Sim-DETR: Unlock DETR for Temporal Sentence Grounding

Jiajin Tang[1*], **Zhengxuan Wei**[1*], **Yuchen Zhu**[1], **Cheng Shi**[1], **Guanbin Li**[2], **Liang Lin**[2], **Sibei Yang**[2†]
[1]ShanghaiTech University [2]School of Computer Science and Engineering, Sun Yat-sen University
{tangjj, weizhx2022}@shanghaitech.edu.cn yangsb3@mail.sysu.edu.cn

Abstract

Temporal sentence grounding aims to identify exact moments in a video that correspond to a given textual query, typically addressed with detection transformer (DETR) solutions. However, we find that typical strategies designed to enhance DETR do not improve, and may even degrade, its performance in this task. We systematically analyze and identify the root causes of this abnormal behavior: (1) conflicts between queries from similar target moments and (2) internal query conflicts due to the tension between global semantics and local localization. Building on these insights, we propose a simple yet powerful baseline, Sim-DETR, which extends the standard DETR with two minor modifications in the decoder layers: (1) constraining self-attention between queries based on their semantic and positional overlap and (2) adding query-to-frame alignment to bridge the global and local contexts. Experiments demonstrate that Sim-DETR unlocks the full potential of DETR for temporal sentence grounding, offering a strong baseline for future research.

1. Introduction

Video has become a dominant media on the internet, with short-form content rapidly expanding and achieving exponential growth in reach over recent years [18–20, 24]. Instead of passively watching entire videos, users now prefer to target specific segments of interest, with natural language descriptions serving as an intuitive and flexible way to convey intent [30, 61, 68]. This amplifies interest in the research topic of temporal sentence grounding [22, 25–27, 32, 40, 46, 47, 57, 71, 81], which aims to locate one or more semantically aligned segments within an untrimmed video according to a given natural language sentence, as shown in Figure 1a.

Early methods either align sentences with predefined temporal proposals for selection [44, 70] or directly predict moment spans through cross-modal interactions between language and frames [36, 37, 72]. Recent advances center on integrating the Detection Transformer (DETR) [2] for object detection into the temporal sentence grounding, leveraging its query-based proposal detection framework to eliminate hand-crafted proposal components and deliver superior grounding performance with high efficiency [22, 25, 26, 46, 47, 57]. They leverage moment queries—often several times more numerous than ground truth segments—in the multi-layer decoder, which are initialized randomly [26, 46, 57], generated from sentences [22, 47], or extracted as event units from the video [25], to locate and search the most suitable matches for target segments. A one-to-one label assignment is applied in the training objective to facilitate non-redundant predictions for each ground-truth segment. The overall architecture of the DETR-based temporal sentence grounding framework is shown in Figure 6a.

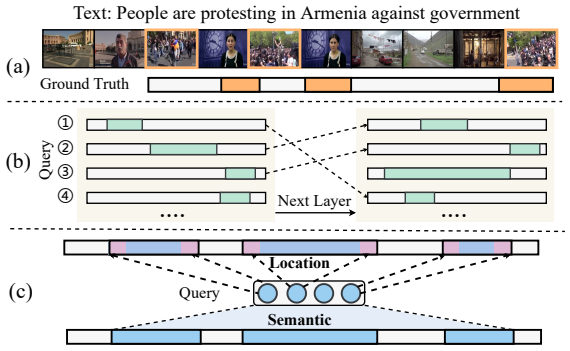

Figure 1. Illustration of (a) temporal sentence grounding task, (b) phenomenon of random matching across multiple layers, and (c) the challenge of achieving the balance between focus on boundary for span prediction and alignment with global semantics.

However, we observe that strategies generally effective for DETR in other tasks, such as object detection, like increasing the number of queries or decoder layers, do not improve performance in temporal sentence grounding and may even degrade it (Sec 3.1). This motivates us to revisit the specific characteristics of temporal sentence grounding that give rise to the distinct challenges in applying DETR to this task. In a series of investigations on query-segment alignment (Sec 3.2), we verify that performance degradation primarily stems from the difficulty in identifying ***dis-***

* Equal contribution. † Corresponding author is Sibei Yang.

tinct local boundaries for target segments that share ***highly similar global semantics***, ***leading to inherent conflicts both between and within queries.*** By diagnosing the similarities between queries and between queries and frames, we uncover two conflicts that result in learning puzzles for queries:

- *Queries corresponding to distinct target segments exhibit high similarity, leading to random matching with various targets and creating learning puzzles for the queries.* The presence of multiple semantically similar target segments results in the high similarity between their corresponding queries, making them difficult to clearly distinguish (Sec 3.3). This causes each query to potentially match well with multiple segments, leading to significant randomness in the one-to-one matching for each target segment (Figure 1b). We observe that across different decoder layers, the queries matched for optimization can vary considerably (Sec 3.2).
- *Each query faces an inherent conflict between encoding the global semantics of its corresponding segment and decoding its local boundary.* To align well with a segment, a query must encode the segment's global semantics from the start to the end frame, such as the entire process of a person holding a cup. However, the query also needs to directly predict the segment's boundary, requiring focus on the local frames near the boundary (Figure 1c). This creates a trade-off: the query either prioritizes global semantics or local boundaries, making it challenging to optimize both simultaneously (Sec 3.3).

Building on these findings, we propose a simple yet strong baseline for video temporal grounding, Sim-DETR. The framework builds on the standard DETR-based temporal sentence grounding architecture, ***with two minor yet straightforward design modifications to the decoder layers*** to address the two conflicts between and within queries (Figure 6), respectively. First, we adjust the self-attention weights between queries based on their pairwise correlation, encouraging indistinguishable queries to focus on different contexts, reducing their similarity, and allowing the most suitable query to gather information from related queries to refine its prediction (Figure 6b). Second, we introduce a query-to-frame matching and loss term, accounting for the alignment between the query and each frame within the segment, as shown in Figure 6c. This ensures that all frames, not just those near the boundary, contribute to segment localization. The full sequence of all target frames, from start to finish, acts as a bridge connecting the global semantics to the local boundary.

Empirically, we indeed observe that our minor modifications (1) significantly reduce conflicts between queries corresponding to different target segments Figure 3), (2) improve the alignment between the query's global semantic attention to frames and its boundary prediction (Figure 5), and (3) lead to more consistent query predictions across layers (Figure 4). Furthermore, (4) increasing the number of queries and decoder layers no longer causes performance degradation (Figure 2). Finally, due to the consistent matching and prediction of queries, (5) our Sim-DETR not only achieves substantial performance gains but also accelerates convergence (as shown in Appendix). All the above observations and advantages, including the minor modifications, are intended to unlock the potential of query-based frameworks for temporal sentence grounding and provide a simple yet strong baseline for future research.

