \rightarrow [S, E]$, we model a score $g_{i,j}$ ($1 \leq i \leq j \leq K$) for each moment candidate, with a higher $g_{i,j}$ indicating better correspondence between $\boldsymbol{V}$ and $\boldsymbol{Q}$. The ground-truth overlapping score $\mu_{i,j}$ of a candidate proposal $[i, j]$ with respect to the answer moment $[S, E]$ is defined via the intersection-over-union (IoU) score: $\mu_{i,j} = \frac{[i,j] \cap [S,E]}{[i,j] \cup [S,E]}$.

The goal of global proposal ranking is to learn a model that predicts $g_{i,j}$ to align with $\mu_{i,j}$ globally, maintaining the partial order relation: when $\mu_{i,j} \geq \mu_{i',j'}$, then $g_{i,j} \geq g_{i',j'}$. We utilize a loss function from Su et al. (2022), originally designed for multi-label classification[1], which is defined as:

$$\begin{aligned}\mathcal{L}_{span} &= \log\Big(1 + \sum_{i \leq j} \mu_{i,j} \exp(-g_{i,j})\Big) \\ &+ \log\Big(1 + \sum_{i \leq j} (1.0 - \mu_{i,j}) \exp(g_{i,j})\Big)\end{aligned} \tag{1}$$

To minimize the loss function above, the model needs to increase the value of $g_{i,j}$ when $\mu_{i,j}$ is large and decrease the value of $g_{i,j}$ when $\mu_{i,j}$ is small. When $\mu_{i,j}$ becomes a binary variable, the loss function is identical to circle loss (Sun et al., 2020). As shown in the appendix, the loss function reaches a local minimum when

$$\hat{\mu}_{i,j} = \sigma(2g_{i,j}), \tag{2}$$

where σ denotes the sigmoid function. This indicates that for prediction, the probability of the span $[i, j]$ being the target span can be approximated using $\sigma(2g_{i,j})$.

We then propose a combination of predictions from both local boundary classifiers and the global span ranking module. Local boundary classifiers learn two mapping functions, $f_s : f_s(\boldsymbol{V}, \boldsymbol{Q}) \rightarrow S$ and $f_e : f_e(\boldsymbol{V}, \boldsymbol{Q}) \rightarrow E$, for recognizing the start and the end points, respectively. The span score obtained from the local boundary classifier is defined as:

$$l_{i,j} = P_s(i) \times P_e(j) \tag{3}$$

where $P_s(i)$ and $P_e(j)$ indicate the probabilities of i and j being the start and end points, respectively. A hyper-parameter λ ($0 \leq \lambda \leq 1$) balances the span scores from the local boundary classifier and the global span ranking module:

$$s_{i,j} = \lambda \log \sigma(2g_{i,j}) + (1.0 - \lambda) \log l_{i,j}. \tag{4}$$

The final answer of the target moment is given by:

$$\hat{S}, \hat{E} = \underset{1 \leq i \leq j \leq K}{\arg\max}\, s_{i,j}. \tag{5}$$

3 Model Architecture

In this section, we describe the specific model to obtain the global ranking score $g_{i,j}$ and the local span classification score $l_{i,j}$ defined in Eq. 3. We highlight specific designs to integrate the pretrained text features from Roberta and modules for the global loss calculation. Figure 2 shows the overview of our model architecture. It mainly contains three stages: InputFusion, ContentFusion and PredictionFusion. The InputFusion stage is responsible

[1]https://spaces.ac.cn/archives/9064

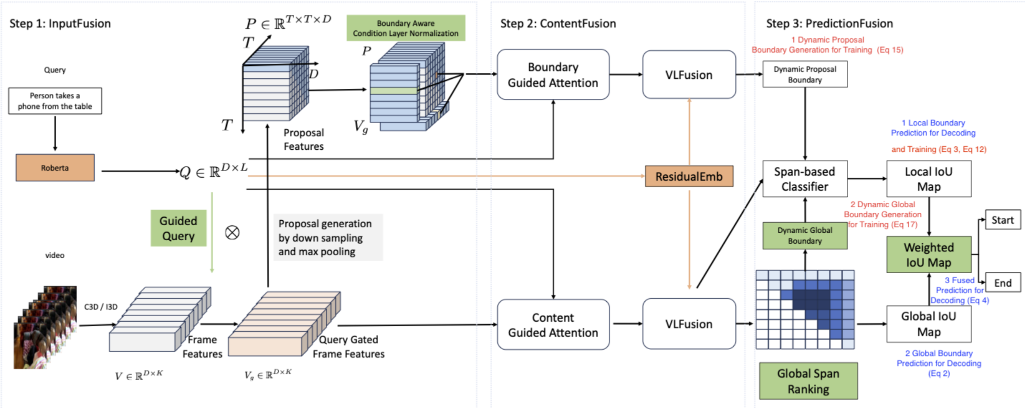

Figure 2: The overview of GPRank architecture.

for fuse pretrained contextual embeddings into the video features using *GuidedQuery*, generate proposals, and inject boundary information to proposals using *condition layer normalization*. The ContentFusion stage captures the interaction among the pretrained contexutal text features, video features and proposal features using guided attention. Particularly, in the fusion stage, a *ResidualEmb* module is used to directly emphasize the input pretrained text features before video-language fusion. The PredictionFusion stage combines the dynamic boundaries from the proposal module and the *global proposal ranking* module to train the span-based classifier. In the end, it combines the local predictions and global predictions to make the final decisions.

3.1 Input Fusion

Video frames are encoded using a pretrained 3D-CNN model (Carreira and Zisserman, 2017), producing $\boldsymbol{V} = \{\boldsymbol{f}_t\}_{t=1}^{K} \in \mathbb{R}^{D_v \times K}$. Queries use the Roberta-base encoder (Liu et al., 2019). After Roberta tokenizes into sub-words, their representations are averaged for word-level representations: $\boldsymbol{Q} = \{\boldsymbol{q}_l\}_{l=1}^{L} \in \mathbb{R}^{D_q \times L}$. The `[CLS]` token's vector is $\mathbf{q}_{[cls]}$. For cross-modal interactions, three linear projections map video and sentence representations to space D, giving $\boldsymbol{V} = \mathrm{FC}(\boldsymbol{V}) \in \mathbb{R}^{D \times K}$, $\boldsymbol{Q} = \mathrm{FC}(\boldsymbol{Q}) \in \mathbb{R}^{D \times L}$, and $\mathbf{q}_{[cls]} = FC(\mathbf{q}_{[cls]}) \in \mathbb{R}^{D}$.

GuidedQuery The video frame features are then used to generate moment proposals. To make these dependent on the powerful pre-trained text features, a `GuidedQuery` module is designed to assign higher weights to the input video frames that share higher similarity with the input query. Specifically, this is achieved by using the global semantic vector of the input query via a gating function,

$$\boldsymbol{V}_g = \sigma(\mathrm{repeat}(\mathbf{q}_{[cls]}, K) \otimes \boldsymbol{V}) \otimes \boldsymbol{V}, \tag{6}$$

where the `repeat` function repeats an input vector multiple times and $\otimes$ denotes Hadamard product.

