Diversifying Query: Region-Guided Transformer for Temporal Sentence Grounding

Xiaolong Sun[1*], Liushuai Shi[1*], Le Wang[1†]
Sanping Zhou[1], Kun Xia[1], Yabing Wang[1], Gang Hua[2]
[1]National Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
National Engineering Research Center for Visual Information and Applications,
and Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
[2]Multimodal Experiences Research Lab, Dolby Laboratories
{sunxiaolong,shiliushuai,xiakun}@stu.xjtu.edu.cn, {lewang,spzhou}@xjtu.edu.cn, {wyb7wyb7,ganghua}@gmail.com

Abstract

Temporal sentence grounding is a challenging task that aims to localize the moment spans relevant to a language description. Although recent DETR-based models have achieved notable progress by leveraging multiple learnable moment queries, they suffer from overlapped and redundant proposals, leading to inaccurate predictions. We attribute this limitation to the lack of task-related guidance for the learnable queries to serve a specific mode. Furthermore, the complex solution space generated by variable and open-vocabulary language descriptions complicates optimization, making it harder for learnable queries to adaptively distinguish each other, leading to more severe overlapped proposals. To address this limitation, we present the Region-Guided TRansformer (RGTR) for temporal sentence grounding, which introduces regional guidance to increase query diversity and eliminate overlapped proposals. Instead of using learnable queries, RGTR adopts a set of anchor pairs as moment queries to introduce explicit regional guidance. Each moment query takes charge of moment prediction for a specific temporal region, which reduces the optimization difficulty and ensures the diversity of the proposals. In addition, we design an IoU-aware scoring head to improve proposal quality. Extensive experiments demonstrate the effectiveness of RGTR, outperforming state-of-the-art methods on three public benchmarks and exhibiting good generalization and robustness on out-of-distribution splits.

Code — https://github.com/TensorsSun/RGTR

1 Introduction

Temporal sentence grounding (TSG) aims at localizing the moment spans semantically aligned with the given language description in an untrimmed video. Early methods address the TSG task by designing predefined dense proposals (Gao et al. 2017; Wang et al. 2022b) or directly learning sentence-frame interactions (Liu et al. 2022a; Yang and Wu 2022). The recent success of detection transformer (DETR) has inspired the integration of transformers into the TSG framework (Moon et al. 2023b; Xiao et al. 2024). By decoding moment spans from a set of learnable queries, they streamline the complicated grounding pipeline.

Although DETR-based approaches have achieved notable performance in TSG task, we still observe some unique limitations of the DETR structure compared to other fields (*e.g.*, object detection). Specifically, they suffer from limited query distribution and overlapped proposals, leading to inaccurate predictions. As shown in Fig. 1, we present the center-length distribution of moment queries in three DETR-based methods, each query learns to predict different temporal regions (*e.g.*, the lower left area represents a short moment near the video's start and the higher middle area represents a long moment). In previous methods, each query includes numerous overlapped and redundant proposals for the same region (*e.g.*, the short moments in the lower part), resulting in ineffective predictions. We attribute this limitation to the lack of task-related guidance (*e.g.*, category constraints, spatial distribution prior, etc.) for the learnable queries to serve a specific mode. Although task-related guidance is crucial to reducing the overlapped proposals, it has been scarcely explored in TSG task. Furthermore, the complex solution space generated by variable and open-vocabulary language descriptions exacerbates the optimization difficulty, making it harder for learnable queries to distinguish each other adaptively and resulting in more severe overlapped proposals. Another limitation is that the proposal scoring in previous methods is purely based on the classification confidence, ignoring the quality of the predicted boundary. Instead, we argue that correctly classified proposals that better overlap with the ground-truth should be assigned higher scores. The above limitations significantly restrict the accurate localization of the DETR structure in TSG task.

In this paper, we introduce an effective Region-Guided TRansformer (RGTR) framework to cope with the aforementioned limitations in TSG task. To address the issue of overlapped proposals, we introduce regional priors based on the distribution of ground-truth moment spans as task-related guidance. This regional guidance can eliminate overlapped proposals by increasing query diversity. Specifically, we design a region-guided decoder with a new concept of anchor pairs as moment queries to provide regional guidance. Each moment query consists of a static anchor and a

*These authors contributed equally.
†Corresponding author.

