RGNet: A Unified Clip Retrieval and Grounding Network for Long Videos

Tanveer Hannan[1,2⋆] Md Mohaiminul Islam[3] Thomas Seidl[1,2] Gedas Bertasius[3]

[1] LMU Munich
[2] MCML
[3] UNC Chapel Hill

Abstract. Locating specific moments within long videos (20–120 minutes) presents a significant challenge, akin to finding a needle in a haystack. Adapting existing short video (5–30 seconds) grounding methods to this problem yields poor performance. Since most real-life videos, such as those on YouTube and AR/VR, are lengthy, addressing this issue is crucial. Existing methods typically operate in two stages: clip retrieval and grounding. However, this disjoint process limits the retrieval module's fine-grained event understanding, crucial for specific moment detection. We propose RGNet which deeply integrates clip retrieval and grounding into a single network capable of processing long videos into multiple granular levels, e.g., clips and frames. Its core component is a novel transformer encoder, RG-Encoder, that unifies the two stages through shared features and mutual optimization. The encoder incorporates a sparse attention mechanism and an attention loss to model both granularity jointly. Moreover, we introduce a contrastive clip sampling technique to mimic the long video paradigm closely during training. RGNet surpasses prior methods, showcasing state-of-the-art performance on long video temporal grounding (LVTG) datasets MAD and Ego4D. The code is released at https://github.com/Tanveer81/RGNet.

Keywords: Long Video Temporal Grounding · Moment Localization

1 Introduction

The exponential rise in online videos has created a demand for effective content retrieval systems. The retrieval results of user-defined queries, particularly in long videos (20-120 minutes), often require locating specific moments. Most existing solutions [5, 41, 53, 55], tailored for short videos (5-30 seconds), struggle when applied to hour-long videos, slowing down the retrieval and impacting the quality of the results [40, 54]. This challenge emerges because it is very difficult to locate short moments in a long video based on text queries, a task known as Long Video Temporal Grounding (LVTG) [14, 40].

[⋆] Corresponding author: hannan@dbs.ifi.lmu.de

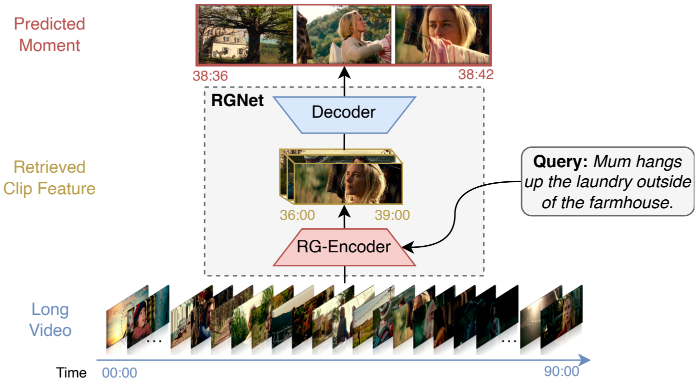

Fig. 1: Overview of RGNet. It predicts the moment boundary specified by textual queries from an hour-long video. First, our proposed **RG-Encoder** maps the video and text features to a joint space and retrieves the relevant clip feature. The subsequent grounding decoder processes the retrieved features to predict the beginning and end times of the moment. The encoder parallelly operates at multiple levels of granularity (e.g., clip and frame) to achieve an end-to-end-solution.

A straightforward solution [16, 37] for the LVTG task is to divide the video into shorter clips, retrieve the most relevant one, and apply a grounding network to predict the moment. However, the grounding network cannot rectify failure in the clip retrieval phase.

Empirically, we evaluate the impact of the two stages in Sec. 5.3 and identify clip retrieval as the primary factor for poor performance.

The existing methods have a separate module for clip retrieval, which is disjoint from their grounding network (see Fig. 2). They typically follow a text-video retrieval [13, 23, 29] technique to select the relevant clip. However, video retrieval only requires a high-level understanding of video topics. For example, a typical video retrieval query is "A movie about a family surviving in a farmhouse.". In contrast, an example clip retrieval query in Fig. 1 is "Find the moment when the mum hangs up laundry outside the farmhouse". These specific moments from a long video require fine-grained event understanding. Thus, video retrieval models are suboptimal for these fine-grained moment localization tasks.

We attribute this disjoint retrieval to the poor performance of these models. However, unifying clip selection and grounding is challenging due to their distinct setups. The former is a retrieval task, whereas the latter is a regression task. Moreover, it requires modeling two levels of video granularity: clip and frame. To address these technical challenges, we propose **RGNet**, a unified clip Retrieval and Grounding Network (see Fig. 1). The single network enables end-to-end training of both stages. This improves the retrieval module's fine-grained event understanding by directly optimizing it with the moment annotations. Parallelly, the grounding network also benefits from the strong multimodal features produced by our encoder, further enhancing the localization capability.

To achieve this unification, we propose a novel transformer encoder, RG-Encoder, which enables the grounding network to perform clip retrieval. This

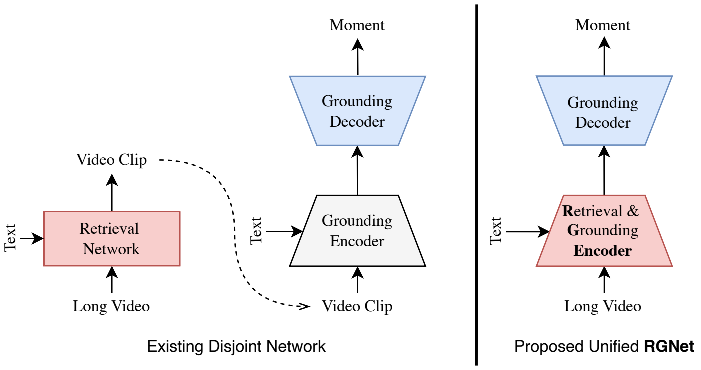

Fig. 2: Unified Solution. (left) Existing methods involve a separate retrieval and grounding network. The disjoint retrieval lacks fine-grained event understanding, which is crucial for moment localization. **(right)** Our unified network architecture overcomes it by deeply integrating the retrieval module with the grounding objective.

eliminates the suboptimal video retrieval network and effectively models long video clip retrieval. To enable the retrieval module to understand fine-grained events, we design a sparse-attention mechanism in the encoder. The sparse attention enables the retrieval module to focus on specific events in the video, which is more synergistic with the grounding task. Thus RG-Encoder is capable of operating both at clip and frame level granularity. We propose an Intra-Clip Attention Loss, which motivates the sparse attention to focus more on the video frames aligned with the specified event. Furthermore, we propose a negative clip mining technique to simulate clip retrieval from long videos during training. This negative mining enables our network to train on a large batch size with an inter-clip contrastive loss. The large number of negative clips in the batch closely mimics the long video paradigm and reduces the gap between the training and test phases.