Local Proposal Generation A 2D feature map is generated as in (Zhang et al., 2020b), by enumerating all pairwise start-end frames, yielding $K = T \times T$ video segments. With $T = \lfloor K/m \rfloor$ and m as the downsampling rate, these segments, $\boldsymbol{P} = \{\boldsymbol{v}_n^p\}_{n=1}^{N} \in \mathbb{R}^{D \times N}$, serve as moment proposals. We simplify by flattening this map. Each segment proposal uses a max-pooled feature vector from its frames. The k-th proposal with boundaries (t_k^s, t_k^e) is:

$$\boldsymbol{P}_k = \mathrm{MaxPool}(\boldsymbol{V}_g[t] | \forall t \in [t_k^s, t_k^e]), \tag{7}$$

where MaxPool is max-pooling over the proposal's frame range, capturing its key features.

Conditional Layer Normalization Downsampling of m frames can lose boundary information vital for temporal endpoints. To mitigate this, we use a conditional layer normalization layer (Chen et al., 2021; Li et al., 2022a) to infuse boundary features from $\boldsymbol{V}_g$ to $\boldsymbol{P}$. For segment $[i, j]$, the normalized representation is:

$$\boldsymbol{P}_{i,j} = \mathrm{CLN}(\boldsymbol{P}_{i,j}, \boldsymbol{V}_g[i] \oplus \boldsymbol{V}_g[j]) = \gamma_{i,j} \otimes (\frac{\boldsymbol{P}_{i,j} - \mu}{\sigma}) + \lambda_{i,j} \tag{8}$$

Here, $\oplus$ is concatenation. The gain and bias parameters, $\gamma_{i,j}$ and $\lambda_{i,j}$, are conditioned on $\boldsymbol{V}_g[i]$ and $\boldsymbol{V}_g[j]$:

$$\gamma_{i,j} = \mathbf{w}_\alpha[\boldsymbol{V}_g[i] \oplus \boldsymbol{V}_g[j]] + b_\alpha, \lambda_{i,j} = \mathbf{w}_\beta[\boldsymbol{V}_g[i] \oplus \boldsymbol{V}_g[j]] + b_\beta \tag{9}$$

μ and σ represent the mean and standard deviation of $\boldsymbol{P}_{i,j}$ elements:

$$\mu = \frac{1}{D}\sum_{k=1}^{D} \boldsymbol{P}_{i,j,k}, \sigma = \sqrt{\frac{1}{D}\sum_{k=1}^{D} (\boldsymbol{P}_{i,j,k} - \mu)^2} \tag{10}$$

with $\boldsymbol{P}_{i,j,k}$ as the k-th dimension of $\boldsymbol{P}_{i,j}$.

3.2 Prediction Fusion

In this section, we consider the boundary prediction from a span-based classifier, the proposal-based prediction from the local proposal module, and the ranking based prediction from the global proposal ranking module. Except for prediction score interpolation during inference as shown in Eq 4, the boundary predictions from the proposal ranking model can be beneficial for training the span-based local classification models by constructing dynamic boundary sets.

Local Boundary Prediction. Given video ($\boldsymbol{V} \in \mathbb{R}^{D\times K}$) and sentence ($\boldsymbol{Q} \in \mathbb{R}^{D\times L}$) representations, we estimate frame-wise endpoint probabilities by computing context-query attention, following the approach of previous work (Seo et al., 2016; Xiong et al., 2016; Yu et al., 2018; Zhang et al., 2020a). The endpoint probabilities are given by:

$$\begin{gathered}\mathbf{h}_s, \mathbf{h}_e = \text{LSTM}(\hat{\boldsymbol{V}} \otimes \boldsymbol{h}) \\ (\boldsymbol{T}_s, \boldsymbol{T}_e) = \text{Softmax}(FC(\mathbf{h}_s)), \text{Softmax}(FC(\mathbf{h}_e)), \\ \text{where } \boldsymbol{h} = \sigma(\text{Conv1d}(\hat{\boldsymbol{V}} \| \boldsymbol{q})) \in \mathbb{R}^{1\times T}, \\ \hat{\boldsymbol{V}} = H(\boldsymbol{V}, \boldsymbol{Q}) = \\ \text{FC}(\boldsymbol{V} \| \boldsymbol{X}^{v2q} \| \boldsymbol{F} \otimes \boldsymbol{X}^{v2q} \| \boldsymbol{F} \otimes \boldsymbol{X}^{q2v}) \in \mathbb{R}^{D\times T};\ \text{and} \\ \boldsymbol{A} = \frac{\text{FC}(\boldsymbol{F})^{\top}\,\text{FC}(\boldsymbol{Q})}{\sqrt{D}}, \boldsymbol{X}^{v2q} = \boldsymbol{Q}\boldsymbol{A}^{r\top}, \ \boldsymbol{X}^{q2v} = \boldsymbol{F}\boldsymbol{A}^{r}\boldsymbol{A}^{c\top}\end{gathered} \tag{11}$$

In Eq. (11), frame-wise endpoint probabilities $\boldsymbol{T}_s$, $\boldsymbol{T}_e$ are predicted using a two-layer LSTM network, which are the probability distributions used in Eq 3. This LSTM network operates on a fusion of the video and sentence modalities, achieved through function H, and per-frame fused feature $\hat{\boldsymbol{V}} \in \mathbb{R}^{D\times K}$ is then rescaled using an estimated likelihood $\boldsymbol{h} \in \mathbb{R}^{1\times T}$ of being foreground.

Matrix $\boldsymbol{A} \in \mathbb{R}^{K\times L}$ contains frame-to-word correlation scores. $\boldsymbol{A}^r$ and $\boldsymbol{A}^c$ are row and column-wise softmax normalised versions of this matrix. The sentence-level representations $\boldsymbol{q}$ are obtained via a weighted sum of words (Bahdanau et al., 2015). The symbols $(\cdot\|\cdot)$ denotes concatenation.

After generating $\boldsymbol{T}_s$ and $\boldsymbol{T}_e$, we predict a specific boundary encompassed by a single start and end frame, based on the outputs of the bounding branch. This is done in a maximum likelihood manner:

$$\hat{S} = \arg\max_t \boldsymbol{T}_s, \quad \hat{E} = \arg\max_t \boldsymbol{T}_e, \tag{12}$$

Here, $\hat{S}$ and $\hat{E}$ represent the predicted start and end frame indices of a video that correspond to the given query.

Dynamic Proposal Boundary Prediction. The proposal-level video representations $\boldsymbol{P}$ are fused with the sentence features using the function H, as defined in Eq. (11). This fused representation is rearranged into a 2D feature map, and per-proposal alignment scores are then predicted using a 2D convolution layer:

$$\boldsymbol{p}^a = \sigma(\text{Conv2d}(H(\boldsymbol{V}, \boldsymbol{Q})))\ s.t.\ p^a_k \in (0,1)\ \forall k \in [1, N]. \tag{13}$$

Here, σ activates the segment-wise alignment scores $\boldsymbol{p}^a$. These scores are supervised based on the temporal overlap between each proposal and the manual boundary:

$$\begin{gathered}\alpha_k = \text{IoU}((t^s_k, t^e_k), (S, E)) \\ y^a_k = \begin{cases} 1, & \text{if } \alpha_k \geq \tau_u \\ 0, & \text{if } \alpha_k < \tau_l \\ \alpha_k, & \text{otherwise} \end{cases} \\ \mathcal{L}_{\text{align}}(\boldsymbol{V}, \boldsymbol{Q}, S, E) = \text{BCE}(\boldsymbol{y}^a, \boldsymbol{p}^a).\end{gathered} \tag{14}$$

In the above equation, τ_u and τ_l represent the upper and lower overlap thresholds, which regulate the flexibility of video-text alignment. These are set to 0.7 and 0.3 respectively, as in (Zhang et al., 2020b).