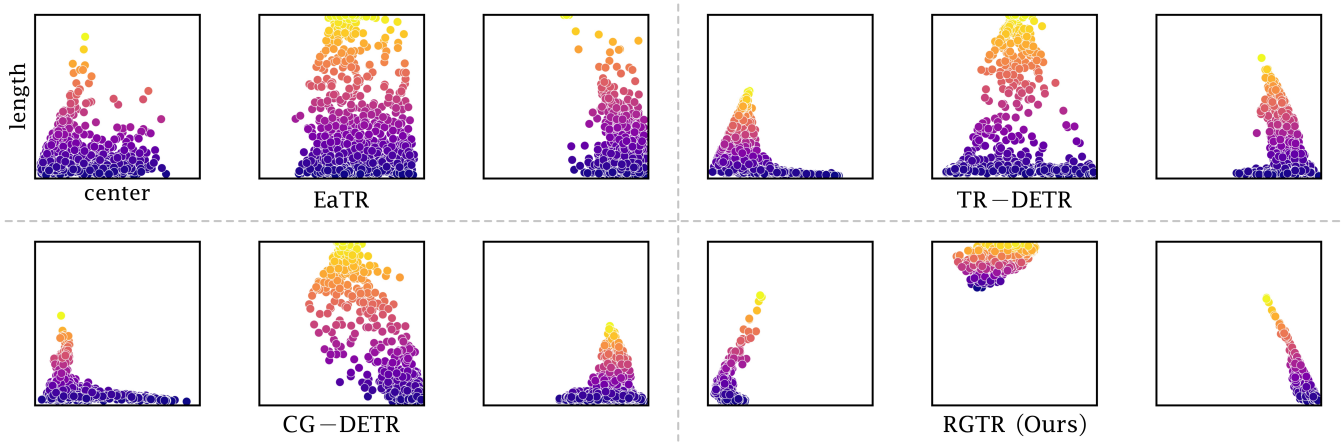

Figure 1: Visualization comparison of all moment predictions on QVHighlights *val* split, for the 3 representative moment queries in EaTR (Jang et al. 2023), TR-DETR (Sun et al. 2024), CG-DETR (Moon et al. 2023a) and RGTR (Ours). x-axis denotes the normalized moment span center coordinate, y-axis denotes the normalized moment span length. All queries in previous methods generate numerous overlapped proposals. For example, the second query tends to predict long moments near the middle of the videos (higher middle area), but the proposals of short moments (lower area) conflict with this purpose, leading to ineffective predictions. In contrast, the predicted region of each query in our RGTR is distinct and more concentrated.

dynamic anchor, both of which are initialized by different clustering centers on the ground-truth moment spans. Such explicit initialization imposes regional priors as guidance on each moment query, enhancing the diversity of query distribution. The two types of anchors serve different roles in the decoder. The static anchor is designed to maintain the regional guidance, so it is not updated during decoding. With the help of the fixed static anchor, the dynamic anchor continuously updates to make diverse predictions for various temporal regions. They collaboratively guide localization with explicit regional guidance and eliminate overlapped proposals. In addition, to improve the scoring of high-quality proposals, we propose an IoU-aware scoring head. By supervising the IoU score with L2 loss, the prediction head considers both classification confidence and localization quality.

Extensive experiments on three TSG benchmarks demonstrate the effectiveness of RGTR framework. As shown in Fig. 1, RGTR eliminates redundant proposals and exhibits diverse query distributions compared to previous methods. Our main contributions are summarized as follows: (1) We design a novel region-guided decoder, which adopts a set of explicitly initialized anchor pairs as moment queries to introduce regional priors as task-related guidance. (2) We propose an IoU-aware scoring head that incorporates localization quality to enhance classification confidence estimation and distinguish high-quality proposals. (3) By employing these techniques, we introduce a Region-Guided TRansformer that eliminates overlapped proposals and improves localization quality. RGTR achieves state-of-the-art performance on three challenging benchmarks and exhibits good generalization and robustness on out-of-distribution splits.

2 Related Work

Temporal Sentence Grounding. Temporal sentence grounding aims at predicting the moment spans of the described activity given an untrimmed video and a language description, which is first proposed in (Gao et al. 2017). Early methods fall into proposal-based methods and proposal-free methods. Proposal-based methods (Liu et al. 2018; Xia et al. 2022; Wang et al. 2022b) initially generate multiple candidate proposals and rank them based on their similarity with the description. Proposal-free methods (Lu et al. 2019; Chen et al. 2020; Yang and Wu 2022) are proposed to avoid the need for predefined candidate moments. Instead of relying on segment candidates, they directly predict the start and end boundaries of the target moments. The recent success of detection transformer (DETR) (Carion et al. 2020) has inspired the integration of transformers into the temporal sentence grounding framework (Lei, Berg, and Bansal 2021; Liu et al. 2022c; Lee and Byun 2023). DETR-based methods simplify the whole process into an end-to-end manner by removing handcrafted techniques. However, due to the lack of task-related guidance for the learnable queries to serve a specific mode, almost all previous methods generate numerous overlapped and redundant proposals. In contrast, our method eliminates overlapped proposals by introducing regional guidance.

Detection Transformers. Recently, the adoption of transformers to object detection (DETR) (Carion et al. 2020) builds a fully end-to-end object detection system based on transformers. The formulation of decoder queries has also been widely studied in previous work (Zhu et al. 2020; Shi et al. 2022, 2023). Anchor DETR (Wang et al. 2022a) initializes queries based on anchor points for specific detection modes. DAB-DETR (Liu et al. 2022b) formulates decoder queries with content and action embeddings. DINO (Zhang et al. 2022) adds position priors for the positional query and randomly initializes the content query. Motivated by their great success, we introduce a set of anchor pairs to introduce explicit regional guidance for accurate prediction.