Our contributions are summarized as follows:

- We systematically analyze the root causes of abnormal behavior in the DETR-based temporal sentence grounding framework, identifying two key conflicts: (1) target segments with highly similar semantics but distinct temporal localization create learning puzzles between queries, and (2) a trade-off between relying on global semantics for matching and local boundaries for localization within queries.
- Based on our analysis, we propose two modifications to form Sim-DETR: (1) a simple adjustment to the self-attention to resolve conflicts between queries and (2) the introduction of a query-to-frame matching to reconcile global semantics and local localization within queries.
- Our Sim-DETR, a simple yet powerful baseline, achieves consistent and significant improvements across all benchmarks. More importantly, it eliminates observed anomalies and exhibits faster convergence.

2. Related Work

Temporal Sentence Grounding (TSG) aims to identify specific video moments that align with given language expressions. As computer vision advances across neural architectures, generation, detection, segmentation, and multimodal tasks [4, 8, 21, 33, 53–55, 73, 82], temporal understanding requires task-specific optimizations. Early methods fall into proposal-free and proposal-based categories. Proposal-free methods use end-to-end frameworks by directly regressing boundary coordinates [6, 43, 76] or predicting frame-level boundary likelihood [16, 78]. Proposal-based approaches densely sample candidate segments via sliding windows [1, 12, 15, 23, 38, 80], temporal anchors [5, 32, 41, 75, 77, 81], or multimodal feature similarity [7, 34, 35, 66, 69]. Recently, DETR has shown exceptional effectiveness in detection tasks. [27, 65] pioneered DETR for TSG. Subsequent studies enhance cross-modal representations [46, 47, 57, 67], improve decoder queries with semantic/spatial information [22, 25, 58], or explore joint training with other temporal tasks [57, 71].

Detection Transformers (DETR) [2] is a pioneering end-to-end framework for object detection and segmentation that employs transformers to make direct set-based pre-

dictions [14, 52, 54, 55, 60, 61, 85]. However, DETR has notable limitations, especially its high computational complexity and slow convergence, impacting its practical efficiency. To address these issues, several derivative models have emerged. Some methods [51, 63, 74, 84] improve computational efficiency by reducing redundant calculations within the DETR architecture. Other models [9, 13, 59] refine the attention mechanism for more efficient information processing and resource utilization. Additionally [39, 45, 64] enhance convergence by embedding spatial information directly into the query design, which accelerates object localization. Recent models [28, 79] tackle optimization challenges in DETR's bipartite matching process by integrating denoising strategies during training, improving both convergence and performance. Despite recent advancements in enhancing DETR for object detection, there is a lack of systematic studies addressing the unique challenges of applying DETR to TSG and the potential difficulties in query optimization.

3. Probing DETR's Inherent Behavior for TSG

This study begins with the observation that enhancements effective for object detection in DETR [2] do not apply to temporal sentence grounding (TSG), as discussed in Sec 3.1. To explore the reasons behind DETR's failure to enhance, this section presents preliminary studies to reveal the behavior of queries in the DETR decoder, as queries are central to predicting proposals. Specifically, we examine the similarity between queries and outputs in Sec 3.2, and the attention between queries and inputs (*i.e.*, frames) in Sec 3.3. These help reveal the underlying cause, and our modification addressing it not only eliminates the abnormal phenomena but also has a positive side effect (Sec 3.4).

Experimental Settings. Unless otherwise specified, our default experimental setup involves conducting statistical analyses on the full validation set of QVHighlights [27]. CG-DETR [46] and TaskWeave [71], which serve as baselines for DETR-based TSG due to their strong performance while retaining the core vanilla DETR mechanism, are used with their standard configurations.

3.1. Abnormal Phenomena

Motivation. DETR is originally proposed for object detection, and increasing the number of queries and decoder layers typically improves performance. We explore whether this trend also applies to TSG.

Setting. Both CG-DETR and TaskWeave default to 10 queries, with 3 and 2 decoder layers, respectively. This experiment specifically increases the number of queries and decoder layers to isolate their effects, keeping all other hyperparameters and training protocols constant.

Result & Discussion. As shown in Figure 2, increasing the number of queries to 15 and 20 led to progressively larger performance declines, with a drop of over -2% with TaskWeave [71] at 20 queries. Similarly, increasing the number of decoder layers caused a drop from -1% to -2.5%. In contrast, our Sim-DETR mitigates this phenomenon and achieves a slight improvement, demonstrating its robustness. This observation motivates an in-depth exploration of decoder layer, particularly the role of queries (see Sec 3.2).

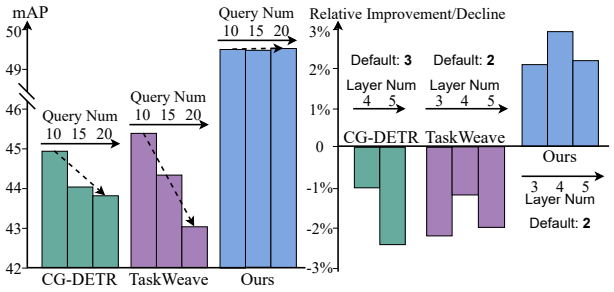

Figure 2. The impact of the number of queries and decoder layers.

3.2. Similar Target Segments Cause Query Conflicts

Motivation. Initially, we hypothesize that the lack of performance improvement with increased queries is due to query redundancy. However, performance declined, suggesting that conflicts between queries, rather than redundancy, prevent compatibility with existing queries. Similarly, conflicts between queries across layers may explain the ineffectiveness of increasing the number of layers. This motivates us to explore (1) the relationships between queries to identify their potential conflicts and (2) their variations across layers to reveal the impact at different layers.

Setting. (1) To evaluate query relationships, we assess similarities between queries corresponding to the same segment (intra-segment similarity) and those from different segments (inter-segment similarity). Specifically, we identify the corresponding segment for each query based on the maximum IoU between the query's prediction and all ground-truth (GT) segments during inference. Queries whose IoU with the corresponding segment is 0.5 or greater are included in our analysis to ensure meaningful corresponding. We then compute feature similarities between queries associated with the same segment (intra) and those matched to different segments (inter). (2) To analyze variations across layers, we introduce the metric of cross-layer matching consistency. Specifically, during training optimization, each GT segment is uniquely matched to a query through bipartite matching at each decoding layer. We assess match consistency by measuring the proportion of segments that retain their corresponding queries across two consecutive decoder layers.

Result & Discussion. (1) As shown in Figure 3, CG-DETR fails to distinguish query similarities between inter- or intra-segment groups, indicating that queries are not clearly grouped with their corresponding segments. ***Consequently, the same query oscillates between segments,***

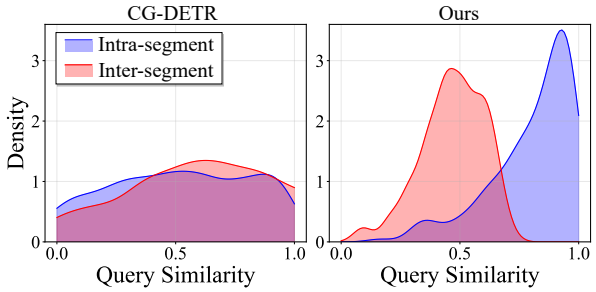

Figure 3. Distribution of query similarity for intra-segment (blue) and inter-segment (red) pairs. Our method can effectively distinguish intra-segment and inter-segment queries, ensuring more stable query-segment associations and reducing conflicts in query assignments across different segments.