Given the learned segment-wise alignment scores $\boldsymbol{p}^a$, the boundary (t^s_k, t^e_k) of the most confident proposal, with the highest predicted score $p^a_{k*} \geq p^a_k\ \forall k \in [1, N]$, is considered as the pseudo boundary. From this, we construct the candidate endpoint sets as:

$$\begin{aligned}\tilde{\boldsymbol{S}} &= [\min(t^s_{k*}, S), \max(t^s_{k*}, S)], \\ \tilde{\boldsymbol{E}} &= [\min(t^e_{k*}, E), \max(t^e_{k*}, E)].\end{aligned} \tag{15}$$

We customize the candidate endpoint sets for each individual activity by exploring content alignments between video segments and query sentences, thereby creating an dynamic boundary.

Dynamic Global Boundary Prediction To compute the global ranking score $s_{i,j}$ as defined in Eq.4, we leverage the vector $\mathbf{h}_e$ as defined in Eq.11. We first split this vector into two equal parts, represented as $\mathbf{z}^s$ and $\mathbf{z}^e$, using the chunk function:

$$\mathbf{z}^s, \mathbf{z}^e = \text{chunk}(\mathbf{h}_e, 2), \ s_{i,j} = FC(\mathbf{z}^s_i)FC(\mathbf{z}^e_j)^T \tag{16}$$

Here $\text{chunk}(\mathbf{h}_e, 2)$ splits $\mathbf{h}_e$ to two equal parts. This calculation allows $s_{i,j}$ to effectively utilize the information encapsulated in $\mathbf{h}_e$, which is also used in Eq. 11 for end point prediction, thereby enabling effective scoring of spans.

Given $s_{i,j}$, we can select the optimal moment by maximizing $s_{i,j}$ over all valid intervals $1 \leq i \leq j \leq K$: Given $s_{i,j}$, we can select the optimal moment by $\hat{S}_g, \hat{E}_g = \arg\max_{1\leq i\leq j\leq K} s_{i,j}$. Based on this prediction, we update the candidate endpoint sets as:

$$\tilde{\boldsymbol{S}} = [\min(\hat{S}_g, \tilde{\boldsymbol{S}}), \max(\hat{S}_g, \tilde{\boldsymbol{S}})],$$
$$\tilde{\boldsymbol{E}} = [\min(\hat{E}_g, \tilde{\boldsymbol{E}}), \max(\hat{E}_g, \tilde{\boldsymbol{E}})]. \tag{17}$$

These updated sets $\tilde{\boldsymbol{S}}$ and $\tilde{\boldsymbol{E}}$ represent the final dynamic boundary that we use to train the span-based local classifier.

Learning from dynamic boundary. Using the dynamic boundary $(\tilde{\boldsymbol{S}}, \tilde{\boldsymbol{E}})$, we create boundary supervisions to maximize the candidate endpoint probabilities in Eq. (18):

$$\mathcal{L}_{\text{bound}}(\boldsymbol{F}, \boldsymbol{Q}, \tilde{\boldsymbol{S}}, \tilde{\boldsymbol{E}}) = -\log(\sum_{t \in \tilde{\boldsymbol{S}}} p_t^s) - \log(\sum_{t \in \tilde{\boldsymbol{E}}} p_t^e). \tag{18}$$

Unlike the common frame-wise supervision that trains $\boldsymbol{p}^s$ and $\boldsymbol{p}^e$ to be one-hot (Zhang et al., 2020a), this method provides a flexible boundary for target moments. It lets the model discern endpoints beyond manual boundaries and ignore irrelevant actions, enhancing learning of the video's temporal structure relative to the query.

3.3 The Overall Training Loss

Besides the global proposal ranking loss $\mathcal{L}_{\text{span}}$, boundary loss $\mathcal{L}_{\text{bound}}$, and alignment loss $\mathcal{L}_{\text{align}}$, we incorporate a highlight loss (Zhang et al., 2020a) to train $\boldsymbol{h}$ as in Eq. (11). This loss emphasizes key video content:

$$\mathcal{L}_{\text{high}}(\boldsymbol{F}, \boldsymbol{Q}, S, E) = \text{BCE}(\boldsymbol{y}^h, \boldsymbol{h}),$$
$$y_t^h = \mathbb{1}[\min(\tilde{\boldsymbol{S}}) \le t \le \max(\tilde{\boldsymbol{E}})]. \tag{19}$$

The overall loss function for our proposed model is then formulated as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{bound}} + \lambda_2 \mathcal{L}_{\text{align}} + \lambda_3 \mathcal{L}_{\text{high}} + \lambda_4 \mathcal{L}_{\text{span}} \tag{20}$$

The model is trained to minimize the weighted sum loss. The weighting factors $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ allow for adjusting the relative importance of each loss term, providing flexibility to tune the model based on specific performance objectives.

4 Experiments

4.1 Data and Settings

We evaluate the robustness of GPRank on three settings. The first is the standard setting, including TACoS (Regneri et al., 2013; Rohrbach et al., 2012), Charades-STA (Gao et al., 2017; Sigurdsson et al., 2016) and ActivityNet-Captions (Krishna et al., 2017; Heilbron et al., 2015). Table 1 in the appendix shows the data statistics of the standard setting. The second is the distribution changing setting including Charades-CD and ActivityNet-CD (Yuan et al., 2021). The performance of the models is evaluated in both in-distribution (test-iid) and out-of-distribution (test-ood) scenarios. The third is the compositional generalization setting, including Charades-CG and ActivityNet-CG (Li et al., 2022b). The evaluations span three settings: Test-Trivial, Novel-Composition, and Novel-Word, with the last two involving unseen semantic composition and words outside the training set.

Evaluation Metrics. We follow common practices in the field (Zhang et al., 2020a; Wang et al., 2021a; Nan et al., 2021) and measures the average recall rate at three temporal IoU thresholds (IoU@μ) for $\mu = 0.3$, $\mu = 0.5$, and $\mu = 0.7$. A higher IoU indicates a better performance. We also report the mean IoU (mIoU), a metric representing the average overlap between the predicted and the ground-truth boundaries.

Implementation Details. The implementation details are described in the appendix.

4.2 Main Results

Table 1 shows the main results on the standard distribution-consistent setting. Various baselines listed in the table include VSLNet (Zhang et al., 2020a), IVG (Nan et al., 2021), 2D-LGI (Mun et al., 2020), BPNet (Xiao et al., 2021), EBM (Huang et al., 2022) and MS-DETR (Wang et al., 2023). GPRank outperforms all the other models in terms of mIoU across all three datasets. Specifically, GPRank achieved the highest mIoU of 37.93, 54.39, and 47.30 on the TACoS, Charades-STA, and ActivityNet-Captions datasets, respectively. When IoU=0.7, MS-DETR outperforms the others on the TACoS and ActivityNet-Captions datasets, while GPRank retains the top spot in the Charades-STA dataset. In the lower IoU thresholds (0.3 and 0.5), GPRank again excels across all three datasets. In fact, for IoU=0.3, GPRank significantly outperforms the baselines, with gains over MS-DETR of 6.48, 6.08, and 4.16 points on TACoS, Charades-STA, and ActivityNet-Captions, respectively. This advantage of GPRank can be attributed to its approach of ranking all the proposals from a global perspective, which enables it to better model the long-tailed low IoU proposals.