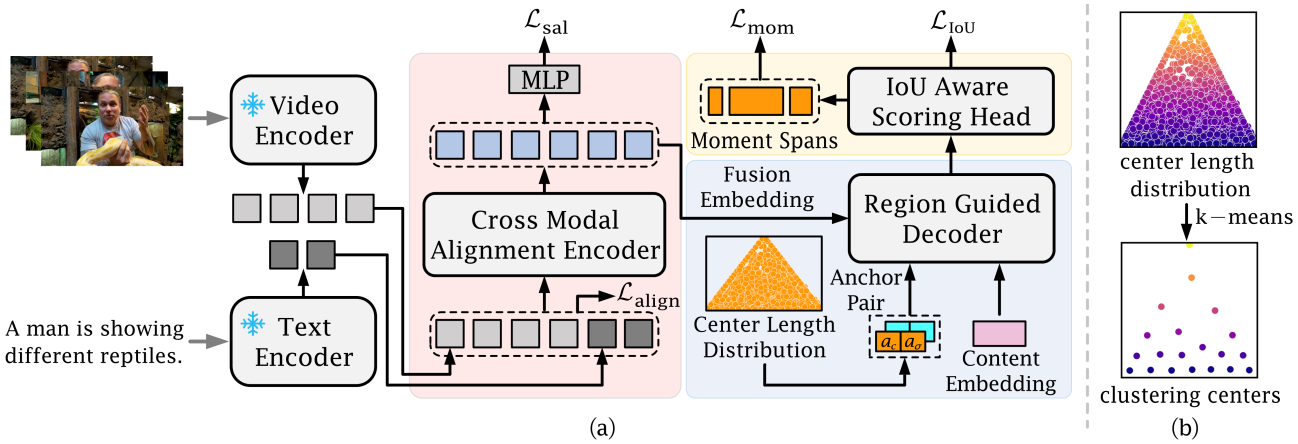

Figure 2: (a) Overview of the proposed RGTR architecture. Given a video and a text description, we first employ two frozen pre-trained models to extract visual and textual features. Subsequently, the cross-modal alignment encoder is constructed to align and fuse the visual and textual features effectively. Then, we design a region-guided decoder to introduce the regional guidance for decoding process through a set of explicitly initialized anchor pairs. Finally, the IoU-aware scoring head generates high-quality proposals by incorporating localization quality to enhance the classification confidence estimation. (b) The clustering centers with regional priors are obtained by adopting k-means algorithm on the distribution of all ground-truth moment spans.

3 Method

3.1 Overview

Given an untrimmed video $\mathcal{V} = \{v_t\}_{t=1}^{L}$ with L frames and an associated natural language description $\mathcal{T} = \{t_n\}_{n=1}^{N}$ with N words, TSG aims to accurately predict the moment span $m = (m_c, m_\sigma)$ that is most relevant to the given description, where m_c and m_σ represent the center time and duration length of the moment span.

Recent DETR-based methods replace hand-crafted components with learnable positional queries to predict target moments. These positional queries, representing a set of learnable referential search areas, are initialized as random learnable embeddings in the previous methods (Moon et al. 2023b; Yang et al. 2024; Xiao et al. 2024). However, due to the lack of task-related guidance (*e.g.*, categories constraints, spatial distribution prior, etc.) and the extensive variability of language descriptions, the random initialization of positional queries greatly exacerbates the optimization difficulty and produces numerous overlapped proposals.

To address this problem, we propose the Region-Guided TRansformer (RGTR), which adopts a set of explicitly initialized anchor pairs as moment queries to replace randomly initialized learnable queries without guidance. In our framework, we construct a region-guided decoder through anchor pairs to provide directive and diverse reference search areas for decoding process. In addition, we introduce an IoU-aware scoring head to distinguish high-quality proposals. The overall architecture is shown in Fig. 2a.

3.2 Cross-Modal Alignment Encoder

Following previous methods (Moon et al. 2023b; Li et al. 2024), we use the pre-trained CLIP (Radford et al. 2021) and Slowfast model (Feichtenhofer et al. 2019) to extract clip-level visual features $F_v \in \mathbb{R}^{L \times d_v}$, where L represents the number of clips and d_v is the dimension of visual features. Furthermore, we utilize the CLIP model to extract word-level textual features $F_t \in \mathbb{R}^{N \times d_t}$, where N is the number of words and d_t is the dimension of textual features.

Given the clip-level visual features F_v and the word-level textual features F_t, they are first projected into the common multimodal space using multi-layer perceptrons (MLPs) to produce the corresponding features $\bar{F}_v \in \mathbb{R}^{L \times D}$ and $\bar{F}_t \in \mathbb{R}^{N \times D}$, where D is the embedding dimension. As highlighted in previous work (Li et al. 2021; Sun et al. 2024), aligning modalities before interaction could reduce the modal gap and obtain better modal representations. Therefore, we employ an alignment loss $\mathcal{L}_{\text{align}}$ to facilitate the alignment between videos and sentences.

$$\mathcal{L}_{\text{align}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp((G_v^i)(G_t^i)^\top)}{\sum_{i=1}^{B}\sum_{j=1}^{B}\exp((G_v^i)(G_t^j)^\top)}, \quad (1)$$

where B represents the batch size, $G_v^i \in \mathbb{R}^D$ and $G_t^i \in \mathbb{R}^D$ denote the global feature of the i-th video and the i-th sentence in a training batch, respectively.

After alignment, we adopt a text-to-video encoder to obtain text-aware video representations. Specifically, three cross-attention layers are utilized to integrate textual features into the visual features:

$$\hat{F}_v = \text{Attention}(Q_v, K_t, V_t) = \text{Softmax}(\frac{Q_v K_t^\top}{\sqrt{D}})V_t. \quad (2)$$

where $Q_v = \text{Linear}_q(\bar{F}_v)$, $K_t = \text{Linear}_k(\bar{F}_t)$ and $V_t = \text{Linear}_v(\bar{F}_t)$. Subsequently, three self-attention layers are leveraged to enhance the representations to help the model better understand the video sequence relations. Here, we project $\hat{F}_v$ to $Q_{\hat{v}}$, $K_{\hat{v}}$ and $V_{\hat{v}}$ and use them to obtain the final cross-modal fusion embedding F, which is imposed by saliency score constraints $\mathcal{L}_{\text{sal}}$ (Moon et al. 2023b).