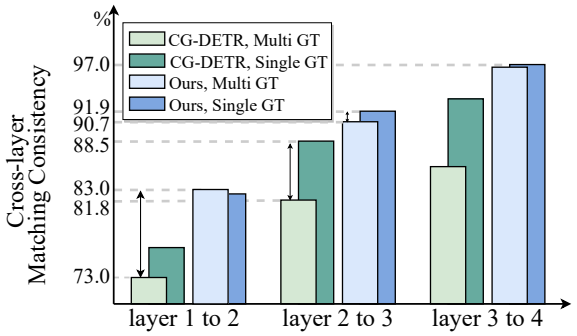

Figure 4. Measuring cross-layer matching consistency (y-axis, defined in Line-260) of segments across two consecutive decoding layers (x-axis). Our method achieves higher consistency.

lacking stable and consistent corresponding. Figure 4 further supports our hypothesis, demonstrating that cross-layer matching consistency in CG-DETR is lower than ours. (2) We attribute this to multiple GT segments in a video sharing the same linguistic semantics. ***The high semantic similarity between GT segments results in high similarity between their corresponding queries.*** Figure 4 supports this, showing that the consistency for cases with multiple GT segments is lower than for those with a single segment.

3.3. Global Matching vs. Local Localization

Motivation. We further question whether the inconsistency in matching is solely due to query conflicts from multiple GT segments. If so, why does the previous method still show roughly 25% mismatched queries in cases with a single target segment? Therefore, we shift our focus from conflicts between queries to conflicts within individual queries. A query serves dual roles: aligning with global linguistic semantics (global matching) and precisely localizing the segment, particularly its boundaries (local localization). This prompts us to explore whether conflicts arise from these roles within a single query.

Setting. (1) To evaluate global matching, we use the query's classification confidence score, as it reflects the confidence of the query being referenced by the global semantics of the linguistic description. (2) To evaluate local localization, we extract attention scores between each query and frame from the cross-attention module in the last decoder layer, as they indicate query-based frame localization. We then compute the IoU score between this attention and the query's corresponding GT segment, reflecting the accuracy of the query's localization at the frame level.

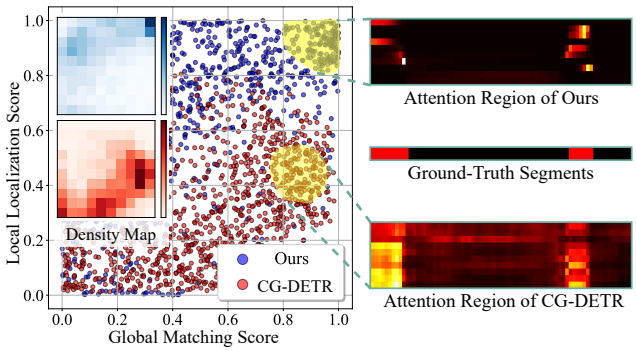

Figure 5. Global matching score (defined in Line-290) vs. local localization score (defined in Line-293). Compared to CG-DETR, our method concentrates a query's attention significantly more within a single GT segment, rather than dispersing it across multiple segments, leading to improved local localization scores.

Result & Discussion. As shown in Figure 5, CG-DETR (marked in red) may yield low local localization scores despite high global matching scores, indicating that ***while queries align well with linguistic descriptions, they fail to achieve precise localization.*** Further analysis reveals that query attention may span frames in multiple GT segments (see Figure 5 "Attention"), leading to a low local score for the query's specific GT segment it corresponds to.

3.4. Positive Side Effect

Convert Abnormal to Normal: Our Sim-DETR effectively (1) distinguishes queries within the same target segment and across different segments in Figure 3, (2) achieves consistent matching across decoder layers in Figure 4, and (3) aligns with global semantics while ensuring precise localization in Figure 5.

Positive Side Effect: (1) More importantly, Sim-DETR maintains stable and robust performance with increased queries and decoder layers (see Figure 2). (2) By eliminating the abnormal phenomena, it also achieves significantly faster convergence compared to previous methods, as shown in Appendix C.

4. Method

We present a simple yet effective TSG baseline that resolves query conflicts in existing DETR-based TSG frameworks [22, 25, 27, 46, 47, 57, 67, 71] through two minor yet key modifications. First, we introduce the standard DETR-based [2] TSG architecture (Sec 4.1). Sec 4.2 introduces our first modification, which adjusts the self-attention

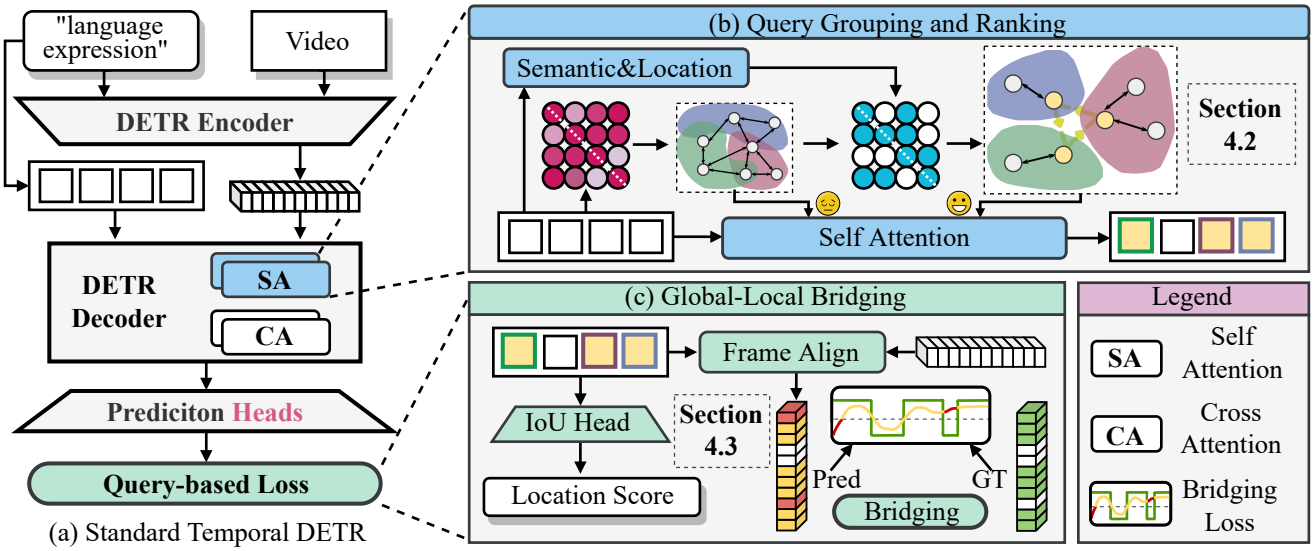

Figure 6. Framework of the proposed Sim-DETR, featuring two simple yet powerful modifications: Query Grouping and Ranking to mitigate cross-query conflicts and Global-Local Bridging to address global-local conflicts within each query.

mechanism between queries in the decoder to avoid conflicts among them, based on the observations presented in Sec 3.2. Next, Sec 4.3 presents another improvement that addresses the conflict between global matching and local localization within the query (discussed in Sec 3.3) by introducing query-to-frame matching loss, which serves as the bridge between the global and local perspectives.