4.3 Out-of-domain Generalization

Table 2 shows the out-of-domain generalization results. We ran the source code of EMB (Huang et al.,

Method	TACoS				Charades-STA				ActivityNet-Captions			
	mIoU	μ =0.3	μ =0.5	μ =0.7	mIoU	μ =0.3	μ =0.5	μ =0.7	mIoU	μ =0.3	μ =0.5	μ =0.7
VSLNet (Zhang et al., 2020a)	24.11	29.61	24.27	20.03	45.15	64.30	47.31	30.19	43.19	63.16	43.22	26.16
IVG (Nan et al., 2021)	28.26	38.84	29.07	19.05	48.02	67.63	50.24	32.88	44.21	63.22	43.83	27.10
LGI (Mun et al., 2020)	-	-	-	-	51.38	72.96	59.46	35.48	41.13	58.52	41.51	23.07
BPNet (Xiao et al., 2021)	19.53	25.93	20.96	14.08	46.34	65.48	50.75	31.64	42.11	58.98	42.07	24.69
EMB (Huang et al., 2022)	35.49	50.46	37.82	22.54	53.09	72.50	58.33	39.25	45.59	64.13	44.81	26.07
MS-DETR (Wang et al., 2023)	35.09	47.66	37.36	**25.81**	50.12	68.68	57.72	37.40	46.82	62.12	**48.69**	**31.15**
GPRank (ours)	**37.93**	**54.14**	**38.42**	24.12	**54.39**	**74.76**	60.78	**40.86**	**47.30**	**66.28**	46.68	27.84

Table 1: Performances on distribution-consistent settings. μ: IoU. Here we only include baselines which report $\mu = \{0.3, 0.5, 0.7\}$ and mIoU at the same time for fair comparisons. More baselines are shown in the Appendix. Bold and underline denotes the best and the second best in a column, respectively.

	Charades-CD				ActivityNet-CD			
	test-iid		test-ood		test-iid		test-ood	
	μ=0.5	μ=0.7	μ=0.5	μ=0.7	μ=0.5	μ=0.7	μ=0.5	μ=0.7
Bias-based	16.87	9.34	5.04	2.21	19.81	12.27	0.26	0.11
PredictAll	0.00	0.00	0.06	0.00	20.05	12.45	0.00	0.00
CTRL	29.80	11.86	30.73	11.97	11.27	4.29	7.89	2.53
ACRN	31.77	12.93	30.03	11.89	11.57	4.41	7.58	2.48
ABLR	41.13	23.50	31.57	11.38	35.45	20.57	20.88	10.03
2D-TAN	46.48	28.76	28.18	13.73	40.87	28.95	18.86	9.77
SCDM	47.36	30.79	41.60	22.22	35.15	22.04	19.14	9.31
DRN	41.91	26.74	30.43	15.91	39.27	25.71	25.15	14.33
TSP-PRL	35.43	17.01	19.37	6.20	33.93	19.50	16.63	7.43
WSSL	14.06	4.27	23.67	8.27	17.20	6.16	7.17	1.82
TCN-DCM	52.50	35.28	40.51	21.02	42.15	29.69	20.86	11.07
MDD	52.78	34.71	40.39	22.70	43.63	31.44	20.80	11.66
DD+MD	55.66	38.87	40.88	28.11	50.37	32.70	25.05	**14.67**
Shuffle	57.59	37.79	46.67	27.08	48.07	32.15	24.57	13.21
EMB†	62.33	43.14	48.68	30.02	48.80	31.27	21.80	10.63
GPRank†	**64.52**	**44.47**	**54.87**	**34.55**	**52.99**	**35.03**	**28.44**	13.95

Table 2: Performance comparisons on Charades-CD and ActivityNet-CD. † denotes our implementation.

2022) on these benchmarks, showing commendable performance compared to all the other baselines as shown in Table 2. GPRank shows the best performance overall. It achieved the highest scores in both datasets, across both in-distribution and out-of-distribution tests. Specifically, in the out-of-distribution tests, GPRank scored 54.87 (IoU=0.5) and 34.55 (IoU=0.7) on Charades-CD, which outperforms the baselines by large margins. GPRank is slightly worse than DD+MD (Zhang et al., 2021a) when looking at IoU=0.7 on ActivityNet-CD, which is reasonable since DD+MD applies video data augmentation techniques, which are not considered in our method. These results suggest that the GPRank model performs well under varying distributions, effectively grounding videos in both familiar (in-distribution) and unfamiliar (out-of-distribution) scenarios.

4.4 Compositional Generalization

Table 3 presents the performance of various temporal grounding methods on on ActivityNet-CG. As shown in Table 3, our GPRank method achieves the highest score in 8 out of 9 metrics, while taking the second position in the remaining one. Compared to MS-2D-TAN+SSL (Li et al., 2023a), GPRank boasts of mIOU improvements of +3.31, +4.54, and +4.50 across the Test-Trivial, Novel-Composition, and Novel-Word settings, respectively. When compared to VISA+ASSL (Li et al., 2023b), the mIOU performances of GPRank on the Novel-Composition and Novel-Word settings are comparable, although GPRank secures higher IoU=0.5 scores. On Charades-CG (presented in the appendix), GPRank markedly outperforms the leading state-of-the-art method, VISA+ASSL (47.44 v.s. 43.89), in the Novel-Composition setting. In addition, in terms of the IoU=0.7 metric across all settings on Charades-CG, GPRank outdoes all considered baselines.

The overall results underscore the robustness of our method in generalizing to novel semantic combinations and new words on both datasets. Intriguingly, both VISA+ASSL and MS-2D-TAN+SSL employ specially designed self-supervised learning modules to better align the semantic space of the two input modalities. The fact that GPRank does not yet incorporate a self-supervised method suggests a potential avenue for further improvement.