3.3 Region-Guided Decoder

Given the fusion embedding F, we aim to localize moment spans semantically aligned with the description in the decoder. As discussed in Sec. 3.1, previous methods employ randomly initialized learnable queries without task-related guidance, leading to increasing optimization difficulty and numerous overlapped proposals. In contrast, we design a region-guided decoder, which adopts a set of explicitly initialized anchor pairs as moment queries to provide directive and diverse regional guidance. Each anchor pair consists of a static anchor and a dynamic anchor, both of which are initialized by clustering centers on the ground-truth moment spans. The two types of anchors serve different roles in the decoder, where static anchors maintain regional guidance without updating and dynamic anchors make diverse predictions. They collaboratively guide localization with explicit regional guidance. The structure of the region-guided decoder is described in Fig. 3. We elaborate on the detailed process in the following.

Anchor Explicit Initialization. Due to the specificity of the TSG task, we lack the task-related guidance (*e.g.*, category constraints) present in other detection tasks. Nonetheless, we can still provide regional guidance for the decoding process by considering the distribution of ground-truth moment spans. Specifically, the forms of static anchors and dynamic anchors are first defined as $a = (a_c, a_\sigma)$, where a_c is the center coordinate and the a_σ is the duration of the moment. Then, as shown in Fig. 2b, we generate $\mathcal{K}$ clustering centers $A \in \mathbb{R}^{\mathcal{K}\times 2}$ by adopting k-means clustering algorithm on the distribution of all ground-truth moment spans. These clustering centers represent explicit temporal regions with diverse center coordinates and durations. Since events described in the text can occur anywhere in videos, generating diverse temporal regions as guidance is crucial. Therefore, the static and dynamic anchors are initialized by $\mathcal{K}$ clustering centers: $A_s^0 = A_d^0 = A \in \mathbb{R}^{\mathcal{K}\times 2}$, and the positional embeddings of anchor pairs are generated by:

$$P_s^0 = P_d^0 = \text{MLP}(\text{PE}(A)), \tag{3}$$

where $\text{PE}(\cdot)$ means positional encoding to generate sinusoidal embeddings. For clarity, we use A_s^j and P_s^j to sign the static anchor and its positional embedding in j-th decoder layer, even though it is never updated. With the explicit initialization, regional priors are introduced to guide the decoder in generating non-overlapped proposals.

Anchor Pair Update. Although introducing regional guidance by explicit initialization, maintaining the guidance during decoding iterations is also important. Following this idea, static anchors are designed to maintain guidance without updating, while dynamic anchors are designed to update for localization as shown in Fig. 3. For static anchors,

$$A_s^{j+1} = A_s^0 = A, \quad P_s^{j+1} = P_s^0 = \text{MLP}(\text{PE}(A)). \tag{4}$$

Given dynamic anchors $A_d^j = (a_c^j, a_\sigma^j)$ in j-th decoder layer and the relative positions $\Delta A_d^j = (\Delta a_c^j, \Delta a_\sigma^j)$ from a prediction head, the dynamic anchors are updated as:

$$\begin{aligned} A_d^{j+1} &= A_d^j + \Delta A_d^j = (a_c^j + \Delta a_c^j,\ a_\sigma^j + \Delta a_\sigma^j), \\ P_d^{j+1} &= \text{MLP}(\text{PE}(A_d^{j+1})). \end{aligned} \tag{5}$$

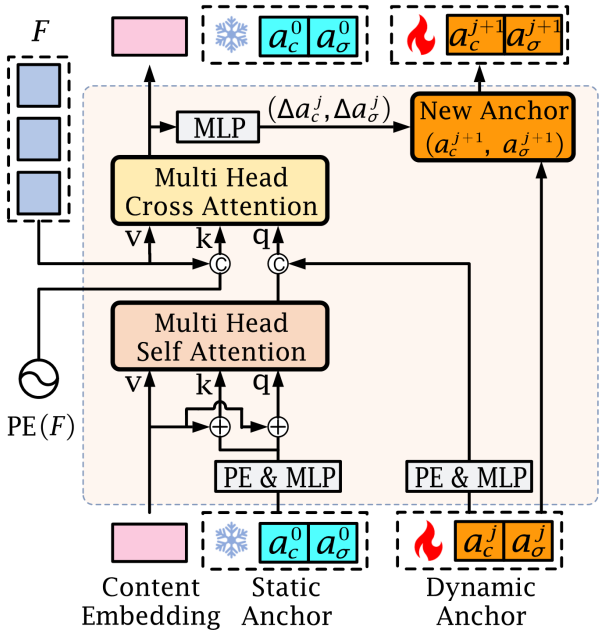

Figure 3: The structure of our proposed region-guided decoder with anchor pair (static anchor and dynamic anchor), where PE means positional encoding.

Note that all prediction heads share the same parameters.

Region-Guided Attention Module. Similar to the general decoder, our region-guided decoder also includes two parts: self-attention module and cross-attention module. However, we employ different anchors in two modules for varying roles, as shown in Fig. 3. In the self-attention module, static anchors are utilized to focus content embeddings on preset representative temporal regions and share information across different regions. Specifically, we utilize static anchors as the positional embedding of self-attention module, such that the updated content embedding C_s^j is as follows:

$$C_s^j = \text{MultiHeadAttn}(C^{j-1} + P_s^0, C^{j-1} + P_s^0, C^{j-1}), \tag{6}$$

where $C^{j-1} \in \mathbb{R}^{\mathcal{K}\times D}$ is the content embedding from $(j-1)$-th decoder layer, and C^0 is initialized to zeros. In the cross-attention module, we employ dynamic anchors as query positional embedding to aggregate region-specific features from fusion embedding F with the assistance of C_s^j. Therefore, the content embedding is updated as:

$$C^j = \text{MultiHeadAttn}([C_s^j, P_d^j],\ [F, \text{PE}(F)],\ F), \tag{7}$$

where $[\cdot,\cdot]$ means concatenation function. By adopting anchor pairs with regional guidance, the decoder reduces the optimization difficulty and eliminates overlapped proposals.