4.1. Standard DETR-based TSG Architecture

In TSG, the goal is to identify video segments with temporal spans that correspond to a given language expression.

Feature Extraction. Previous DETR-based works [25, 47, 57, 71] employ CLIP [49] and SlowFast [10] to extract video features, represented as $\mathcal{T} \in \mathbb{R}^{N\times C}$, where N and C stands for the number of frames and the hidden dimension, respectively. For the language expression, typically a sentence with L tokens, the CLIP text encoder is used to obtain token-level features, resulting in $\mathcal{W} \in \mathbb{R}^{L\times C}$.

Multimodal Encoder. After feature extraction, DETR-based TSG methods [22, 46, 47, 57] employ a multimodal encoder to fuse these video and language features, producing multimodal video features, denoted as $\hat{\mathcal{T}} \in \mathbb{R}^{N\times C}$. The key component of the multimodal encoder is a cross-attention mechanism [62], where video features $\mathcal{T}$ act as queries, while word features $\mathcal{W}$ serve as keys and values.

Temporal Sentence Decoding. In the decoder, the model initializes a set of query to gather global information from the video, which is then decoded into local spans in the span prediction head. Specifically, given initialized queries $\mathcal{Q} \in \mathbb{R}^{M\times C}$, where M is the query number, and the multimodal video features $\hat{\mathcal{T}}$, the decoder facilitates interaction between queries and gathers global information through interleaved self-attention and cross-attention modules, as follows:

$$
\begin{aligned}
&\begin{cases}
\mathcal{Q}_q^{\text{SA}} = \text{FC}_q^{\text{SA}}(\mathcal{Q}), \mathcal{Q}_k^{\text{SA}} = \text{FC}_k^{\text{SA}}(\mathcal{Q}), \mathcal{Q}_v^{\text{SA}} = \text{FC}_v^{\text{SA}}(\mathcal{Q}), \\
\mathcal{Q}^{\text{SA}} = \text{FC}_o^{\text{SA}}\left(\text{softmax}\left(\frac{\mathcal{Q}_q^{\text{SA}}(\mathcal{Q}_k^{\text{SA}})^\intercal}{\sqrt{C}}\right)\mathcal{Q}_v^{\text{SA}}\right).
\end{cases} \\
&\begin{cases}
\mathcal{Q}_q^{\text{CA}} = \text{FC}_q^{\text{CA}}(\mathcal{Q}^{\text{SA}}), \hat{\mathcal{T}}_k = \text{FC}_k^{\text{CA}}(\hat{\mathcal{T}}), \hat{\mathcal{T}}_v = \text{FC}_v(\hat{\mathcal{T}}), \\
\mathcal{Q}^{\text{CA}} = \text{FC}_o^{\text{CA}}\left(\text{softmax}\left(\frac{\mathcal{Q}_q^{\text{CA}}(\hat{\mathcal{T}}_k)^\intercal}{\sqrt{C}}\right)\hat{\mathcal{T}}_v\right).
\end{cases} \\
&\hat{\mathcal{Q}} = \text{MLP}(\text{LayerNorm}(\mathcal{Q}^{\text{CA}}) + \mathcal{Q}), \\
&B = \mathcal{H}_{\text{span}}(\hat{\mathcal{Q}}), \quad B = \{b_1, \cdots, b_M\},
\end{aligned}
\tag{1}
$$

where $\text{FC}(\cdot)$ denotes the fully connected projection layer, with SA and CA representing self-attention and cross-attention, respectively. Subscripts q, k, and v indicate query, key, and value computations. $\text{MLP}(\cdot)$ represents a nonlinear mapping with an activation function. $\mathcal{H}_{\text{span}}(\cdot)$ denotes the span prediction head, which takes the updated query features $\hat{\mathcal{Q}}$ to predict the span set B, where each $b_i \in \mathbb{R}^2$ represents the (start, end) of a predicted span.

4.2. Query Grouping and Ranking

In this section, we address the query conflicts identified in Sec 3.2: (1) the inability to distinguish between intra- and inter-segment queries (Figure 3), leading to query oscillation between segments without stable correspondence (Line-266), and (2) inconsistencies in segment-query matching across layers, where different queries may be matched to ground-truth (GT) segments at each layer for optimization, hindering effective learning (Figure 4).

To differentiate queries across segments, the challenge arises from GT segments that inherently share identical lin-

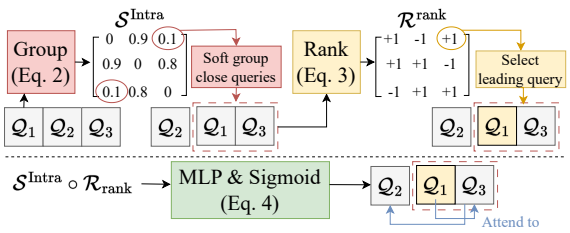

Figure 7. Overview of our Query Grouping and Ranking.

guistic semantics, leading to highly similar queries. Therefore, a natural solution is to directly group queries, assigning each group to a specific GT segment for training and prediction. Instead of hard grouping with predefined rules or extra hyperparameters, we learn a soft grouping based on predicted temporal spans of queries, where closer spans indicate potential corresponding for the same segment. Specifically, for any two queries q_i and q_j, we compute their span distance, which is then used in Eq. 4 to learn their grouping:

$$\mathcal{S}^{\text{intra}}_{i,j} = \|b_i - b_j\|_2, \tag{2}$$

where $\|\cdot\|_2$ denotes the L2 norm. Notably, we use L2 distance instead of the commonly used L1 because, when two spans are close (normalized distances ≤ 1), L2 distance decreases more sharply than L1, imposing a smaller penalty on minor differences (see Appendix G).