5 Analysis

Ablation Study Table 4 presents the results of an ablation study conducted on the Charades-CD dataset. In the study, the base model, EMB (Huang et al., 2022), is progressively augmented with various components: `Roberta`, `ResidualEmb`, `GuidedQuery`, and `CLN`. By comparing each row

	Method	*Test-Trivial*			*Novel-Composition*			*Novel-Word*		
		μ=0.5	μ=0.7	mIoU	μ=0.5	μ=0.7	mIoU	μ=0.5	μ=0.7	mIoU
WeakSup	WSSL (Duan et al., 2018)	11.03	4.14	15.07	2.89	0.76	7.65	3.09	1.13	7.10
RL-based	TSP-PRL (Wu et al., 2020)	34.27	18.80	37.05	14.74	1.43	12.61	18.05	3.15	14.34
Proposal-free	LGI (Mun et al., 2020)	43.56	23.29	41.37	23.21	9.02	27.86	23.10	9.03	26.95
	VLSNet (Zhang et al., 2020a)	39.27	23.12	42.51	20.21	9.18	29.07	21.68	9.94	29.58
	VISA (Li et al., 2022b)	47.13	29.64	44.02	31.51	16.73	35.85	30.14	15.90	35.13
	VISA+ASSL (Li et al., 2023b)	49.37	31.18	46.15	<u>33.22</u>	<u>17.83</u>	<u>37.56</u>	<u>32.04</u>	<u>17.24</u>	**36.87**
Proposal-based	TMN (Liu et al., 2018)	16.82	7.01	17.13	8.74	4.39	10.08	9.93	5.12	11.38
	2D-TAN (Zhang et al., 2020b)	44.50	26.03	42.12	22.80	9.95	28.49	23.86	10.37	28.88
	2D-TAN+SSL (Li et al., 2023a)	46.58	29.65	45.60	27.18	12.60	30.98	26.58	12.55	30.09
	DeCo (Yang et al., 2023)	43.98	24.25	43.47	27.35	11.66	31.27	-	-	-
	LGI+DeCo (Yang et al., 2023)	47.38	28.43	46.03	28.69	12.98	32.67	-	-	-
	MS-2D-TAN (Zhang et al., 2021b)	48.80	31.52	46.58	29.86	14.40	31.80	28.90	13.83	31.01
	MS-2D-TAN+SSL (Li et al., 2023a)	<u>49.63</u>	<u>31.73</u>	<u>47.22</u>	30.80	15.39	33.18	30.15	14.97	32.14
Hybrid	GPRank (ours)	**52.37**	**33.10**	**50.53**	**34.76**	**17.97**	**37.72**	**33.44**	**17.28**	<u>36.64</u>

Table 3: Performances on ActivityNet-CG. **Bold** and <u>underlined</u> denote the best and second best results, respectively.

	test-iid		test-ood	
	0.5	0.7	0.5	0.7
EMB	62.33	43.14	48.68	30.02
+Roberta	59.66	42.41	47.34	28.62
+Roberta+Res	62.45	42.77	50.61	30.64
+Roberta+Res+GuidedQuery	61.24	41.19	51.64	30.99
+Roberta+Res+GuidedQuery+CLN	64.28	40.71	52.47	31.41
EMB w/ global ranking loss	63.00	44.00	51.80	31.00
GPRank w/o global ranking loss	61.48	40.83	52.30	32.53
GPRank	**64.52**	**44.47**	**54.87**	**34.55**

Table 4: Ablation study on Charades-CD.

with its predecessor, we can observe the impact of each component on the system's performance.

However, upon adding `Roberta` to EMB, we notice a **decrease** in performance for both test-iid and test-ood conditions. This implies that the integration of Roberta into this system does not enhance the results. While this finding might seem counterintuitive given Roberta's strong performance in language tasks, it aligns with the results obtained by Shimomoto et al. (2023). One potential explanation could be that Roberta embeddings, unlike the GloVe embeddings used in the EMB model, are more context-specific and dynamic. These properties might make it challenging to establish a robust mapping function necessary for bridging the gap between text and video modalities.

When we add the `ResidualEmb` component to the `EMB+Roberta` model, an improvement is seen in the IoU scores for both test-iid and test-ood conditions at the 0.5 and 0.7 thresholds. This suggests that the `ResidualLLM` contributes positively to the model's performance. The inclusion of `GuidedQuery` in the `EMB+Roberta+Res` model further enhances the IoU scores under the test-ood condition, but slightly reduces the scores under the test-iid condition. This might indicate a trade-off situation. The addition of `CLN` to the `EMB+Roberta+Res+GuidedQuery` model improves the IoU scores under both test-iid and test-ood conditions, signifying that the `CLN` component positively contributes to the model's effectiveness.

Finally, we compare the performance of the GPRank model with and without the global ranking loss. The GPRank model without global ranking loss shows lower IoU scores under both conditions compared to the version with the global ranking loss. This suggests that the global ranking loss is a valuable contribution to the model's performance.

Effect of λ The impact of the λ parameter is investigated by incrementing its value from 0 to 1 in steps of 0.1. A λ value of 1 implies the exclusive use of local boundary-based classifiers, while a value of 0 indicates sole reliance on global span ranking scores. As illustrated in Figure 3, the global span ranking model's pure form yields lower results compared to the pure local boundary classifiers. We posit that this is because the global span ranking model requires effective span representations for successful training. At present, we utilize only boundary-based features, neglecting the internal features of spans. When $\lambda < 0.5$, the performance remains relatively strong, whereas it deteriorates for $\lambda > 0.5$. The model exhibits optimal performance at $\lambda = 0.5$. This suggests that the global ranking scores account for a non-negligible role.

More analysis and discussions are included in the appendix.

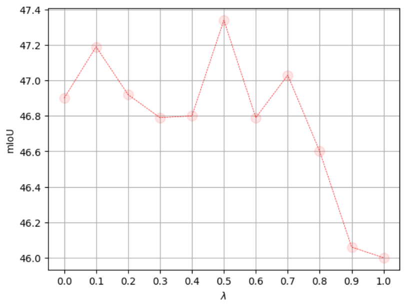

Figure 3: The effect of λ (Eq 4) on ActivityNet.

6 Related Work

Temporal Sentence Video Grounding Models Temporal sentence video grounding (TSVG) was introduced by Gao et al. (2017) and quickly gained research community's attention. Methodologies for this task are typically proposal-free or proposal-based. Proposal-free methods target recognizing start and end boundaries of moments (Chen et al., 2018; Ghosh et al., 2019; Zeng et al., 2020; Zhang et al., 2020a; Li et al., 2021; Zhou et al., 2021; Nan et al., 2021; Xu et al., 2021). They train models using ground-truth endpoints but can be biased due to annotation uncertainties (Otani et al., 2020; Zhou et al., 2021; Huang et al., 2022). Proposal-based methods generate candidate proposals from video segments, aligning them with query sentences (Gao et al., 2017; Anne Hendricks et al., 2017; Ge et al., 2019; Zhang et al., 2019, 2020b, 2021b; Zheng et al., 2022; Li et al., 2023a). The top-ranked proposal is chosen as the prediction. While less boundary-sensitive, their success hinges on proposal quality and ranking efficiency. Hybrid methods blend proposal-free and proposal-based advantages, using both segment and frame-level data for deeper video insight (Wang et al., 2020, 2021a; Xiao et al., 2021; Huang et al., 2022). Notably, Huang et al. (2022) address the uncertain boundary issue by generating a set of elastic boundaries that are dynamically built using proposal-based methods. Despite these advancements, Huang et al. (2022) generate proposals using sampled moments, whereas our model considers all possible moments.

TSVG with Pretrained Language Models The use of TSVG with GloVE embeddings (Pennington et al., 2014) still remarkably dominates the field. The exploration of TSVG with pretrained language models such as BERT (Devlin et al., 2019) and Roberta (Liu et al., 2019) is less prevalent. Although Yang et al. (2022) used Roberta for spatio-temporal video grounding, it is a different task. Further, Wang et al. (2021b) and Zheng et al. (2023) utilize DistillBERT (Sanh et al., 2020) as the text encoder, a distilled version of BERT that may not fully leverage BERT's capabilities. Recent work by Shimomoto et al. (2023) successfully employ efficient adapter-based pretrained language models (PLMs) for TSVG. Despite their efforts, fine-tuning the pretrained encoder on Charades-STA (Gao et al., 2017) yields limited improvements or occasionally reduces performance across different backbone models, indicating the challenge of integrating PLMs for TSVG. We diverge from these methods by designing a backbone-specific integration that enables better interaction between the pretrained text features and video features.