Together, these components of RGNet bridge the gap between the two stages of LVTG, enhancing the fine-grained event understanding in hour-long videos. In summary, our contributions are fourfold:

- We systematically deconstruct existing LVTG methods into clip retrieval and grounding stages. Through empirical evaluations, we discern that disjoint retrieval is the primary factor contributing to poor performance.
- Based on our observations, we introduce RGNet, which integrates clip retrieval with grounding through parallel clip and frame-level modeling. This obviates the necessity for a separate video retrieval network, replaced instead by an end-to-end clip retrieval module tailored specifically for long videos.
- We introduce sparse attention to the retriever and a corresponding loss to model fine-grained event understanding in long-range video. We propose a contrastive negative clip-mining strategy to simulate clip retrieval from a long video during training.
- RGNet achieves state-of-the-art performance across both LVTG datasets, Ego4D [14] and MAD [40]. For instance, RGNet outperforms the previous best Ego4D method [39] by a substantial margin (**9.7%**).

2 Related Literature

Short Video Temporal Grounding. The recent grounding methods mainly focus on short videos. The best-performing ones utilize transformer variants [3, 27, 34, 56]. For example, Moment-DETR [22] adopted transformers for combined video and text features. Subsequent works, such as UMT [28] and QD-DETR [35], split the cross-modal and unimodal modeling of video and text features for enhanced performance. EATR [18] improves its decoder with video-specific moment queries. These models were initially designed for shorter videos. However, when applied to hour-long videos, the moments become extremely tiny, posing a challenge akin to finding a needle in a haystack. Hence, these methods perform poorly on long videos, as noted by the authors of the MAD dataset [40]. In contrast, we effectively address the task by simultaneously processing long videos at two granular levels (video clips and frames) within a single network.

Long Video Temporal Grounding. Recently, video temporal grounding has been adapted for long videos with MAD [40] and Ego4D [14] datasets. Typically, the methods designed for long videos involve two stages. Proposal-free methods [2, 26, 41, 53, 55] segment lengthy videos into smaller parts to predict candidate moments and then rank them to obtain the final predictions. Proposal-based methods [16, 37] generate proposal clips or anchors and subsequently apply their grounding model to the retrieved proposals. Some methods, like CONE [16], and M-Guidance [2], utilize the detection transformer [3] as the grounding network to improve the grounding phase. These proposal-based models are more efficient and perform better than their counterparts. However, they need a separate retrieval module to select the proposal clips. This disjoint two-stage architecture is what we find to be the main reason for the poor performance of these models. In contrast, we propose a novel transformer encoder that solves both clip retrieval and grounding in a unified way, enabling end-to-end optimization.

Video-text Retrieval. The early retrieval methods [4, 7, 21, 49, 50] designed multimodal fusion techniques to align pre-trained video and text models. To bridge the gap between pre-training and downstream tasks, large-scale pre-trained video-language representations [1, 6, 9, 11, 31, 42, 45–47] have been introduced in various works. Several studies [8, 10, 19, 20, 32, 43, 44, 48, 51] have transferred the knowledge from CLIP [38] to text-video retrieval tasks. In recent studies, various sampling strategies [13, 23, 29] have been explored, enabling the model to selectively concentrate on pertinent video frames based on provided text inputs. For instance, Clip-BERT [23] introduces a sparse sampling strategy, X-Pool [13] utilizes a cross-modal attention model, and TS2-Net [29] adapts CLIP to capture temporal information and eliminate unimportant tokens. While existing methods aim to retrieve high-level video topics from a pool of video collections, moment localization demands a fine grained understanding to identify specific events within a video. Despite the similar setup to video retrieval, retrieving the corresponding clip from a long video necessitates deeper event comprehension. Therefore, we tailored our encoder to enhance event understanding and simulated precise clip retrieval conditions during training.

3 RGNet

This section presents a detailed description of our proposed unified LVTG method. As illustrated in Fig. 1, RGNet takes a long video and a text query as input and predicts the precise moment semantically correlated with the query in an end-to-end manner. The whole network is divided into the proposed RG-Encoder (Fig. 3) and a decoder (Sec. 3.3). The encoder retrieves the most relevant clip feature from the input long video. The decoder then processes the retrieved clip feature to predict the precise moment boundary. Additionally, an Intra-Clip Attention Loss and an Inter-Clip Contrastive Loss are incorporated to facilitate low-level event understanding in long-range videos.

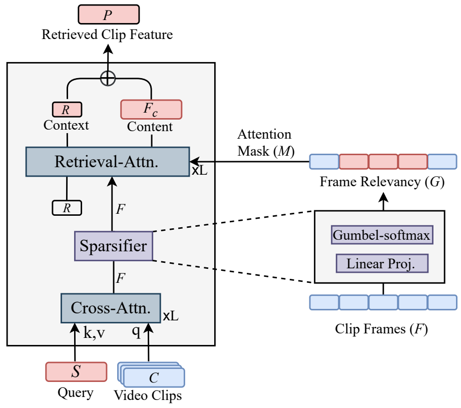

Fig. 3: Overview of RG-Encoder. It takes video clips and textual query as input and retrieves the relevant clip features. First, a cross-attention fuses the clips with text, and the sparsifier masks the out-of-moment frames. Based on the mask, the retrieval attention focuses on in-moment frames (colored red) and generates clip-level context and frame-level content features. We combine the context and content to generate the retrieved clip feature.