3.4 IoU-Aware Scoring Head

The region-guided decoder improves the quality of proposals by reducing overlapped and redundant proposals, while high-quality proposals demand not only fewer duplications but also accurate boundaries. In the previous DETR-based methods (Jang et al. 2023; Sun et al. 2024), classification

Method	test					val				
	R1		mAP			R1		mAP		
	@0.5	@0.7	@0.5	@0.75	Avg.	@0.5	@0.7	@0.5	@0.75	Avg.
M-DETR (Lei, Berg, and Bansal 2021)	52.89	33.02	54.82	29.40	30.73	53.94	34.84	-	-	32.20
QD-DETR (Moon et al. 2023b)	62.40	44.98	62.52	39.88	39.86	62.68	46.66	62.23	41.82	41.22
UniVTG (Lin et al. 2023)	58.86	40.86	57.60	35.59	35.47	59.74	-	-	-	36.13
TR-DETR (Sun et al. 2024)	64.66	48.96	63.98	43.73	42.62	67.10	51.48	66.27	46.42	45.09
TaskWeave (Yang et al. 2024)	-	-	-	-	-	64.26	50.06	65.39	46.47	45.38
UVCOM (Xiao et al. 2024)	63.55	47.47	63.37	42.67	43.18	65.10	51.81	-	-	45.79
CG-DETR (Moon et al. 2023a)	65.43	48.38	64.51	42.77	42.86	67.35	52.06	65.57	45.73	44.93
LLMEPET† (Jiang et al. 2024)	**66.73**	**49.94**	65.76	43.91	44.05	66.58	51.10	-	-	46.24
RGTR (Ours)	65.50	49.22	**67.12**	**45.77**	**45.53**	**67.68**	**52.90**	**67.38**	**48.00**	**46.95**

Table 1: Performance Comparison on QVHighlights *test* and *val* splits. † indicates LLM-based method.

Method	TACoS				Charades-STA			
	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
M-DETR (Lei, Berg, and Bansal 2021)	37.97	24.67	11.97	25.49	65.83	52.07	30.59	45.54
MomentDiff (Li et al. 2024)	44.78	33.68	-	-	-	55.57	32.42	-
UniVTG (Lin et al. 2023)	51.44	34.97	17.35	33.60	70.81	58.01	35.65	50.10
CG-DETR (Moon et al. 2023a)	52.23	39.61	22.23	36.48	70.43	**58.44**	36.34	50.13
LLMEPET† (Jiang et al. 2024)	52.73	-	22.78	36.55	70.91	-	**36.49**	50.25
RGTR (Ours)	**53.04**	**40.31**	**24.32**	**37.44**	**72.04**	57.93	35.16	**50.32**

Table 2: Performances Comparison on TACoS and Charades-STA. † indicates LLM-based method.

confidence (foreground or background) is adopted to rank all proposals. However, a single binary classification score may inadequately assess proposal quality by overlooking temporal boundary accuracy. To distinguish high-quality proposals, we introduce an IoU-aware scoring head, which considers both localization quality and classification confidence.

Specifically, the output of the decoder is fed to an FFN and a linear layer to predict the moment span and the confidence score p_c. Additionally, we add a linear layer to predict the expected IoU p_{IoU}. Instead of scoring proposals by classification confidence alone, we score them by a joint combination of confidence and IoU score, *i.e.*, the product between p_c and p_{IoU}. We supervise the IoU score with an L2 loss to the ground-truth IoU, denoted as $\hat{g}_{\text{IoU}}$,

$$\mathcal{L}_{\text{IoU}} = || \, p_{\text{IoU}} - \hat{g}_{\text{IoU}} \, ||^2 . \quad (8)$$

This additional IoU score can explicitly incorporate localization quality to enhance classification confidence estimation, thereby generating high-quality proposals. Additionally, non maximum suppression (NMS) is applied during inference.

3.5 Training Objectives

The objective losses of RGTR include four parts: moment loss $\mathcal{L}_{\text{mom}}$, saliency loss $\mathcal{L}_{\text{sal}}$, alignment loss $\mathcal{L}_{\text{align}}$ and IoU loss $\mathcal{L}_{\text{IoU}}$. The overall objective is defined as:

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{mom}} + \lambda_{\text{sal}}\mathcal{L}_{\text{sal}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{IoU}}\mathcal{L}_{\text{IoU}}, \quad (9)$$

where λ_* are the balancing parameters. $\mathcal{L}_{\text{mom}}$ and $\mathcal{L}_{\text{sal}}$ are consistent with QD-DETR (Moon et al. 2023b).

4 Experiments

4.1 Datasets and Metrics

Datasets. We evaluate the proposed method on three temporal sentence grounding benchmarks, including the QVHighlights (Lei, Berg, and Bansal 2021), Charades-STA (Gao et al. 2017), and TACoS (Regneri et al. 2013). QVHighlights spans various themes, Charades-STA comprises intricate daily human activities, and TACoS mainly showcases long-form videos focusing on culinary activities.
Metrics. We adopt the Recall@1 (R1) under the IoU thresholds of 0.3, 0.5, and 0.7. Since QVHighlights contains multiple ground-truth moments per sentence, we also report the mean average precision (mAP) with IoU thresholds of 0.5, 0.75, and the average mAP over a set of IoU thresholds [0.5: 0.05: 0.95]. For Charades-STA and TACoS, we compute the mean IoU of top-1 predictions.