Global Proposal Ranking Liu et al. (2021) and Zhang et al. (2021b) proposed methods for ranking candidate proposals using cross-entropy loss. Liu et al. (2021) introduced a contextual biaffine scoring network, while Zhang et al. (2021b) employs multi-scale 2D temporal feature maps. However, both methods use cross-entropy as their training objective and do not explicitly consider the ranking of all candidate moments from a global perspective. Our method adopts a global ranking loss function, originally designed for multi-label classification (Su et al., 2022). The ranking score of a moment proposal is directly calculated based on its overlap with the ground truth, thus enabling a global ranking of the candidate moments.

7 Conclusion

In this paper, we presented an exploration of integrating the pre-trained language model Roberta for temporal video grounding models. Our focus was not only to enhance the model's performance but also to ensure robustness in varying conditions. Contrary to expectations, the direct incorporation of Roberta resulted in a slight performance decrease in a dataset, emphasizing the importance of thoughtfully integrating these models. To address this, we proposed architecture modifications, which positively impacted the IoU scores in both in-distribution and out-of-distribution, and compositional generalization testing scenarios. We also leveraged a global proposal ranking loss, which further augmented our model's performance, indicating its effectiveness in enhancing the model's robustness. The approach and findings from this study offer valuable guidance for future research in effectively combining large-scale language models and video grounding models.

Limitation

One key limitation is that we only consider one specific temporal video grounding model, EMB, in our work. While EMB is an effective baseline model in this domain, there is a range of other models available in the temporal video grounding literature, each with its unique strengths and features. These models include VSLNet and MS-2D-TAN, among others, which offer different mechanisms for understanding and grounding temporal video content.

Another limitation is our model-specific architecture and global ranking loss, designed to work optimally with the EMB model and Roberta embeddings, might not be directly compatible with other temporal video grounding models or other pre-trained language models and large language models such as LLaMA (Touvron et al., 2023). Therefore, our proposed architecture may require significant adaptations or the development of new components to be compatible with other models.

References

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *ICCV*, pages 5803–5812.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Da Cao, Yawen Zeng, Meng Liu, Xiangnan He, Meng Wang, and Zheng Qin. 2020a. Strong: Spatio-temporal reinforcement learning for cross-modal video moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 4162–4170, New York, NY, USA. Association for Computing Machinery.

Da Cao, Yawen Zeng, Xiaochi Wei, Liqiang Nie, Richang Hong, and Zheng Qin. 2020b. Adversarial video moment retrieval by jointly modeling ranking and localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 898–906, New York, NY, USA. Association for Computing Machinery.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308.

Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. 2018. Temporally grounding natural sentence in video. In *EMNLP*.

Mingjian Chen, Xu Tan, Bohan Li, Yanqing Liu, Tao Qin, Sheng Zhao, and Tie-Yan Liu. 2021. Adaspeech: Adaptive text to speech for custom voice.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. 2018. Weakly supervised dense event captioning in videos. In *NeurIPS*, volume 31.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *ICCV*, pages 5267–5275.

Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. 2019. Mac: Mining activity concepts for language-based temporal localization. In *WACV*.

Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann. 2019. ExCL: Extractive Clip Localization Using Natural Language Descriptions. In *NAACL*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970.

Jiabo Huang, Hailin Jin, Shaogang Gong, and Yang Liu. 2022. Video activity localisation with uncertainties in temporal boundary. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *ICCV*.

Chuanhao Li, Zhen Li, Chenchen Jing, Yunde Jia, and Yuwei Wu. 2023a. Exploring the effect of primitives for compositional generalization in vision-and-language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19092–19101.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022a. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10965–10973.

Juncheng Li, Siliang Tang, Linchao Zhu, Wenqiao Zhang, Yi Yang, Tat-Seng Chua, Fei Wu, and Yueting Zhuang. 2023b. Variational cross-graph reasoning and adaptive structured semantics learning for compositional temporal grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–16.

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

Kun Li, Dan Guo, and Meng Wang. 2021. Proposal-free video grounding with contextual pyramid network. In *AAAI*, volume 35, pages 1902–1910.

Bingbin Liu, Serena Yeung, Edward Chou, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. 2018. Temporal modular networks for retrieving complex compositional activities in videos. In *ECCV*.

Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. 2021. Context-aware biaffine localizing network for temporal sentence grounding. In *CVPR*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Jonghwan Mun, Minsu Cho, and Bohyung Han. 2020. Local-global video-text interactions for temporal grounding. In *CVPR*, pages 10810–10819.

Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. 2021. Interventional video grounding with dual contrastive learning. In *CVPR*, pages 2765–2775.

Mayu Otani, Yuta Nakahima, Esa Rahtu, and Janne Heikkilä. 2020. Uncovering hidden challenges in query-based video moment retrieval. In *BMVC*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP*.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36.

Cristian Rodriguez, Edison Marrese-Taylor, Fatemeh Sadat Saleh, HONGDONG LI, and Stephen Gould. 2020. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *WACV*.

Cristian Rodriguez-Opazo, Edison Marrese-Taylor, Basura Fernando, Hongdong Li, and Stephen Gould. 2021. Dori: Discovering object relationships for moment localization of a natural language query in a video. In *WACV*.

Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. 2012. Script data for attribute-based recognition of composite activities. In *ECCV*, pages 144–157. Springer.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Erica K. Shimomoto, Edison Marrese-Taylor, Hiroya Takamura, Ichiro Kobayashi, Hideki Nakayama, and Yusuke Miyao. 2023. Towards parameter-efficient integration of pre-trained language models in temporal video grounding.

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, pages 510–526. Springer.

Jianlin Su, Mingren Zhu, Ahmed Murtadha, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2022. Zlpr: A novel loss for multi-label classification.

Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. 2020. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6398–6407.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Hao Wang, Zheng-Jun Zha, Xuejin Chen, Zhiwei Xiong, and Jiebo Luo. 2020. Dual path interaction network for video moment localization. In *ACM MM*, pages 4116–4124.

Hao Wang, Zheng-Jun Zha, Liang Li, Dong Liu, and Jiebo Luo. 2021a. Structured multi-level interaction network for video moment localization via language query. In *CVPR*, pages 7026–7035.

Jing Wang, Aixin Sun, Hao Zhang, and Xiaoli Li. 2023. Ms-detr: Natural language video localization with sampling moment-moment interaction. *arXiv preprint arXiv:2305.18969*.

Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. 2021b. Negative sample matters: A renaissance of metric learning for temporal grounding. *ArXiv*, abs/2109.04872.

Jie Wu, Guanbin Li, Si Liu, and Liang Lin. 2020. Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In *AAAI*, volume 34.

Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. 2021. Boundary proposal network for two-stage natural language video localization. In *AAAI*, volume 35, pages 2986–2994.

Caiming Xiong, Victor Zhong, and Richard Socher. 2016. Dynamic coattention networks for question answering. *arXiv preprint arXiv:1611.01604*.