Meanwhile, to enhance cross-layer matching consistency, we draw inspiration from EASE-DETR [14]'s query ranking strategy in object detection. By ranking queries within a segment at each layer, the leading query is prioritized for segment matching in subsequent layers. However, directly ranking queries by prediction scores, as in EASE-DETR, proves ineffective (see Appendix D). As analyzed in Sec. 3.3, high prediction/confidence scores do not guarantee precise predictions and may correspond to poor localization. Instead, we introduce an IoU head to assess localization accuracy and rank queries based on both factors, as follows:

$$R_{\text{rank}}(q_i, q_j) = \begin{cases} +1 & \mathcal{P}^{\text{cls}}_i \circ \mathcal{P}^{\text{IoU}}_i \geq \mathcal{P}^{\text{cls}}_j \circ \mathcal{P}^{\text{IoU}}_j \\ -1 & \mathcal{P}^{\text{cls}}_i \circ \mathcal{P}^{\text{IoU}}_i < \mathcal{P}^{\text{cls}}_j \circ \mathcal{P}^{\text{IoU}}_j \end{cases}, \tag{3}$$

where $\mathcal{P}^{\text{cls}}$ and $\mathcal{P}^{\text{IoU}}$ denote the query's confidence score and IoU prediction, respectively, with subscripts i and j indicating queries q_i and q_j. $R_{\text{rank}}(q_i, q_j) = 1$ indicates that query q_i has a higher priority than query q_j, considering both global alignment and local localization.

Finally, we incorporate query grouping within segments and intra-segment ranking to infer relative relationships, then reshape query-to-query self-attention following [14] to mitigate conflicts, as follows:

$$\mathcal{S}^{\text{attn}} = \text{sigmoid}\left(\text{MLP}(\mathcal{S}^{\text{intra}} \circ \mathcal{R}_{\text{rank}}\right),$$
$$\mathcal{Q}^{\text{SA}} = \text{FC}^{\text{SA}}_o\left(\text{softmax}\left(\left(\frac{\mathcal{Q}^{\text{SA}}_q(\mathcal{Q}^{\text{SA}}_k)^\intercal}{\sqrt{C}}\right)\mathcal{S}^{\text{attn}}\right)\mathcal{Q}^{\text{SA}}_v\right) \tag{4}$$

where $\mathcal{S}^{\text{attn}}$ represents query relationships, with higher values promoting high-quality queries to dominate predictions by integrating similar queries within the same group.

4.3. Global-Local Bridging

In this section, we address the query conflicts identified in Sec 3.3, specifically the conflict between a query's global semantic matching and local localization. Since queries overly rely on global semantics while neglecting local localization, our core idea is to reinforce fine-grained frame-level localization within the segment. Specifically, we introduce a query-to-frame matching mechanism and loss term to enhance localization beyond segment boundary prediction, ensuring alignment with in-segment frames while avoiding matches with those from other segments. The full sequence of frames within the segment, from start to finish, serves as a bridge linking global semantics to local boundaries.

During training, we compute the semantic similarity between each query q_i and all frames within its corresponding ground-truth span, maximizing similarity to ensure full alignment with in-segment frames while minimizing similarity to out-segment frames. The computation process is outlined as follows,

$$z = \text{sigmoid}(\tau \cdot \cos(q_i, \hat{\mathcal{T}})),$$
$$\mathcal{L}_{\text{bridge}} = \lambda_{\text{bridge}} \frac{-\sum_j z_j \mathbb{I}[b^{\text{gt}}_i]_j}{\sum_j z_j(1 - \mathbb{I}[b^{\text{gt}}_i]_j) + \sum_j \mathbb{I}[b^{\text{gt}}_i]_j}, \tag{5}$$

where $z \in \mathbb{R}^N$ donates query-to-frame similarity, calculated using the cosine similarity between the query q_i and the N frame features $\hat{\mathcal{T}}$ defined in Line-347, and τ is a learnable scaling coefficient. b^{gt}_i is the ground-truth span corresponding to query q_i, while $\mathbb{I}[b^{\text{gt}}_i] \in \mathbb{R}^N$ is an indicator function that assigns 1 to frames within the ground truth and 0 otherwise, with j representing the frame index. Thus, the numerator in $\mathcal{L}_{\text{bridge}}$, *i.e.*, $-\sum_j z_j \mathbb{I}[b^{\text{gt}}_i]_j$, encourages the query to be similar to every frame within its corresponding segment. Meanwhile, the denominator's first term, $\sum_j z_j(1 - \mathbb{I}[b^{\text{gt}}_i]_j)$, represents query q_i's similarity to out-segment frames, promoting its minimization. The denominator's second term, $\sum_j \mathbb{I}[b^{\text{gt}}_i]_j$, represents the length of the ground-truth segment, serving as a normalization factor. And λ_{bridge} is a hyperparameter that controls the loss weight.

4.4. Overall Loss

The overall loss is: $\mathcal{L} = \mathcal{L}_{\text{MD}} + \lambda_{\text{bridge}}\mathcal{L}_{\text{bridge}} + \lambda_{\text{iou}}\mathcal{L}_{\text{iou}}$, where $\mathcal{L}_{\text{MD}}$ includes the L1, gIoU, classification, and saliency losses, as in Moment DETR [26]. $\mathcal{L}_{\text{iou}}$ is the loss for the IoU head introduced in Line-406, designed to assist in ranking queries using $\mathcal{P}^{\text{IoU}}_i$ and $\mathcal{P}^{\text{IoU}}_j$ in Eq 3. The IoU head implementation follows [25, 58]. Ablations for hyperparameters λ_{bridge} and λ_{iou} are detailed in Appendix E.

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 [27]*NeurIPS'21*	52.89	33.02	54.82	29.40	30.73	53.94	34.84	-	-	32.20
UMT [40]*CVPR'22*	56.23	41.18	53.83	37.01	36.12	60.26	44.26	-	-	38.59
QD-DETR [47]*CVPR'23*	62.40	44.98	62.52	39.88	39.86	62.68	46.66	62.23	41.82	41.22
UniVTG [32]*ICCV'23*	58.86	40.86	57.60	35.59	35.47	59.74	-	-	-	36.13
EaTR [22]*ICCV'23*	-	-	-	-	-	61.36	45.79	61.86	41.91	41.74
MomentDiff [29]*NeurIPS'23*	57.42	39.66	54.02	35.73	35.95	-	-	-	-	-
TR-DETR [57]*AAAI'24*	64.66	48.96	63.98	43.73	42.62	67.10	51.48	66.27	46.42	45.09
UVCOM [67]*CVPR'24*	63.55	47.47	63.37	42.67	43.18	65.10	51.81	-	-	45.79
TaskWeave [71]*CVPR'24*	-	-	-	-	-	64.26	50.06	65.39	46.47	45.38
BAM-DETR [25]*ECCV'24*	62.71	48.64	64.57	46.33	45.36	65.10	51.61	65.41	48.56	47.61
CG-DETR [46]*Arxiv'24*	65.43	48.38	64.51	42.77	42.86	67.35	52.06	65.57	45.73	44.93
SpikeMba [31]*Arxiv'24*	64.13	49.42	-	43.67	43.79	65.32	51.33	-	44.96	44.84
Sim-DETR(Ours)	**67.64**	**50.91**	**67.81**	**47.59**	**46.93**	**69.48**	**54.06**	**69.70**	**51.11**	**49.50**

Table 1. Experimental results on the QVHighlights benchmark, comparing ours with state-of-the-art methods.