4.2 Implementation Details

Following previous methods (Moon et al. 2023b), we use SlowFast and CLIP to extract visual features and CLIP to extract textual features. We set the embedding dimension D to 256. The number of anchor pairs $\mathcal{K}$ is set to 20 for QVHighlights, 10 for Charades-STA and TACoS. The NMS threshold is set to 0.8. The balancing parameters are set as: $\lambda_{\text{align}} = 0.3$, $\lambda_{\text{iou}} = 1$, and λ_{sal} is set as 1 for QVHighlights, 4 for Charades-STA and TACoS. We train all models with batch size 32 for 200 epochs using the AdamW optimizer with weight decay 1e-4. The learning rate is set to 1e-4.

Method	R0.5	R0.7	mAP$_{avg}$
Charades-STA-Len			
2D-TAN (Zhang et al. 2020)	28.68	17.72	22.79
MMN (Wang et al. 2022b)	34.31	19.94	26.85
QD-DETR† (Moon et al. 2023b)	54.06	32.53	36.37
MomentDiff (Li et al. 2024)	38.32	23.38	28.19
RGTR	**61.17**	**40.23**	**44.30**
Charades-STA-Mom			
2D-TAN (Zhang et al. 2020)	20.44	10.84	17.23
MMN (Wang et al. 2022b)	27.20	14.12	19.18
QD-DETR† (Moon et al. 2023b)	46.31	28.65	30.46
MomentDiff (Li et al. 2024)	33.59	15.71	21.37
RGTR	**49.81**	**29.77**	**33.19**

Table 3: Results on two out-of-distribution splits of Charades-STA. The VGG and Glove features are employed for all models. † indicates reproduced by official codebase.

Setting	AEI	RGAM	IASH	R0.5	R0.7	mAP$_{avg}$
(a)				65.35	48.97	43.12
(b)	✓			64.65	50.58	44.82
(c)			✓	66.19	49.61	44.03
(d)	✓	✓		65.55	51.29	45.36
(e)	✓		✓	66.13	51.68	46.51
(f)	✓	✓	✓	**67.68**	**52.90**	**46.95**

Table 4: Ablation study on the components of RGTR on QVHighlights *val* split. It investigates the anchor explicit initialization (AEI), the region-guided attention module (RGAM), and the IoU-aware scoring head (IASH).

4.3 Performance Comparison

As shown in Tab. 1, we compare RGTR to previous methods on QVHighlights. For a fair comparison, we report numbers for both the test and validation splits. Our method achieves new state-of-the-art performance on almost all metrics. Specifically, RGTR outperforms the latest methods like LLMEPET (Jiang et al. 2024), achieving 67.12% at mAP@0.5 and 45.53% at mAP$_{avg}$ on the test split. On the validation split, RGTR also maintains its lead. The notable performance advantages of RGTR demonstrate the effectiveness of anchor pairs with explicit regional guidance.

Tab. 2 presents comparisons on TACoS and Charades-STA. Our method achieves the best performance on TACoS. On Charades-STA, RGTR also maintains its competitiveness. However, we observe that while our results are notably superior on QVHighlights, the margin is slightly reduced on TACoS and Charades-STA. We attribute this to the biased distribution of the two datasets compared to QVHighlights, resulting in less query diversity learned by anchor pairs.

4.4 Experiments on Out-of-Distribution Splits

To measure robustness, we also evaluate RGTR on two out-of-distribution splits (Li et al. 2024), Charades-STA-Len and Charades-STA-Mom, with distribution shifts of the length and moment location between training and test sets respectively. Since our anchor pairs are initialized by clustering centers on the training set, performance may degrade when the distribution changes significantly. However, as shown in Tab. 3, RGTR outperforms all previous methods under both out-of-distribution settings. Such surprising results indicate that the regional guidance introduced by anchor pairs works more by increasing the diversity between moment queries, rather than merely relying on the similarity between the training and test set distributions. Ablation experiments on other anchor initialization methods in Tab. 5 also confirm this point. Even with uniform grid points for initialization, which also serve as an initialization method to increase query diversity but are unrelated to the dataset distribution, the model performance improves significantly. Therefore, despite a distribution shift, the query diversity from regional guidance remains crucial for effective localization.

Method	Changes	R0.5	R0.7	mAP$_{avg}$
Initialization Method	random	66.19	49.61	44.03
	uniform grid	67.10	50.97	44.93
	k-means	**67.68**	**52.90**	**46.95**
Scoring Method	IoU superv.	**67.87**	52.84	46.54
	cls + IoU	67.23	52.39	46.92
	cls × IoU	67.68	**52.90**	**46.95**

Table 5: Ablation study on initialization and scoring method.

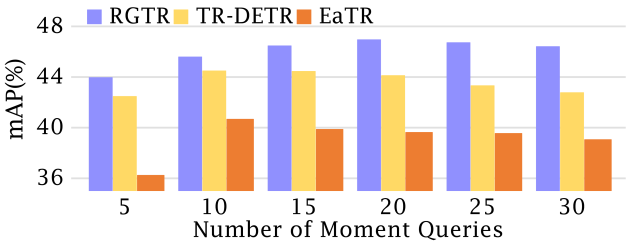

Figure 4: Ablation study on number of moment queries $\mathcal{K}$.