Mengmeng Xu, Juan-Manuel Pérez-Rúa, Victor Escorcia, Brais Martínez, Xiatian Zhu, Li Zhang, Bernard Ghanem, and Tao Xiang. 2021. Boundary-sensitive pre-training for temporal localization in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7220–7230.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2022. Tubedetr: Spatio-temporal video grounding with transformers. In *CVPR*.

Lijin Yang, Quan Kong, Hsuan-Kung Yang, Wadim Kehl, Yoichi Sato, and Norimasa Kobori. 2023. Deco: Decomposition and reconstruction for compositional temporal grounding via coarse-to-fine contrastive ranking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23130–23140.

Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. 2018. Fast and accurate reading comprehension by combining self-attention and convolution. In *ICLR*, volume 2.

Yitian Yuan, Xiaohan Lan, Xin Wang, Long Chen, Zhi Wang, and Wenwu Zhu. 2021. A closer look at temporal sentence grounding in videos. In *Proceedings of the 2nd International Workshop on Human-centric Multimedia Analysis*. ACM.

Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. 2020. Dense regression network for video grounding. In *CVPR*, pages 10287–10296.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2020a. Span-based localizing network for natural language video localization. In *ACL*.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2021a. Towards debiasing temporal sentence grounding in video. *ArXiv preprint arXiv:2111.04321*, abs/2111.04321.

Songyang Zhang, Houwen Peng, Jianlong Fu, Yijuan Lu, and Jiebo Luo. 2021b. Multi-scale 2d temporal adjacency networks for moment localization with natural language. *IEEE TPAMI*.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020b. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, volume 34, pages 12870–12877.

Songyang Zhang, Jinsong Su, and Jiebo Luo. 2019. Exploiting temporal relationships in video moment localization with natural language. In *ACM MM*.

Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. 2017. Temporal action detection with structured segment networks. In *ICCV*.

Minghang Zheng, Yanjie Huang, Qingchao Chen, Yuxin Peng, and Yang Liu. 2022. Weakly supervised temporal sentence grounding with gaussian-based contrastive proposal learning. In *CVPR*.

Minghang Zheng, Sizhe Li, Qingchao Chen, Yuxin Peng, and Yang Liu. 2023. Phrase-level temporal relationship mining for temporal sentence localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Hao Zhou, Chongyang Zhang, Yan Luo, Yanjun Chen, and Chuanping Hu. 2021. Embracing uncertainty: Decoupling and de-bias for robust temporal grounding. In *CVPR*.

A Appendix

A.1 The relationship between μ and g

To minimize the loss function above, the model needs to increase the value of $g_{i,j}$ when $\mu_{i,j}$ is large and decrease the value of $g_{i,j}$ when $\mu_{i,j}$ is small. When $\mu_{i,j}$ becomes a binary variable, the loss function is identical to circle loss (Sun et al., 2020). The direct relation between $\mu_{i,j}$ and $g_{i,j}$ during inference can be derived by considering the partial derivative of $\mathcal{L}_{span}$ with respect to each $g_{i,j}$:

$$\frac{\partial \mathcal{L}_{span}}{\partial g_{i,j}} = \frac{-\mu_{i,j} e^{-g_{i,j}}}{1 + \sum\limits_{i \le j} \mu_{i,j} e^{-g_{i,j}}} + \frac{(1-\mu_{i,j}) e^{g_{i,j}}}{1 + \sum\limits_{i \le j} (1-\mu_{i,j}) e^{g_{i,j}}}, \tag{21}$$

By setting $\mu_{i,j} e^{-g_{i,j}} = (1-\mu_{i,j}) e^{g_{i,j}}$, the partial derivative $\frac{\partial \mathcal{L}_{span}}{\partial g_{i,j}}$ equals zero, indicating local minimums of the loss function. Solving this equation, we get: $\hat{\mu}_{i,j} = \sigma(2g_{i,j})$, where σ denotes the sigmoid function. This indicates that for prediction, we can approximate the probability of the span $[i, j]$ being the target span using $\sigma(2g_{i,j})$.

A.2 Guided Attention

Content Guided Attention. The content-guided attention module incorporates the preceding and subsequent content information of each frame into its representation. This approach emphasizes the importance of discerning changes or differences between sequential frames:

$$\begin{aligned} \boldsymbol{V}_{\text{pre}} &= \{\text{MaxPool}(\{\boldsymbol{f}_i\}_{i=1}^{t})\}_{t=1}^{K} \in \mathbb{R}^{D\times K}, \\ \boldsymbol{V}_{\text{sub}} &= \{\text{MaxPool}(\{\boldsymbol{f}_i\}_{i=t}^{K})\}_{t=1}^{K} \in \mathbb{R}^{D\times K}, \\ \tilde{\boldsymbol{V}} &= \text{Conv2d}(\{\boldsymbol{V}, \boldsymbol{V}_{\text{pre}}, \boldsymbol{V}_{\text{sub}}\}) \in \mathbb{R}^{D\times K}. \end{aligned} \tag{22}$$

Afterwards, the content-guided representations of video frames ($\tilde{\boldsymbol{V}}$) are used for attentive encoding (Eq.11), both within the same modality ($\boldsymbol{V} \leftarrow g(\tilde{\boldsymbol{V}}, \tilde{\boldsymbol{V}})$ and across modalities ($\boldsymbol{V} \leftarrow g(\tilde{\boldsymbol{V}}, \boldsymbol{Q})$). This way, the model pays attention not just to the content of individual frames, but also to how they change over time, aiding in the identification of key moments within the video.

Boundary Guided Attention. Similar to the content-guided attention approach for video frames, we explicitly incorporate the frame-wise boundary features with the content representations of video segments to promote boundary-sensitive content alignment:

$$\begin{gathered} \boldsymbol{P}_{\text{sta}} = \{\boldsymbol{f}_{t_k^s}\}_{k=1}^{N} \in \mathbb{R}^{D\times N},\ \boldsymbol{P}_{\text{end}} = \{\boldsymbol{f}_{t_k^e}\}_{k=1}^{N} \in \mathbb{R}^{D\times N}, \\ \tilde{\boldsymbol{P}} = \text{Conv2d}(\{\boldsymbol{P}, \boldsymbol{P}_{\text{sta}}, \boldsymbol{V}_{\text{end}}\}) \in \mathbb{R}^{D\times N}. \end{gathered} \tag{23}$$

In Eq. (23), the features $\boldsymbol{P}_{\text{sta}}$ and $\boldsymbol{P}_{\text{end}}$ represent the start and end frames of each of the K proposals. These boundary features are stacked and combined with the segment-wise content features $\boldsymbol{P}$ through a 2D convolution layer to generate the boundary-guided segment representations $\tilde{\boldsymbol{P}}$. This boundary-guided attention approach shares the same philosophy as temporal pyramid pooling (Zhao et al., 2017), in that it explicitly encodes the temporal structure into the segment's representation to make it sensitive to the segment's boundaries. The boundary-guided representations $\tilde{\boldsymbol{P}}$ are then used for attentive encoding within the same modality $\boldsymbol{P} \leftarrow g(\tilde{\boldsymbol{P}}, \tilde{\boldsymbol{P}})$ and across different modalities $\boldsymbol{P} \leftarrow g(\tilde{\boldsymbol{P}}, \boldsymbol{Q})$) as defined by Eq.11.

A.3 Data Statistics

Table 5 shows the data statistics on the distribution-consistent settings.