Method	TACoS				Charades-STA			
	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU
2D-TAN [81]	40.01	27.99	12.92	27.22	58.76	46.02	27.50	41.25
M-DETR [27]	37.97	24.67	11.97	25.49	65.83	52.07	30.59	45.54
MomentDiff [29]	44.78	33.68	-	-	-	55.57	32.42	-
QD-DETR [47]	-	-	-	-	-	57.31	32.55	-
UniVTG [32]	51.44	34.97	17.35	33.60	70.81	58.01	35.65	50.10
CG-DETR [46]	52.23	39.61	22.23	36.48	70.43	58.44	36.34	50.13
UVCOM [67]	-	36.39	23.32	-	-	59.25	36.64	-
SpikeMba [31]	51.98	39.34	22.83	35.81	71.24	59.65	36.12	51.74
Sim-DETR(Ours)	**57.06**	**42.79**	**26.82**	**39.44**	**73.09**	**61.34**	**39.62**	**52.56**

Table 2. Comparison results on the TACoS and Charades-STA benchmarks.

5. Experiments

5.1. Datasets and Implementation Details

Datasets. Following [27, 31, 32, 47, 67], we conduct comprehensive evaluation experiments across three datasets: QVHighlights [27], Charades-STA [12], and TACoS [50]. The QVHighlights dataset contains 10,310 text expressions and 18,367 distinct events, with most expressions corresponding to multiple events. Charades-STA is derived from a collection of 9,848 indoor scene videos, encompassing 16,128 events. TACoS includes only 273 videos, with an average length approaching five minutes per video. For a fair comparison, we adopt the same evaluation metrics as in previous methods. On the QVHighlights benchmark, we use Recall and Mean Average Precision (mAP) as primary metrics, while for Charades-STA and TACoS, we employ Recall and Mean Intersection over Union (mIoU).

Implementation Details. We set the input video resolution to 224×224 and represent each video by concatenating the CLIP [49] [CLS] tokens with SlowFast [10] features. For the Charades-STA and TACoS datasets, we additionally constructed two experiments using I3D [3] and VGG [56] as visual encoders, with Glove [48] embeddings for text features. The model is trained for 200 epochs on a single NVIDIA A40 GPU, using the AdamW [42] optimizer with an initial learning rate of 1e-4. Note that, for a fair comparison, the total number of layers in our Sim-DETR is set to 6, the same as recent works [22, 46, 67].

Backbone	Method	R@0.5	R@0.7
VGG [56]	2D-TAN [81]	40.94	22.85
	FVMR [11]	42.36	24.14
	SSRN [83]	46.72	27.98
	UMT [40]	48.31	29.25
	MomentDiff [29]	51.94	28.25
	QD-DETR [47]	52.77	31.13
	TR-DETR [57]	53.47	30.81
	CG-DETR [46]	55.22	34.19
	Sim-DETR(Ours)	**55.97**	**35.38**
I3D [3]	MAN [77]	46.53	22.72
	VSLNet [78]	47.31	30.19
	QD-DETR [47]	50.67	31.02
	TR-DETR [57]	55.51	33.66
	TaskWeave [71]	53.36	31.40
	Sim-DETR(Ours)	**57.55**	**35.73**

Table 3. Experimental results with VGG or I3D as backbone.

5.2. Comparison with State-of-the-Art Methods

As shown in Tables 1-3, we report detailed evaluation results across the QVHighlights, Charades-STA, and TACoS benchmarks. Our Sim-DETR consistently outperforms all state-of-the-art methods (SOTAs) across all metrics.

Results on QVHighlights are shown in Table 1. Compared to the latest published work, BAM-DETR [25], our Sim-

Method	R1		mAP		
	@0.5	@0.7	@0.5	@0.75	Avg.
Baseline	65.48	50.84	65.34	45.82	44.97
+ $\mathcal{L}_{\text{iou}}$	66.58	51.94	65.68	46.30	45.22
+ QGR	68.77	52.26	67.72	48.28	47.03
+ GLB	67.16	52.77	68.58	48.88	48.17
+ QGR,GLB	**69.48**	**54.06**	**69.70**	**51.11**	**49.50**

Table 4. Ablation study on components.

	Inner Relevance	Outer Global	Outer Local	mAP
(1)	span border dist	confidence pred	IoU pred	**49.50**
(2)	center dist	confidence pred	IoU pred	48.59
(3)	span IoU			48.94
(4)	span border dist	w/o	IoU pred	48.93
(5)	span border dist	confidence pred	w/o	48.66
(6)			center pred	48.97

Table 5. Impact of different measures.

DETR outperforms it by an average of +2.88% across all metrics. More importantly, it achieves remarkable gains of +4.93% and +2.27% in R1@0.5 and R1@0.7, respectively. In comparison TR-DETR [57], which uses standard DETR [2] as its basic framework and is used as our baseline model, our Sim-DETR achieves an average improvement of +4.31% and +4.41% in mAP for test and validation, respectively. These improvements underscore the effectiveness of our Sim-DETR in fundamentally addressing conflicts within and between queries. We further conduct a comparison with SpikeMba [31], a contemporary method built on the Mamba [17] framework. Without any complex module designs, the consistent improvements across all metrics highlight the simplicity and effectiveness of our Sim-DETR. Sim-DETR establishes an efficient baseline for the task, significantly advancing research in TSG.

Results on Charades-STA and TACoS are presented in Table 2, where we consistently outperform all methods on both datasets. Compared to the state-of-the-art methods, we achieve a +1.97% improvement on the Charades-STA dataset (with SpikeMba as the SOTA) and a +3.89% improvement on the TACoS dataset (with CG-DETR [46] as the SOTA). These enhancements further demonstrate the outstanding capability of our method in temporal localization. More importantly, they confirm that our Sim-DETR exhibits significant performance advantages across various scenarios, types, and lengths of videos, highlighting its robustness and generalizability.

Results with Different Backbones. Following previous works, we conduct two additional experiments on the QVHighlights dataset. These experiments use VGG and I3D for visual feature extraction and use Glove [48]embedding as text features. Table 3 presents the detailed results of these two experiments, where we achieve state-of-the-art results in both cases. Notably, for the video feature extractor I3D, we surpass the latest published method by +4.26%. These results further demonstrate the excellent generalization capabilities of our Sim-DETR across different backbones.