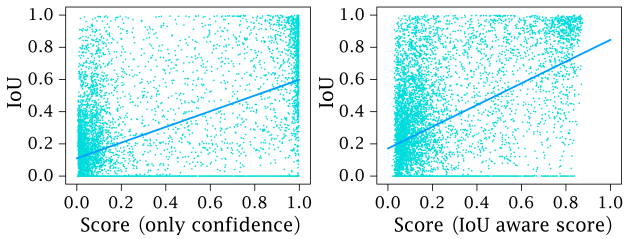

Figure 5: Correlation between scores and ground-truth IoUs.

4.5 Ablation Study

Main Ablation. We first investigate the effectiveness of each component in RGTR. As shown in Tab. 4, we report the impact according to anchor explicit initialization, region-guided attention module, and IoU-aware scoring head. No-

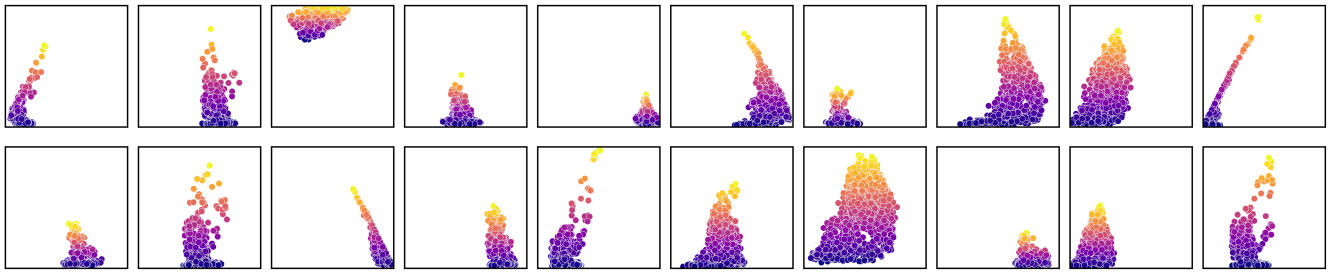

Figure 6: Visualization of moment predictions on QVHighlights *val* split, for all 20 dynamic anchors in region-guided decoder.

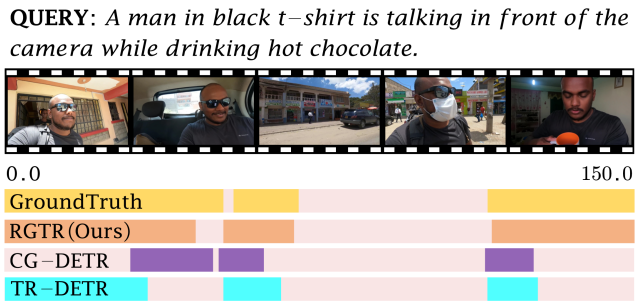

Figure 7: The qualitative result on QVHighlights.

tably, setting (b) represents the decoder only uses the explicitly initialized dynamic anchors, while setting (d) utilizes both static and dynamic anchors during the decoding process. The results demonstrate that each component contributes significantly to overall performance and setting (f) improves performance by 3.93% in terms of R1@0.7 and 3.83% in terms of mAP_{avg} by using all components.

Anchor Initialization Method. We adopt another two simple initialization methods to replace the k-means algorithm. "random" means utilizing random learnable queries as moment queries. "uniform grid" means generating a uniform grid on the normalized $m_c \times m_\sigma$ area, and uniformly sampling $5 \times 5 = 25$ points in a practical temporal region. As shown in Tab. 5, the performance of k-means initialization is significantly better than random initialization and uniform grid initialization. It verifies that k-means algorithm can provide optimal explicit regional priors for decoding process.

Scoring Method. Tab. 5 compares the product fusion with other scoring methods, where IoU superv. means only using confidence score with IoU loss as supervision. All methods have significant performance improvements, among which the product method achieves the best performance.

Number of Moment Queries. In previous methods, the number of moment queries $\mathcal{K}$ is typically limited to 10. This is because increasing $\mathcal{K}$ without explicit guidance only produces more overlapped proposals, resulting in negligible performance improvement or even degradation. In contrast, our method provides explicit regional guidance for each moment query, *i.e.*, each moment query is accountable for a specific temporal region. Therefore, increasing $\mathcal{K}$ allows moment queries to cover more temporal regions, leading to effective prediction. As shown in Fig. 4, we present the performance of EaTR, TR-DETR, and our RGTR in terms of mAP_{avg} according to $\mathcal{K}$. We re-implement the other two methods in different $\mathcal{K}$. As discussed above, for TR-DETR and EaTR, performance peaks when $\mathcal{K}$ reaches 10 and then declines significantly. In contrast, for RGTR, increasing $\mathcal{K}$ to 20 significantly improves performance, demonstrating the effectiveness of anchor pairs with explicit regional guidance.

Correlation between Score and IoU. To compare IoU-aware scoring and classification confidence scoring, we draw scatter plots of the correlation between scores and ground-truth IoUs on the QVHighlights validation set in Fig. 5. It can be observed that our IoU-aware score shows a stronger correlation with the ground-truth IoU, *i.e.*, the slope of the fitted line increases from 0.49 to 0.67, improving the distinction of high-quality proposals.

4.6 Visualization and Qualitative Result

As shown in Fig. 6, we visualize all 20 dynamic anchors in the region-guided decoder on QVHighlights. Compared with previous methods in Fig. 1, RGTR introduces regional guidance through anchor pairs, effectively enhancing query diversity and eliminating numerous overlapped proposals.