3.1 Feature Extraction

Let us assume we have an untrimmed long video V consisting of T frames and a query sentence S with N words corresponding to a target moment's center τ_c and width τ_w. Following previous methods [16, 37], we use pre-trained frozen models to extract visual features $V = \{f^1, f^2, ..., f^T\} \in R^{T \times D_f}$ as well as textual features $S = \{w^1, w^2, ..., w^N\} \in R^{N \times D_w}$, where D_f and D_w represent the feature dimensions of video frames and query words. We slice the long video V into clips of length L_c and feed them into our model. We employ a sliding window of length

L_c and stride $L_c/2$ to obtain a collection of T_c clips $C = \{C^1, C^2, ..., C^{T_c}\} \in R^{T_c \times L_c \times D_f}$. Here, the i^{th} clip, $C^i = \{f^{s_i+1}, f^{s_i+2}, ..., f^{s_i+L_c}\} \in R^{L_c \times D_f}$, where s_i is the start index of the clip.

3.2 RG-Encoder

All the clips C from a single video comprise samples in a batch. The RG-Encoder (Fig. 3) processes them with the query S to retrieve the relevant clip feature P. First, the cross-attention generates query-conditioned frame features F. Then, based on the frame-query correlation calculated from F, the sparsifier determines the relevant frames inside the clip to produce a mask M for retrieval attention. We incorporate a learnable token R to capture the context of the clip. Additionally, the frame features F undergo retrieval attention based on the predicted mask M to generate the content feature F_c. The combination of context R and content F_c forms the retrieved clip feature P.

Cross Attention: To evaluate the relevance of each frame within the clip C^i, we utilize cross-attention, where frame features act as queries and text features serve as keys and values. Cross-attention represents the frame features as a weighted average of the textual features, scaled by their mutual correlation. This process results in text-conditioned frame features F^i, establishing a fine-grained correspondence between the modalities. The query, key and values are calculated as $Q^i = l_Q(C^i) \in R^{L_c \times D_f}$, $K = l_K(S) \in R^{N \times D_q}$ and $V = l_V(S) \in R^{N \times D_q}$ respectively. Here, $l_Q(\cdot)$, $l_K(\cdot)$, and $l_V(\cdot)$ denote the projection layers for the query, key, and value. Then, we apply the cross-attention layer as follows:

$$F^i = \text{softmax}(Q^i K^T)V + Q^i \tag{1}$$

Sparsifier: To retrieve the clip, we need to assess the relevance of each clip based on the text query. Clip relevance is determined by aggregating the relevancy of its frames. However, given the small moment duration, most frames are unrelated to the query. Hence, we classify frames into relevant and non-relevant categories using the sparsifier to assist retrieval attention in focusing on these fine-grained events. We calculate the relevancy $G^j \in [0, 1]$ of the j^{th} frame with Eq. 2. We use a differentiable Gumbel-softmax function [15, 17] for enabling end-to-end training. The detailed calculation is in the supplementary materials.

$$G^j = \text{Gumbel-softmax}(\text{linear}(f^j)) \tag{2}$$

The intra-clip attention loss optimizes this classification using the ground truth frame relevancy information described in Sec. 3.4. To distinguish irrelevant frames from their counterparts, we constrain their updates in the retrieval attention. This restriction is implemented with an attention mask computed from the learned frame relevancy G^j. The attention mask M between j^{th} and k^{th} frame is calculated with Eq. 3. Here, the j^{th} and k^{th} frames are the attention query and key, respectively. This leads to a reduced set of relevant and informative frames for subsequent attention.

$$M(j,k) = \begin{cases} 0 & \text{if } G^j > 0.5 \text{ or } j = k \\ -\infty & \text{otherwise} \end{cases} \tag{3}$$

Retrieval Attention: We aim for a learnable clip retrieval function to train end-to-end with the grounding objective. Thus, we devise the retrieval attention for aggregating the text-conditioned frame features based on their relevancy. To condense the i^{th} clip into a contextual feature embedding, we introduce a randomly initialized learnable vector R^i. We denote this as retrieval token, which is trained with the inter-clip contrastive loss.

First, we concatenate the retrieval token with the text-conditioned frame features to create the queries $Q^i = [R^i, F^i]$ for retrieval attention. We also generate an attention mask containing all zeros for the retrieval token and concatenate it with M. We set the key $K = Q^i$, and the value $V = Q^i$ to calculate the retrieval attention according to Eq. 4. We get the clip-level context feature, $R^i = \tilde{Q}^{i,1}$ and frame-level content features, ${F_c}^{i,j} = [\tilde{Q}^{i,2}, \tilde{Q}^{i,3}, ..., \tilde{Q}^{i,L_c+1}]$. Here, ${F_c}^{i,j}$ is the j^{th} frame feature of the i^{th} clip.

$$\tilde{Q}^i = \text{softmax}(Q^i K^T + M)V + Q^i \tag{4}$$

Further, the context feature R is utilized to retrieve the clip. It undergoes a linear projection layer, $l_s(\cdot)$, to obtain context scores $S_r = \text{sigmoid}(l_s(R))$. We retrieve the most relevant clip based on the context score S_r. Importantly, the context R represents the queried moment by attending to the reduced set of relevant frames. So, we fuse it with the frame features with Eq. 5 to produce stronger multimodal inputs for the decoder. It enhances the decoder's localization capability significantly. The decoder only processes features from the retrieved clip to predict the moment.

$$P^{i,j} = {F_c}^{i,j} + R^i \times G^j, i \in P, \forall j \tag{5}$$

3.3 Grounding Decoder

We process the retrieved clip features with a transformer decoder to predict precise moment boundaries. Previous methods [16, 22] feed both modalities into the decoder, introducing multiple positional information (e.g., time in the video and word order in the text) that complicates the detection task. In contrast, our encoder injects text information directly into the clip, eliminating the requirement to feed both modalities into the decoder. We incorporate learnable anchor queries [27] to represent the moment's center and width as (τ_c, τ_w). The decoder has 2 cross-attention and 2 self-attention layers.

3.4 Loss Functions

Intra-Clip Attention Loss guides our sparsifier to distinguish frames within and outside of the ground truth moment. This enables the retrieval attention to focus on relevant video regions and achieve fine-grained event understanding.