5.3. Ablation Study

Roadmap for building a simple yet effective TSG baseline is presented in Table 4. We employ TR-DETR without MR2HD module as our baseline, which achieves an mAP of 44.97%. Since our Query Grouping and Ranking (QGR) utilizes $\mathcal{L}_{\text{iou}}$ proposed by [25, 58], we additionally conduct an ablation study to isolate the impact of $\mathcal{L}_{\text{iou}}$. (1) QGR: To mitigate conflicts and optimization challenges arising from overly similar semantics between queries, we introduce the QGR, which improves the baseline by +2.07% on mAP, representing a subtle yet significant enhancement. (2) Global-Local Bridging (GLB): To address conflicts and ambiguities between global semantics and local localization within queries, we introduce GLB to enable seamless transformation through query-to-frame alignment. Experimental results show a performance boost to 48.17% on mAP, indicating the effectiveness of GLB in bridging global-to-local transformations. However, without QGR to address semantic conflicts in temporal DETR, the recall predictably decreased by -3.1%. (3) Overall Framework: With these adjustments, our Sim-DETR achieves a remarkable SOTA performance: 49.50% on mAP and 69.48% on R1@0.5.

The Effects of Different Conflict Metrics in QGR. Table 5 summarizes five experiments on conflict metrics in QGR. (1) The first row displays our final implementation, achieving 49.50% mAP. (2) & (3) We evaluate L2 center distance and span IoU for inter-query relationships. Center distance alone yields 48.59% mAP, highlighting its inadequacy in segment representation. Span IoU outperforms center distance but lags by 0.56% compared to boundary distance, likely due to IoU's broader tolerance, which reduces precision in query relevance. (4) To assess outer global semantic alignment, we remove global semantics, resulting in a -0.57% drop, emphasizing their role in query interaction and quality. (5) Finally, we test local location metrics in query prioritization. Excluding the outer local metric significantly lowers performance (-0.84% mAP), indicating that without local information, queries encounter conflicts due to overly generalized global semantics, leading to random matches. Consistent with (2), center localization accuracy shows similar limitations, with a 0.53% decrease.

6. Conclusion

In this paper, we propose a concise and efficient temporal sentence grounding framework Sim-DETR, introducing two simple modifications to the decoder for resolving conflicts both between and within queries. Experimental results demonstrate that our approach significantly outperforms state-of-the-art methods across all benchmarks.

Acknowledgment: This work is supported by the National Natural Science Foundation of China (No.62206174).

References

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

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1, 2, 3, 4, 8

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 7

[4] Chaoqi Chen, Yushuang Wu, Qiyuan Dai, Hong-Yu Zhou, Mutian Xu, Sibei Yang, Xiaoguang Han, and Yizhou Yu. A survey on graph neural networks and graph transformers in computer vision: A task-oriented perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

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

[6] Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, and Xiaolin Li. Rethinking the bottom-up framework for query-based video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10551–10558, 2020. 2

[7] Shaoxiang Chen and Yu-Gang Jiang. Semantic proposal for activity localization in videos via sentence query. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8199–8206, 2019. 2

[8] Qiyuan Dai and Sibei Yang. Curriculum point prompting for weakly-supervised referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13711–13722, 2024. 2

[9] Xiyang Dai, Yinpeng Chen, Jianwei Yang, Pengchuan Zhang, Lu Yuan, and Lei Zhang. Dynamic detr: End-to-end object detection with dynamic attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2988–2997, 2021. 3

[10] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 5, 7

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

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

[13] Peng Gao, Minghang Zheng, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fast convergence of detr with spatially modulated co-attention. *arXiv preprint arXiv:2101.07448*, 2021. 3

[14] Yulu Gao, Yifan Sun, Xudong Ding, Chuyang Zhao, and Si Liu. Ease-detr: Easing the competition among object queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17282–17291, 2024. 3, 6

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

[16] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann. Excl: Extractive clip localization using natural language descriptions. *arXiv preprint arXiv:1904.02755*, 2019. 2

[17] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 8

[18] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Advances in neural information processing systems*, 33:5679–5690, 2020. 1

[19] De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7366–7375, 2018.

[20] Hanzhuo Huang, Yufan Feng, Cheng Shi, Lan Xu, Jingyi Yu, and Sibei Yang. Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. *Advances in Neural Information Processing Systems*, 36:26135–26158, 2023. 1

[21] Hanzhuo Huang, Yuan Liu, Ge Zheng, Jiepeng Wang, Zhiyang Dou, and Sibei Yang. Mvtokenflow: High-quality 4d content generation using multiview token flow. *arXiv preprint arXiv:2502.11697*, 2025. 2

[22] Jinhyun Jang, Jungin Park, Jin Kim, Hyeongjun Kwon, and Kwanghoon Sohn. Knowing where to focus: Event-aware transformer for video grounding. In *ICCV*, pages 13846–13856, 2023. 1, 2, 4, 5, 7

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

[24] Manjin Kim, Heeseung Kwon, Chunyu Wang, Suha Kwak, and Minsu Cho. Relational self-attention: What's missing in attention for video understanding. *Advances in Neural Information Processing Systems*, 34:8046–8059, 2021. 1

[25] Pilhyeon Lee and Hyeran Byun. Bam-detr: Boundary-aligned moment detection transformer for temporal sentence grounding in videos. In *European Conference on Computer Vision*, pages 220–238. Springer, 2025. 1, 2, 4, 5, 6, 7, 8

[26] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858, 2021. 1, 6

[27] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In *Neurips*, pages 11846–11858, 2021. 1, 2, 3, 4, 7

[28] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13619–13627, 2022. 3

[29] Pandeng Li, Chen-Wei Xie, Hongtao Xie, Liming Zhao, Lei Zhang, Yun Zheng, Deli Zhao, and Yongdong Zhang. Momentdiff: Generative video moment retrieval from random to real. In *Neurips*, 2023. 7

[30] Pandeng Li, Chen-Wei Xie, Liming Zhao, Hongtao Xie, Jiannan Ge, Yun Zheng, Deli Zhao, and Yongdong Zhang. Progressive spatio-temporal prototype matching for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4100–4110, 2023. 1

[31] Wenrui Li, Xiaopeng Hong, Ruiqin Xiong, and Xiaopeng Fan. Spikemba: Multi-modal spiking saliency mamba for temporal video grounding. *arXiv preprint arXiv:2404.01174*, 2024. 7, 8

[32] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In *ICCV*, pages 2794–2804, 2023. 1, 2, 7

[33] Liang Lin, Pengxiang Yan, Xiaoqian Xu, Sibei Yang, Kun Zeng, and Guanbin Li. Structured attention network for referring image segmentation. *IEEE Transactions on Multimedia*, 24:1922–1932, 2021. 2

[34] Daizong Liu and Wei Hu. Skimming, locating, then perusing: A human-like framework for natural language video localization. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4536–4545, 2022. 2

[35] Daizong Liu, Xiaoye Qu, Jianfeng Dong, and Pan Zhou. Adaptive proposal generation network for temporal sentence localization in videos. *arXiv preprint arXiv:2109.06398*, 2021. 2