In Fig. 7, we illustrate a qualitative example on QVHighlights, where the sentence corresponds to multiple moment spans. Since our method emphasizes enhancing query diversity, RGTR generates more accurate predictions than other methods, especially in the case of requiring simultaneous attention to different center coordinates and durations.

5 Conclusion

In this paper, we propose a Region-Guided TRansformer (RGTR) framework to address the limitations of DETR structure in TSG task. To eliminate overlapped proposals, we design a region-guided decoder, which adopts a set of anchor pairs as moment queries to introduce explicit regional guidance for decoding process. Each anchor pair takes charge of moment prediction for a specific temporal region, which reduces optimization difficulty and eliminates redundant proposals. To distinguish high-quality proposals, we employ an IoU-aware scoring head that incorporates localization quality to enhance classification confidence estimation. Experiments on three public datasets and two out-of-distribution splits demonstrate the superiority of RGTR.

Acknowledgments

This work was supported in part by National Science and Technology Major Project under Grant 2023ZD0121300, National Natural Science Foundation of China under Grants 62088102, 12326608 and 62106192, Natural Science Foundation of Shaanxi Province under Grant 2022JC-41, and Fundamental Research Funds for the Central Universities under Grant XTR042021005.

References

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chen, L.; Lu, C.; Tang, S.; Xiao, J.; Zhang, D.; Tan, C.; and Li, X. 2020. Rethinking the bottom-up framework for query-based video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 10551–10558.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6202–6211.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, 5267–5275.

Jang, J.; Park, J.; Kim, J.; Kwon, H.; and Sohn, K. 2023. Knowing where to focus: Event-aware transformer for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13846–13856.

Jiang, Y.; Zhang, W.; Zhang, X.; Wei, X.; Chen, C. W.; and Li, Q. 2024. Prior Knowledge Integration via LLM Encoding and Pseudo Event Regulation for Video Moment Retrieval. In *ACM Multimedia 2024*.

Lee, P.; and Byun, H. 2023. BAM-DETR: Boundary-Aligned Moment Detection Transformer for Temporal Sentence Grounding in Videos. *arXiv preprint arXiv:2312.00083*.

Lei, J.; Berg, T. L.; and Bansal, M. 2021. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705.

Li, P.; Xie, C.-W.; Xie, H.; Zhao, L.; Zhang, L.; Zheng, Y.; Zhao, D.; and Zhang, Y. 2024. Momentdiff: Generative video moment retrieval from random to real. *Advances in neural information processing systems*, 36.

Lin, K. Q.; Zhang, P.; Chen, J.; Pramanick, S.; Gao, D.; Wang, A. J.; Yan, R.; and Shou, M. Z. 2023. Univtg: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2794–2804.

Liu, D.; Qu, X.; Di, X.; Cheng, Y.; Xu, Z.; and Zhou, P. 2022a. Memory-guided semantic learning network for temporal sentence grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1665–1673.

Liu, M.; Wang, X.; Nie, L.; Tian, Q.; Chen, B.; and Chua, T.-S. 2018. Cross-modal moment localization in videos. In *Proceedings of the 26th ACM international conference on Multimedia*, 843–851.

Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; and Zhang, L. 2022b. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*.

Liu, Y.; Li, S.; Wu, Y.; Chen, C.-W.; Shan, Y.; and Qie, X. 2022c. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3042–3051.

Lu, C.; Chen, L.; Tan, C.; Li, X.; and Xiao, J. 2019. Debug: A dense bottom-up grounding approach for natural language video localization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5144–5153.

Moon, W.; Hyun, S.; Lee, S.; and Heo, J.-P. 2023a. Correlation-guided Query-Dependency Calibration in Video Representation Learning for Temporal Grounding. *arXiv preprint arXiv:2311.08835*.

Moon, W.; Hyun, S.; Park, S.; Park, D.; and Heo, J.-P. 2023b. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23023–23033.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Regneri, M.; Rohrbach, M.; Wetzel, D.; Thater, S.; Schiele, B.; and Pinkal, M. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1: 25–36.

Shi, L.; Wang, L.; Zhou, S.; and Hua, G. 2023. Trajectory unified transformer for pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9675–9684.

Shi, S.; Jiang, L.; Dai, D.; and Schiele, B. 2022. Motion transformer with global intention localization and local movement refinement. *Advances in Neural Information Processing Systems*, 35: 6531–6543.

Sun, H.; Zhou, M.; Chen, W.; and Xie, W. 2024. TR-DETR: Task-Reciprocal Transformer for Joint Moment Retrieval and Highlight Detection. *arXiv preprint arXiv:2401.02309*.

Wang, Y.; Zhang, X.; Yang, T.; and Sun, J. 2022a. Anchor detr: Query design for transformer-based detector. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 2567–2575.

Wang, Z.; Wang, L.; Wu, T.; Li, T.; and Wu, G. 2022b. Negative sample matters: A renaissance of metric learning for temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2613–2623.

Xia, K.; Wang, L.; Zhou, S.; Zheng, N.; and Tang, W. 2022. Learning to refactor action and co-occurrence features for temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13884–13893.

Xiao, Y.; Luo, Z.; Liu, Y.; Ma, Y.; Bian, H.; Ji, Y.; Yang, Y.; and Li, X. 2024. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18709–18719.

Yang, J.; Wei, P.; Li, H.; and Ren, Z. 2024. Task-Driven Exploration: Decoupling and Inter-Task Feedback for Joint Moment Retrieval and Highlight Detection. *arXiv preprint arXiv:2404.09263*.

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

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Zhang, S.; Peng, H.; Fu, J.; and Luo, J. 2020. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12870–12877.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.