Specifically, the in-moment frames are required to maintain higher relevancy scores than those outside the ground truth moment. The loss is defined as:

$$\mathcal{L}_{\text{attn}} = \max(0, \Delta + S_c(i, j_{\text{out}}) - S_c(i, j_{\text{in}})) \tag{6}$$

In the equation, S_c is the relevancy score, Δ represents the margin, and anchor i is an in-moment frame. We randomly chose in-moment and out-of-moment frames j_{in} and j_{out} from the same clip. To calculate the relevancy score, the context features R_i, and encoded clip $P^{i,j}$ undergoes linear layers , $l_r(\cdot)$ and $l_c(\cdot)$, to obtain projected context $R^i_{proj} = l_r(R^i) \in R^{D_l}$ and content features, $P^{i,j}_{proj} = l_c(P^{i,j}) \in R^{L_c \times D_l}$. Then, we calculate the frame-level relevancy score:

$$S_c(i, j) = R^i_{proj} \cdot P^{i,j}_{proj} \tag{7}$$

Inter-Clip Contrastive loss. A long video contains many clips with similar environments and scenes. However, most of these clips do not contain the specific moment we are searching for. So, high-level video understanding is not enough to distinguish precise moments. During the test phase, we need to handle many such negative clips. To tackle this issue, unlike previous approaches [16], we sample a large number of negative clips and train our retriever on a large batch size. Hence, this negative clip sampling reduces the disparity between the train and test phases.

For this section, we denote the context feature, $R^{i,j}$ for i^{th} proposal and j^{th} text query. We train our model contrastively with positive text-clip pairs $[R^{i,j}]_{i=j}$ and negative pairs $[R^{i,j}]_{i \neq j}$. We use the InfoNCE loss [36] to identify the positive text-clip pair amongst a set of unrelated negative samples. Here, $l_{cont}(\cdot)$ is a linear projection layer that transforms the D_f dimensional context feature into a one-dimensional logit.

$$\mathcal{L}_{\text{cont}} = -\sum_i log \frac{\exp(l_{\text{cont}}(R^{i,i}))}{\sum_j \exp(l_{\text{cont}}(R^{i,j}))} \tag{8}$$

Grounding loss. The objective functions for grounding, which aim to locate desired moments, are adopted from the baseline approach [22]. The grounding loss $\mathcal{L}_{\text{g}}$ measures the discrepancy between the ground truth (GT) and predicted moments. A one-to-one correspondence between GT and predicted moments is established using the Hungarian algorithm. This loss includes both an $\mathcal{L}1$ loss and a generalized IoU loss ($\mathcal{L}$gIoU). Additionally, a cross-entropy loss $\mathcal{L}_{\text{CE}}$ is employed to classify predicted moments as either foreground or background. Thus, $\mathcal{L}_{\text{g}}$ is defined as follows:

$$\mathcal{L}_{\text{g}} = \lambda_{\mathcal{L}1}||\tau - \hat{\tau}|| + \lambda_{\text{gIoU}}\mathcal{L}_{\text{gIoU}}(\tau, \hat{\tau}) + \lambda_{\text{CE}}\mathcal{L}_{\text{CE}}, \tag{9}$$

where τ and $\hat{\tau}$ represent the ground-truth moment and its corresponding prediction, containing center coordinates τ_c and width τ_w. The hyper-parameters λ_* are used for balancing the losses. Finally, with the attention and contrastive loss, the total loss $\mathcal{L}_{\text{total}}$ is defined as follows:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{attn}}\mathcal{L}_{\text{attn}} + \lambda_{\text{cont}}\mathcal{L}_{\text{cont}} + \mathcal{L}_{\text{g}} \tag{10}$$

4 Experimental Setup

4.1 Datasets and Evaluation Metric

MAD [40] is an extensive dataset comprising 1.2K hours of full-length movies and 384K natural language queries, each associated with specific moments in the videos. The videos are orders of magnitude longer than previous datasets, with an average duration of 110 minutes, while the specified text moments are on average 4.1 seconds. This small moment-to-video ratio poses a significant challenge for the grounding task.

Ego4D-NLQ [14] is a large-scale egocentric video data set with multiple challenges. Specifically, we use the episodic memory benchmark Ego4D-NLQ, which requires localizing where the answer to a natural language query can be seen. It contains around 13 template questions, amounting to around 74K queries. The train, val, and test sets contain 11.3K, 3.9K, and 4.0K queries. The video length ranges from 8 to 20 minutes, with an average of 8.25 minutes, and the average moment duration is 8.3 seconds. This means the moments constitute only 2% of the input video on average.

Grounding Metric: For grounding, we follow previous methods [16, 40], and adopt the standard metric Recall@k at IoU=θ (Rk_θ). This metric represents the percentage of testing samples with at least one grounding prediction whose intersection over union (IoU) with the ground truth (GT) is larger than θ among the top-k predictions.

Retrieval Metric: To assess our proposal retrieval stage independently of the grounding, we utilize the standard retrieval metric [13], Recall at Rank k (R@k). This metric calculates the percentage of GT moments present in the top-k retrieved proposals.

4.2 Implementation Details:

We use pre-trained CLIP [38] and EgoVLP [24] models to extract video frames from MAD and Ego4D. Text features for both datasets are extracted using CLIP. The feature extractors are frozen, and their pre-trained weights remain unchanged during training. For Ego4D, we pre-train our model with NaQ annotations from previous work [39]. The proposal length, W_c is set to 180s for MAD and 48s for Ego4D, with only the top-30 and top-5 proposals retrieved, respectively. Training on Ego4D and MAD is conducted on four Nvidia-RTX-A6000 GPUs while fine-tuning on Ego4D is performed on a single GPU. Considering the longer proposals in MAD, we utilize a batch size of 32 for Ego4D and 8 for MAD. The number of moment queries is set to 5. For loss computation, we set hyperparameters as $\lambda_{L1} = 10$, $\lambda_{\text{gIoU}} = 1$, $\lambda_{\text{CE}} = 4$, $\lambda_{\text{samp}} = 1$, $\lambda_{\text{cont}} = 10$, and $\mathit{\Delta} = 0.2$. We use Xavier initialization [12] and employ AdamW [30] with an initial learning rate of 1×10^{-4}. RGNet is trained for 35 and 200 epochs on

MAD and Ego4D datasets. The initial learning rate is reduced by one order of magnitude at epochs 25 for MAD and 120 for Ego4D.