A.4 Implementation Details

We generally follow the settings of Huang et al. (2022). We employed the provided video features

Metric	ActivityNet	Charades	TACoS
#Train	37,421	12,408	10,146
#Val	17,031	-	4,589
#Test	17,505	3,720	4,083
Avg Len of V	117.61s	30.59s	287.14s
Avg Len of M	36.18s	8.22s	5.45s
Avg Words of Q	14.8	7.2	10.1

Table 5: Data Statistics. V: video, M: ground-truth moment, Q: language query.

of Zhang et al. (2020a) to encode video inputs. For text inputs, we use the 300D GloVe (Pennington et al., 2014) embeddings and Roberta-base (Liu et al., 2019) as the pretrained langauge model. We tune our GPRank model for 20 epochs using a batch size of 16. The backbone parameters of EMB and the parameters of Roberta are tuned using separate Adam optimizers. For the backbone parameters, we use a learning rate of 5e-4. For the Roberta parameters, we use 5e-6 for Charades datasets (including Charades-STA, Charades-CD, Charades-CG) and 1e-5 for TACoS and ActivityNet datasets (ActivityNet-Captions, ActivityNet-CD, ActivityNet-CG). To represent the input language query, we use the last output layer of Roberta for Charades-related datasets and sum the last four output layers of Robert for TACoS and ActivityNet-related datasets. For the loss weights, $\lambda_1 = 1.0$, $\lambda_2 = 1.0$, $\lambda_3 = 5.0$, and $\lambda_4 = 1.0$ give the optimal performance.

A.5 Effect of Pretrained Language Models

In this study, we compared our models with different pretrained language models on the Charades-STA test set. Our results were compared with those from models such as TMLGA, DoRi (Shimomoto et al., 2023), MMN (Wang et al., 2021b), and TRM (Zheng et al., 2023). Notably, TMLGA is a less robust backbone model compared to DoRi. Both MMN and TRM, which utilised DistillBERT (Sanh et al., 2020) as their encoder, are based on VGG video features, making a direct comparison less feasible. However, they have been included for reference.

TMLGA (Rodriguez et al., 2020) exhibited similar results across all three pretrained encoders. DoRi (Rodriguez-Opazo et al., 2021) also achieved comparable performance using both BERT (Devlin et al., 2019) and DeBERTa (He et al., 2021), outperforming TMLGA by a significant margin. These results suggest that the choice between BERT,

Method	Encoder	IoU=0.3	IoU=0.5	IoU=0.7	mIoU
MNL	DistillBERT	60.48	47.45	27.15	—
TRM	DistillBERT	60.67	47.77	28.01	42.77
TMLGA	BERT	71.02	52.53	33.52	49.80
TMLGA	Roberta	-	53.84	34.78	49.91
TMLGA	DeBERTa	-	53.49	34.65	49.78
DoRi	BERT	72.50	58.63	40.97	53.29
DoRi	DeBERTa	-	58.39	**41.61**	53.34
GPRank	Roberta	**74.76**	**60.78**	<u>40.86</u>	**54.39**

Table 6: Comparisons with different pretrained language model encoders on Charades-STA test set.

Dataset	Config	mIoU	$\mu = 0.3$	$\mu = 0.5$	$\mu = 0.7$
TACoS	CE	36.01	51.48	37.43	22.41
	CE + combined	36.49	52.00	37.77	22.60
	GPRank	37.93	54.14	38.42	24.12
Charades-STA	CE	53.60	73.57	59.64	39.92
	CE + combined	53.93	74.45	60.18	39.83
	GPRank	54.39	74.76	60.78	40.86

Table 7: Effect of cross-entropy loss

Roberta, and DeBERTa might yield similar performance levels when the same backbone models are used, implying that the backbone models might have a more significant impact.

Drawing from these experiences, we chose to utilise only the Roberta encoder, which provided the best performance for TMLGA among the three encoders. Our method GPRank demonstrated the highest scores across all IoU thresholds, outperforming all other methods and encoders. These results underscore the effectiveness of our unique design approach, which involves a deep integration of pretrained language representations with the EMB backbone, over architecture-agnostic integration.

A.6 Effect of cross-entropy loss

We also investigate performances of cross-entropy (CE) loss with pre-trained language models, combining probabilities predicted by CE (CE+combined) and local boundary classifier probabilities. Table 7 shows the results on TACoS and Charades-STA. Comparable outcomes are observed with ActivityNet as well. Using CE loss is also beneficial with our encoder. A configuration with our prediction fusion further enhances performance. However, both approaches fall short when compared to our proposed method.

A.7 Composition Generalization Results on Charades-CG

Table 8 shows the composition generalization results on Charades-CG. Table 8 reveals that MS-2D-TAN+SSL emerges as the top-performing baseline model, achieving the highest IOU=0.7 score in the Novel-Word setting. VISA+ASSL stands out with its superior IOU=0.5 performance in the Novel-Composition setting. Our GPRank method registers seven top records and two second-place records across the nine metrics. Particularly, GPRank excels in the Test-Trivial setting, surpassing all baseline methods. In the Novel-Composition setting, GPRank markedly outperforms the leading state-of-the-art method, VISA+ASSL (47.44 v.s. 43.89). In terms of the IoU=0.7 metric across all settings, GPRank outdoes all considered baselines.

	Method	*Test-Trivial*			*Novel-Composition*			*Novel-Word*		
		μ=0.5	μ=0.7	mIoU	μ=0.5	μ=0.7	mIoU	μ=0.5	μ=0.7	mIoU
WeakSup	WSSL	15.33	5.46	18.31	3.61	1.21	8.26	2.79	0.73	7.92
RL-based	TSP-PRL	39.86	21.07	38.41	16.30	2.04	13.52	14.83	2.61	14.03
Proposal-free	LGI	49.45	23.80	45.01	29.42	12.73	30.09	26.48	12.47	27.62
	VLSNet	45.91	19.80	41.63	24.25	11.54	31.43	25.60	10.07	30.21
	VISA	53.20	26.52	47.11	45.41	22.71	42.03	42.35	20.88	40.18
	VISA+ASSL	56.14	28.27	48.92	**47.76**	24.85	<u>43.89</u>	44.75	22.31	42.38
Proposal-based	TMN	18.75	8.16	19.82	8.68	4.07	10.14	9.43	4.96	11.23
	2D-TAN	48.58	26.49	44.27	30.91	12.23	29.75	29.36	13.21	28.47
	2D-TAN+SSL	53.91	31.82	46.84	35.42	17.95	33.07	43.60	25.32	39.32
	DeCo	<u>58.75</u>	28.71	49.06	47.39	21.06	40.70	-	-	-
	MS-2D-TAN	57.85	37.63	50.51	43.17	23.27	38.06	45.76	27.19	40.80
	MS-2D-TAN+SSL	58.14	<u>37.98</u>	<u>50.58</u>	46.54	<u>25.10</u>	40.00	<u>50.36</u>	<u>28.78</u>	**43.15**
Hybrid	GPRank (ours)	**59.85**	**40.89**	**53.72**	<u>47.04</u>	**29.46**	**47.44**	**51.80**	**34.53**	<u>43.01</u>

Table 8: Performances on Charades-CG. **Bold** and <u>underlined</u> denote the best and second best results, respectively.