[36] Daizong Liu, Xiaoye Qu, Xing Di, Yu Cheng, Zichuan Xu, and Pan Zhou. Memory-guided semantic learning network for temporal sentence grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1665–1673, 2022. 1

[37] Daizong Liu, Xiaoye Qu, and Wei Hu. Reducing the vision and language bias for temporal sentence grounding. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4092–4101, 2022. 1

[38] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. Cross-modal moment localization in videos. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 843–851, 2018. 2

[39] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022. 3

[40] Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *CVPR*, pages 3042–3051, 2022. 1, 7

[41] Ye Liu, Jixuan He, Wanhua Li, Junsik Kim, Donglai Wei, Hanspeter Pfister, and Chang Wen Chen. r^2-tuning: Efficient image-to-video transfer learning for video temporal grounding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2

[42] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[43] Chujie Lu, Long Chen, Chilie Tan, Xiaolin Li, and Jun Xiao. 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)*, pages 5144–5153, 2019. 2

[44] Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. Generation-augmented retrieval for open-domain question answering. *arXiv preprint arXiv:2009.08553*, 2020. 1

[45] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3651–3660, 2021. 3

[46] WonJun Moon, Sangeek Hyun, SuBeen Lee, and Jae-Pil Heo. Correlation-guided query-dependency calibration in video representation learning for temporal grounding. *arXiv preprint arXiv:2311.08835*, 2023. 1, 2, 3, 4, 5, 7, 8

[47] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *CVPR*, pages 23023–23033, 2023. 1, 2, 4, 5, 7

[48] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 7, 8

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5, 7

[50] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013. 7

[51] Byungseok Roh, JaeWoong Shin, Wuhyun Shin, and Saehoon Kim. Sparse detr: Efficient end-to-end object detection with learnable sparsity. *arXiv preprint arXiv:2111.14330*, 2021. 3

[52] Cheng Shi and Sibei Yang. Edadet: Open-vocabulary object detection using early dense alignment. In *Proceedings of*

the IEEE/CVF international conference on computer vision, pages 15724–15734, 2023. 3

[53] Cheng Shi and Sibei Yang. Logoprompt: Synthetic text images can be good visual prompts for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2932–2941, 2023. 2

[54] Cheng Shi, Yulin Zhang, Bin Yang, Jiajin Tang, Yuexin Ma, and Sibei Yang. Part2object: Hierarchical unsupervised 3d instance segmentation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 3

[55] Cheng Shi, Yuchen Zhu, and Sibei Yang. Plain-det: A plain multi-dataset object detector. In *European Conference on Computer Vision*, pages 210–226. Springer, 2024. 2, 3

[56] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 7

[57] Hao Sun, Mingyao Zhou, Wenjing Chen, and Wei Xie. Tr-detr: Task-reciprocal transformer for joint moment retrieval and highlight detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4998–5007, 2024. 1, 2, 4, 5, 7, 8

[58] Xiaolong Sun, Liushuai Shi, Le Wang, Sanping Zhou, Kun Xia, Yabing Wang, and Gang Hua. Diversifying query: Region-guided transformer for temporal sentence grounding. *arXiv preprint arXiv:2406.00143*, 2024. 2, 6, 8

[59] Zhiqing Sun, Shengcao Cao, Yiming Yang, and Kris M Kitani. Rethinking transformer-based set prediction for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3611–3620, 2021. 3

[60] Jiajin Tang, Ge Zheng, Cheng Shi, and Sibei Yang. Contrastive grouping with transformer for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23570–23580, 2023. 3

[61] Jiajin Tang, Ge Zheng, and Sibei Yang. Temporal collection and distribution for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15466–15476, 2023. 1, 3

[62] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 5

[63] Tao Wang, Li Yuan, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. Pnp-detr: Towards efficient visual analysis with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4661–4670, 2021. 3

[64] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based detector. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2567–2575, 2022. 3

[65] Sangmin Woo, Jinyoung Park, Inyong Koo, Sumin Lee, Minki Jeong, and Changick Kim. Explore-and-match: Bridging proposal-based and proposal-free with transformer for sentence grounding in videos. *arXiv preprint arXiv:2201.10168*, 2022. 2

[66] Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. Boundary proposal network for two-stage natural language video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2986–2994, 2021. 2

[67] Yicheng Xiao, Zhuoyan Luo, Yong Liu, Yue Ma, Hengwei Bian, Yatai Ji, Yujiu Yang, and Xiu Li. 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*, pages 18709–18719, 2024. 2, 4, 7

[68] Chen-Wei Xie, Siyang Sun, Xiong Xiong, Yun Zheng, Deli Zhao, and Jingren Zhou. Ra-clip: Retrieval augmented contrastive language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19265–19274, 2023. 1

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

[70] Vikas Yadav, Steven Bethard, and Mihai Surdeanu. Unsupervised alignment-based iterative evidence retrieval for multi-hop question answering. *arXiv preprint arXiv:2005.01218*, 2020. 1

[71] Jin Yang, Ping Wei, Huan Li, and Ziyang Ren. Task-driven exploration: Decoupling and inter-task feedback for joint moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18308–18318, 2024. 1, 2, 3, 4, 5, 7

[72] Shuo Yang and Xinxiao Wu. Entity-aware and motion-aware transformers for language-driven action localization in videos. *arXiv preprint arXiv:2205.05854*, 2022. 1

[73] Sibei Yang, Meng Xia, Guanbin Li, Hong-Yu Zhou, and Yizhou Yu. Bottom-up shift and reasoning for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11266–11275, 2021. 2

[74] Zhuyu Yao, Jiangbo Ai, Boxun Li, and Chi Zhang. Efficient detr: improving end-to-end object detector with dense prior. *arXiv preprint arXiv:2104.01318*, 2021. 3

[75] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[76] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10287–10296, 2020. 2

[77] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1247–1257, 2019. 2, 7

[78] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. *arXiv preprint arXiv:2004.13931*, 2020. 2, 7

[79] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 3

[80] Songyang Zhang, Jinsong Su, and Jiebo Luo. Exploiting temporal relationships in video moment localization with natural language. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1230–1238, 2019. 2

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

[82] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191, 2023. 2

[83] Jiahao Zhu, Daizong Liu, Pan Zhou, Xing Di, Yu Cheng, Song Yang, Wenzheng Xu, Zichuan Xu, Yao Wan, Lichao Sun, et al. Rethinking the video sampling and reasoning strategies for temporal sentence grounding. *arXiv preprint arXiv:2301.00514*, 2023. 7

[84] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3

[85] Yuchen Zhu, Cheng Shi, Dingyou Wang, Jiajin Tang, Zhengxuan Wei, Yu Wu, Guanbin Li, and Sibei Yang. Rethinking query-based transformer for continual image segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4595–4606, 2025. 3