5 Results and Analysis

We first compare our performance with previous state-of-the-art models. Then, we report detailed ablation studies of our proposed method and visualize the qualitative results compared to the disjoint baseline.

Model	`NaQ`	**Ego4D-NLQ** [14]					**MAD** [40]							**Avg**
		$R1_{.3}$	$R5_{.3}$	$R1_{.5}$	$R5_{.5}$	Avg	$R1_{.1}$	$R5_{.1}$	$R1_{.3}$	$R5_{.3}$	$R1_{.5}$	$R5_{.5}$	Avg	
2D-TAN [55]	✗	5.04	12.89	2.02	5.88	6.46	3.22	11.90	2.52	9.25	1.58	5.69	5.69	6.07
UniVTG [25]	✗	11.74	7.54	3.25	7.88	7.60	-	-	-	-	-	-	-	-
VSLNet [53]	✓	10.26	19.01	5.81	12.67	11.93	-	-	-	-	-	-	-	-
VLG-Net [41]	✗	-	-	-	-	-	3.64	11.66	2.76	9.31	1.65	5.99	5.84	-
M-DETR [22]	✗	8.23	23.23	5.01	13.37	12.46	3.60	12.98	2.81	9.86	1.67	5.58	6.08	9.27
M-Guide [2]	✗	-	-	-	-	-	9.30	18.96	4.65	13.06	2.16	7.40	9.26	-
SOONet [37]	✗	8.00	22.40	3.76	11.09	11.31	11.26	23.21	9.00	**19.64**	5.32	**13.14**	13.59	12.45
H-Hands [52]	✗	13.20	23.30	7.90	15.60	15.00	-	-	-	-	-	-	-	-
CONE [16]	✗	14.15	30.33	8.18	18.02	17.67	8.90	20.51	6.87	16.11	4.10	9.59	11.01	14.34
EgoVLP	✓	15.90	26.38	9.46	17.80	17.38	-	-	-	-	-	-	-	-
ReLeR [39]	✓	19.31	23.62	11.59	15.75	17.57	-	-	-	-	-	-	-	-
Ours `(Default)`	✗	18.28	34.02	12.04	22.89	21.81	**12.43**	**25.12**	**9.48**	18.72	**5.61**	10.86	**13.70**	17.80
Ours	✓	**20.63**	**41.67**	**12.47**	**25.08**	**24.96**	**12.43**	**25.12**	**9.48**	18.72	**5.61**	10.86	**13.70**	**19.33**

Table 1: Main results on Ego4D-NLQ and MAD. RGNet achieves state-of-the-art performance on both datasets. Our default network is trained without NaQ annotations [39] on Ego4D. Even without NaQ annotations, our default network shows a larger improvement, underscoring the effectiveness of our solution with limited data.

5.1 Result on the Ego4D-NLQ Dataset

RGNet demonstrates state-of-the-art performance on all metrics for the Ego4D-NLQ benchmark. As shown in Table 1, RGNet improves its primary evaluation metric, the $R1_{.3}$ and $R5_{.3}$ score average by 9.7%. Separately, we outperform the previous best methods on $R1_{.3}$ and $R5_{.3}$ by **1.32%** and **18.1%**, respectively. This improvement is consistent across other metrics as well. The unified clip and frame-level modeling in RGNet parallelly capture long- and short-range temporal dependencies, contributing to its superior performance. By modeling the entire video as a whole, RGNet learns better cross-modal alignment by contrasting visual scenarios from various events within the same video. Notably, compared to SooNet [37] and CONE [16], two prominent proposal-based methods, RGNet achieves an average score improvement of 13.7% and 7.3%. This substantial improvement is attributed to our unified modeling of clip retrieval, whereas they separately employ a retrieval similarity search between the video and the text.

5.2 Result on the MAD Dataset

We compare our model's performance on the MAD dataset in Tab. 1. RGNet achieves state-of-the-art performance on both the $R1_{.1}$ and $R5_{.1}$ scores by outperforming the previous best SooNet [37] by **1.2%** and **1.9%**. Together with its competitive performance on other metrics, RGNet demonstrates the importance of end-to-end modeling of long videos. Specifically, our performance of R1 is superior at all IoU thresholds, which testifies to the effectiveness of the proposed retrieval method. Notably, prior models rely on heuristic dot product similarity search applied to CLIP [38] features for the retrieval phase. In contrast, our shared retrieval feature and its unified optimization with the grounding objective drastically enhance the first stage retrieval, as seen in Tab. 2. It further helps the network capture more discriminative events crucial for detecting extremely tiny moments in hour-long videos.

5.3 Disjoint vs. Unified Architecture

Since the long video temporal grounding (LVTG) performance relies on the first stage of clip retrieval, it does not perfectly reflect the actual moment localization capability of its grounding network. To assess the standalone grounding performance, we designed an oracle experiment operating exclusively on the clip where the ground truth moment is present. Compared to the oracle grounding, the LVTG $R1_{.3}$ drops by 15.7% and 23.9% for Ego4D and MAD in Tab. 2. The poor grounding in LVTG emerges from the suboptimal retrieval accuracy of the disjoint network. Hence, retrieving clips where the ground truth moment exists is crucial, independent of the subsequent grounding accuracy. Regardless of the effectiveness of moment boundary detection, if the clip containing the moment boundary cannot be accurately retrieved. Our unified approach consistently improves all retrieval metrics compared to the disjoint baseline. Notably, we achieve a staggering 10.4% and 12.6% improved R@1 score for Ego4D and MAD. This improvement in clip retrieval leads to state-of-the-art LVTG performance. Moreover, the grounding also improves because of the stronger clip features generated by our encoder. Compared to the disjoint baseline, the oracle R1$_{.3}$ scores improved by about 6.7% and 4.0% for Ego4D and MAD, respectively. Hence, the unified end-to-end architecture proves more effective at LVTG by mutually improving both stages.

Data	Model	Clip Retrieval		Oracle Grounding		LVTG	
		R@1	R@5	R1$_{.3}$	R5$_{.3}$	R1$_{.3}$	R5$_{.3}$
Ego4d	Baseline	31.71	64.63	29.84	54.01	14.15	30.33
	Ours	**42.08**	**76.28**	**36.53**	**63.64**	**20.63**	**41.67**
MAD	Baseline	12.41	24.50	29.49	53.02	6.87	16.11
	Ours	**25.01**	**50.02**	**33.42**	**63.43**	**9.48**	**18.72**

Table 2: Empirical impact of two stages. To asses standalone grounding capability, we run an oracle experiment on the clip where the ground truth moment is present. The grounding capability degrades in LVTG evaluation because of incorrect selection by the disjoint clip retrieval network. Our unified model improves retrieval significantly, which leads to more effective temporal grounding in long videos.

5.4 Ablation Studies

This section reports detailed ablation studies on the proposed modules and loss functions. We also ablate with various clip lengths and retrieved clip numbers. Unless stated, the experiments are done with the `Default` network w/o NaQ annotations [39] for the Ego4D-NLQ dataset and CONE [16] as the baseline.

Model	$R1_{.3}$	$R5_{.3}$	$R1_{.5}$	$R5_{.5}$
RGNet (`Default`)	**18.28**	**34.02**	**12.04**	**22.89**
w/o Retrieval Token	17.80	33.99	11.25	22.32
w/o Sparsifier	16.12	31.57	9.91	20.47
w/o RG-Encoder*	14.15	30.33	8.18	18.02
w/o Contrastive Loss	17.41	32.12	10.79	22.80
w/o Attention Loss	16.21	31.59	9.91	20.53

Table 3: Cumulative ablation study for the proposed modules and losses of RGNet. The experiment is conducted w/o NaQ augmentation on the Ego4D dataset. In each step, we remove one of our proposed modules or losses to evaluate their individual impact. *denotes the baseline model.

Modules: In our initial experiments, we validate the effectiveness of each proposed module with a cumulative ablation in Tab. 3. The unified network demonstrates superior performance across all evaluation metrics, notably achieving an $R1_{.3}$ score of 18.28%. Note that the disjoint baseline performs 4.2% worse than our unified network. This performance difference can be attributed to our models unified architecture.

With the removal of the **retrieval token** for generating the clip feature in Eq. 5, considerably drops the score by about 0.5%. This proves that modeling both clip level context and frame level content are necessary to produces strong multimodal features for the grounding decoder. Subsequently, eliminating the **sparsifier**, there is a decline in $R1_{.3}$ performance of approximately 1.7%. This verifies the impact of the sparsifier in modeling the fine-grained event boundaries. Finally, removing the **RG-Encoder** yields a notable degradation of approximately 2.0% in the $R1_{.3}$ score, which substantiates its capability to model clip and frame level granularity jointly.

Loss Functions: Training RGNet without our **negative clip sampling** strategy for the contrastive loss (Tab. 3) results in an almost 0.9% decline in the $R1_{.3}$ score. Our sampling strategy mimics long-video setup closely during training, which improves event discrimination capability inside the same scene and surroundings.

By eliminating the **attention loss**, we observe a further reduction of 1.2% in grounding performance. The attention loss enables our encoder to model fine-grained events in long videos which is crucial for specific moment detection.

Number of Clips: The number of retrieved proposal clips significantly influences the Long Video Temporal Grounding (LVTG) performance. Our model consistently outperforms the baseline across various choices of top-k retrieved clips, as shown in Fig. 4. Even with only a single retrieved clip, our performance surpasses the baseline's optimal result for both datasets. Moreover, our model exhibits a less steep performance decline when fewer clips are retrieved. Importantly, with fewer clips, the grounding network runs fewer times, leading to an overall speedup. Therefore, our model strikes a favorable balance between efficiency and performance, showcasing superior performance than the baseline.

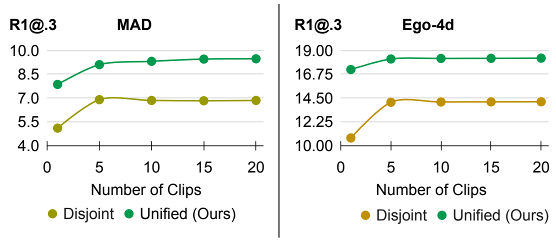

Fig. 4: Impact of number of retrieved clips. Reducing the number of clips speeds up the network execution time. While the baseline model experiences a significant drop in performance with this reduction, RGNet shows a noticeably smaller decline in performance under the same conditions.

Clip Length: Varying clip length impacts both retrieval and grounding performance. The number of clip decreases with longer clips, making the retrieval task easier. Fig. 5 reports a monotonically improving retrieval with longer clips. However, after a certain length, moment localization quality starts to deteriorate. With longer clips, the task becomes extremely difficult. We see the decline after the 180s and 48s for the MAD and Ego4D datasets. Hence, we set the default clip length of our method to these optimal lengths. Importantly, the unified network helps outscore the disjoint baseline on all tested clip lengths (inversely proportional to the clip numbers in Fig. 4).

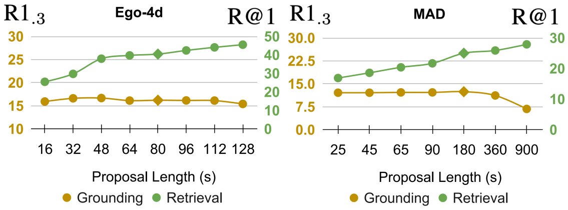

Fig. 5: Impact of retrieved clip length. Longer clips result in improved retrieval due to fewer candidates. However, grounding becomes exceedingly difficult with longer clips. For example, the grounding performance drops after 180 seconds and 48 seconds for the MAD and Ego4D datasets.

5.5 Qualitative Analysis

We visualize the qualitative predictions of RGNet compared to the disjoint baseline in Fig. 6. For the first and second queries, the baseline struggles to retrieve the correct clip where the scissor appears. RGNet accurately retrieves the clip with the searched object and precisely localizes the moment it appeared in the video. Disjoint retrieval often leads to incorrect clip selection, which the final grounding network cannot recover from. For the third query, the baseline selects the correct clip but fails to localize the precise moment the man closes the car door, indicating a gap between the two stages. RGNet mitigates this gap by leveraging features from the retrieved clip to improve moment localization. More visual outputs are included the supplementary materials.

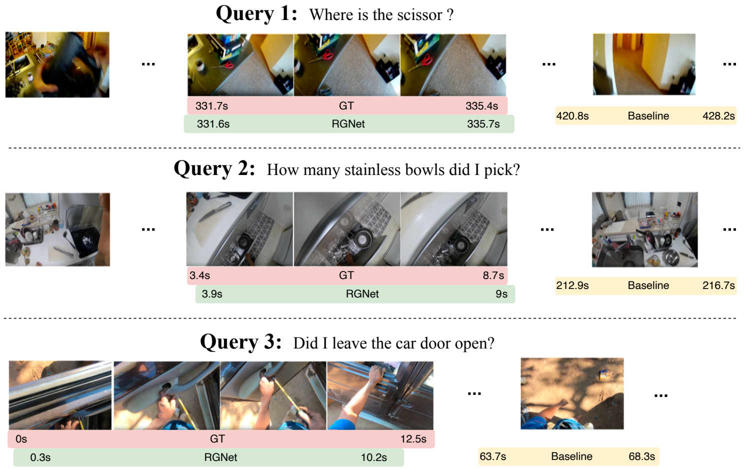

Fig. 6: Qualitative Results. The baseline fails to retrieve the correct clip in the first two queries. Since they primarily depict the same indoor room throughout the whole video, precise event discrimination is vital for accurate clip retrieval. In the third query, the baseline cannot identify the moment within the correctly retrieved clip, detecting it only after it has concluded. Precise event localization demands improved alignment between visual events and the queried text. RGNet correctly localizes all the moments.

6 Conclusion

We introduce an end-to-end model for long video temporal grounding, unifying the two stages of prevailing methods with shared features and mutual optimization. We conduct independent analyses of the two stages, shaping our solution based on their distinct impacts. This leads to a better understanding of fine-grained events in long videos. Our approach demonstrates state-of-the-art performance on challenging long video grounding datasets, validating its effectiveness. Like all LVTG methods, we rely on a pre-trained image encoder to extract the visual features. A promising future direction includes eliminating the need for an image encoder and training directly on the raw video data.

References

1. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: ICCV. pp. 1728–1738 (2021)
2. Barrios, W., Soldan, M., Ceballos-Arroyo, A.M., Heilbron, F.C., Ghanem, B.: Localizing moments in long video via multimodal guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13667–13678 (2023)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
4. Chen, S., Zhao, Y., Jin, Q., Wu, Q.: Fine-grained video-text retrieval with hierarchical graph reasoning. In: CVPR. pp. 10638–10647 (2020)
5. Chen, Z., Jiang, X., Xu, X., Cao, Z., Mo, Y., Shen, H.T.: Joint searching and grounding: Multi-granularity video content retrieval. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 975–983 (2023)
6. Cheng, F., Wang, X., Lei, J., Crandall, D., Bansal, M., Bertasius, G.: Vindlu: A recipe for effective video-and-language pretraining. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2023)
7. Dong, J., Li, X., Xu, C., Yang, X., Yang, G., Wang, X., Wang, M.: Dual encoding for video retrieval by text. IEEE TPAMI pp. 4065–4080 (2021)
8. Fang, B., Liu, C., Zhou, Y., Yang, M., Song, Y., Li, F., Wang, W., Ji, X., Ouyang, W., et al.: Uatvr: Uncertainty-adaptive text-video retrieval. arXiv preprint arXiv:2301.06309 (2023)
9. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal transformer for video retrieval. In: ECCV. pp. 214–229 (2020)
10. Gao, Z., Liu, J., Sun, W., Chen, S., Chang, D., Zhao, L.: Clip2tv: Align, match and distill for video-text retrieval. arXiv preprint arXiv:2111.05610 (2021)
11. Ge, Y., Ge, Y., Liu, X., Li, D., Shan, Y., Qie, X., Luo, P.: Bridging video-text retrieval with multiple choice questions. In: CVPR. pp. 16167–16176 (2022)
12. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 249–256. JMLR Workshop and Conference Proceedings (2010)
13. Gorti, S.K., Vouitsis, N., Ma, J., Golestan, K., Volkovs, M., Garg, A., Yu, G.: X-pool: Cross-modal language-video attention for text-video retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5006–5015 (2022)
14. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: CVPR. pp. 18995–19012 (2022)
15. Hannan, T., Koner, R., Bernhard, M., Shit, S., Menze, B., Tresp, V., Schubert, M., Seidl, T.: Gratt-vis: Gated residual attention for auto rectifying video instance segmentation. arXiv preprint arXiv:2305.17096 (2023)
16. Hou, Z., Zhong, W., Ji, L., Gao, D., Yan, K., Chan, W.K., Ngo, C.W., Shou, Z., Duan, N.: Cone: An efficient coarse-to-fine alignment framework for long video temporal grounding. arXiv preprint arXiv:2209.10918 (2022)
17. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144 (2016)
18. Jang, J., Park, J., Kim, J., Kwon, H., Sohn, K.: Knowing where to focus: Event-aware transformer for video grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13846–13856 (2023)

19. Jiang, H., Zhang, J., Huang, R., Ge, C., Ni, Z., Lu, J., Zhou, J., Song, S., Huang, G.: Cross-modal adapter for text-video retrieval. arXiv preprint arXiv:2211.09623 (2022)
20. Jin, P., Huang, J., Liu, F., Wu, X., Ge, S., Song, G., Clifton, D.A., Chen, J.: Expectation-maximization contrastive learning for compact video-and-language representations. NeurIPS (2022)
21. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539 (2014)
22. Lei, J., Berg, T.L., Bansal, M.: Detecting moments and highlights in videos via natural language queries. Advances in Neural Information Processing Systems **34**, 11846–11858 (2021)
23. Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T.L., Bansal, M., Liu, J.: Less is more: Clipbert for video-and-language learning via sparse sampling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7331–7341 (2021)
24. Lin, K.Q., Wang, A.J., Soldan, M., Wray, M., Yan, R., Xu, E.Z., Gao, D., Tu, R., Zhao, W., Kong, W., et al.: Egocentric video-language pretraining. arXiv preprint arXiv:2206.01670 (2022)
25. Lin, K.Q., Zhang, P., Chen, J., Pramanick, S., Gao, D., Wang, A.J., Yan, R., Shou, M.Z.: Univtg: Towards unified video-language temporal grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2794–2804 (2023)
26. Liu, N., Wang, X., Li, X., Yang, Y., Zhuang, Y.: Reler@ zju-alibaba submission to the ego4d natural language queries challenge 2022. arXiv preprint arXiv:2207.00383 (2022)
27. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: Dab-detr: Dynamic anchor boxes are better queries for detr. arXiv preprint arXiv:2201.12329 (2022)
28. Liu, Y., Li, S., Wu, Y., Chen, C.W., Shan, Y., Qie, X.: 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. pp. 3042–3051 (2022)
29. Liu, Y., Xiong, P., Xu, L., Cao, S., Jin, Q.: Ts2-net: Token shift and selection transformer for text-video retrieval. In: ECCV. pp. 319–335 (2022)
30. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
31. Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Li, J., Bharti, T., Zhou, M.: Univl: A unified video and language pre-training model for multimodal understanding and generation. arXiv preprint arXiv:2002.06353 (2020)
32. Ma, Y., Xu, G., Sun, X., Yan, M., Zhang, J., Ji, R.: X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In: ACM MM. pp. 638–647 (2022)
33. Maddison, C.J., Mnih, A., Teh, Y.W.: The concrete distribution: A continuous relaxation of discrete random variables. CoRR **abs/1611.00712** (2016), `http://arxiv.org/abs/1611.00712`
34. Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., Sun, L., Wang, J.: Conditional detr for fast training convergence. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3651–3660 (2021)
35. Moon, W., Hyun, S., Park, S., Park, D., Heo, J.P.: Query-dependent video representation for moment retrieval and highlight detection. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23023–23033 (2023)

36. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
37. Pan, Y., He, X., Gong, B., Lv, Y., Shen, Y., Peng, Y., Zhao, D.: Scanning only once: An end-to-end framework for fast temporal grounding in long videos. arXiv preprint arXiv:2303.08345 (2023)
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. pp. 8748–8763 (2021)
39. Ramakrishnan, S.K., Al-Halah, Z., Grauman, K.: Naq: Leveraging narrations as queries to supervise episodic memory. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6694–6703 (2023)
40. Soldan, M., Pardo, A., Alcázar, J.L., Caba, F., Zhao, C., Giancola, S., Ghanem, B.: Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In: CVPR. pp. 5026–5035 (2022)
41. Soldan, M., Xu, M., Qu, S., Tegner, J., Ghanem, B.: Vlg-net: Video-language graph matching network for video grounding. In: ICCV. pp. 3224–3234 (2021)
42. Wang, A.J., Ge, Y., Yan, R., Ge, Y., Lin, X., Cai, G., Wu, J., Shan, Y., Qie, X., Shou, M.Z.: All in one: Exploring unified video-language pre-training. arXiv preprint arXiv:2203.07303 (2022)
43. Wang, Q., Zhang, Y., Zheng, Y., Pan, P., Hua, X.S.: Disentangled representation learning for text-video retrieval. arXiv preprint arXiv:2203.07111 (2022)
44. Wang, Z., Sung, Y.L., Cheng, F., Bertasius, G., Bansal, M.: Unified coarse-to-fine alignment for video-text retrieval. In: The IEEE International Conference on Computer Vision (ICCV) (October 2023)
45. Xu, H., Ghosh, G., Huang, P.Y., Arora, P., Aminzadeh, M., Feichtenhofer, C., Metze, F., Zettlemoyer, L.: Vlm: Task-agnostic video-language model pre-training for video understanding. ACLliu2021hit (2021)
46. Xu, H., Ghosh, G., Huang, P.Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., Feichtenhofer, C.: Videoclip: Contrastive pre-training for zero-shot video-text understanding. In: EMNLP. pp. 6787–6800 (2021)
47. Xue, H., Hang, T., Zeng, Y., Sun, Y., Liu, B., Yang, H., Fu, J., Guo, B.: Advancing high-resolution video-language representation with large-scale video transcriptions. In: CVPR. pp. 5036–5045 (2022)
48. Xue, H., Sun, Y., Liu, B., Fu, J., Song, R., Li, H., Luo, J.: Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. ICLR (2023)
49. Yu, Y., Kim, J., Kim, G.: A joint sequence fusion model for video question answering and retrieval. In: ECCV. pp. 471–487 (2018)
50. Yu, Y., Ko, H., Choi, J., Kim, G.: End-to-end concept word detection for video captioning, retrieval, and question answering. In: CVPR. pp. 3165–3173 (2017)
51. Zhang, B., Jin, X., Gong, W., Xu, K., Zhang, Z., Wang, P., Shen, X., Feng, J.: Multimodal video adapter for parameter efficient video text retrieval. arXiv preprint arXiv:2301.07868 (2023)
52. Zhang, C., Gupta, A., Zisserman, A.: Helping hands: An object-aware ego-centric video recognition model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13901–13912 (2023)
53. Zhang, H., Sun, A., Jing, W., Zhou, J.T.: Span-based localizing network for natural language video localization. arXiv preprint arXiv:2004.13931 (2020)

54. Zhang, H., Sun, A., Jing, W., Zhou, J.T.: The elements of temporal sentence grounding in videos: A survey and future directions. arXiv preprint arXiv:2201.08071 **1**(2) (2022)
55. Zhang, S., Peng, H., Fu, J., Luo, J.: Learning 2d temporal adjacent networks for moment localization with natural language. In: AAAI. pp. 12870–12877 (2020)
56. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)