A Supplementary

A.1 Further Technical Deatails

Motivation of contextual feature R. Most prior works use a heuristic-based scoring function borrowed from the short video-text retrieval methods and apply them directly to longer videos. This often leads to poor performance in fine-grained event retrieval within very long videos. Instead, we use a learnable retrieval token **R**, facilitating more effective fine-grained event understanding in long videos. Our introduced retrieval token unifies the retrieval and grounding objectives since it is used as input into the grounding decoder. This enables our model to maximize the synergy between retrieval and grounding stages, significantly improving both clip retrieval (+**12.6% R1**) and grounding (+**4.2% R1@.3**) on Ego4D (Tab. 4). For both training and inference, we randomly initialize **R** and then replicate it for all clips. Thus, the initial features $\mathbf{R}_i$ and $\mathbf{R}_j$ of clips i and j are replicated copies of a randomly initialized feature **R**.

Method	Grounding		Retrieval	
	R1@.3 (↑)	R5@.3 (↑)	R1 (↑)	R5 (↑)
Heuristic	14.15	30.33	12.41	24.50
Learned	**18.28**	**34.02**	**25.01**	**50.02**

Table 4: Performance Comparison on Ego4D (ref. Tab 2, 3)

Runtime Analysis. We divide a video of length L into clips of length C. The number of text tokens is T, and the hidden dimension is D. The time complexity of self-attention is $O(LCD)$, cross-attention is $O(LTD)$, and the full model is $O(L(C+T)D)$. The runtime of our model scales linearly with video length. We compared our runtime with the CONE baseline on a GTX A100 and reported the total time for the Ego4D validation set. CONE achieves a 14.15% R1@0.3 score with a 39.9-second inference time. In contrast, RGNet scores 20.63% R1@0.3 in just 24.2 seconds, making it **1.7x** faster than CONE. Our retrieval module achieves superior performance with fewer retrieved clips, substantially reducing the runtime, as shown in the leftmost subfigure below.

Convergence. RGNet converges in 35 and 200 epochs on MAD and Ego4D. The rightmost figure above shows that on MAD our model converges earlier due to a larger dataset.

Number of the Negative Samples. We trained Ego4D with a batch size of 32, resulting in 1024 clip-text pairs (32 positives and 992 negatives). We use 8 positive and 56 negative samples for MAD to fit its longer clips in a GPU.

A.2 Further Ablation Studies

This section provides an in-depth analysis of various components of RGNet, such as the sparsifier and transformer decoder. Furthermore, we conduct ablation studies on pre-training data and frame rate. All experiments are carried out on the Ego4D-NLQ dataset.

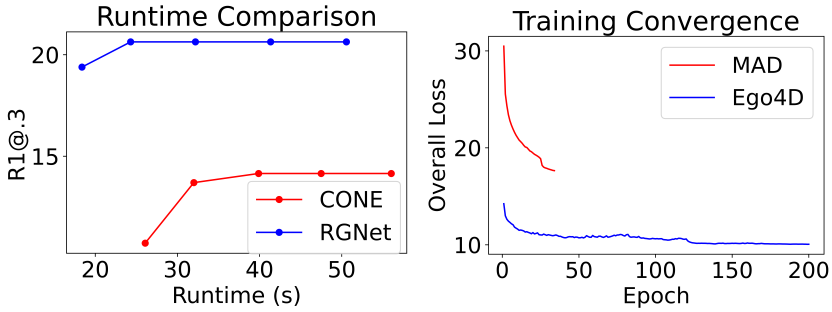


Fig. 7: (left) Runtime comparison w.r.t. baseline model CONE. **(right)** Training convergence curve for both MAD and Ego4D datasets.

Sparsifier: We use a differentiable softmax function to enable end-to-end training of the sparsifier. Let's denote p^j as the linear projection of f^j from Eq. 2.

$$p^j = \text{linear}(f^j) \tag{11}$$

To derive a categorical variable G^j characterized by probabilities ${\pi_1}^j = \sigma(p^j)$ and ${\pi_0}^j = 1 - \sigma(p^j)$, where σ is the sigmoid operation, we can reframe the sampling procedure for G^j through the utilization of the Gumbel-Max trick, outlined as follows:

$$G^j = \arg\max_k \left\{ \log\left(\pi_k^j\right) + g_k : k = 0, 1 \right\} \tag{12}$$

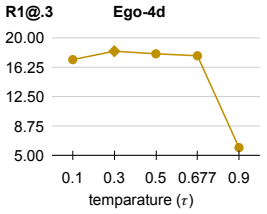
Here, the set $\{g_k\}_{k=0,1}$ consists of independently and identically distributed (i.i.d.) random variables sampled from the $Gumbel(0, 1)$ distribution. Considering the non-differentiable characteristic of the argmax operation, we employ an approximation for G^j using a differentiable, soft version $\hat{G}^j$, derived from the Gumbel-Softmax relaxation [17, 33].

$$\hat{G}^j = \frac{\exp\left(\left(\log\left(\pi_1^j\right) + g_1\right)/\tau\right)}{\Sigma_{k\in 0,1} \exp\left(\left(\log\left(\pi_k^j\right) + g_k\right)/\tau\right)} \tag{13}$$

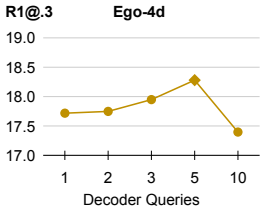
To ensure differentiability with respect to the discrete samples G^j, we employ the straight-through trick [17] and utilize the gradients of $\hat{G}^j$ as an approximation for the gradients of G^j in the backward pass.

The Gumbel-Softmax distribution serves as an interpolation between discrete one-hot-encoded categorical distributions and continuous categorical densities. For low temperatures ($\tau = 0.3$), the expected value of a Gumbel-Softmax random variable approaches the expected value of a categorical random variable with the same logits. As the temperature increases ($\tau = 0.9$), the expected value converges to a uniform distribution over the categories. Fig. 8a shows that we achieve the

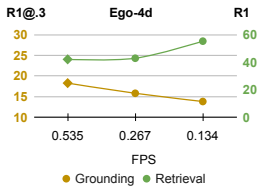
best performance with τ to be 0.3, which means the differentiation between the relevant and irrelevant frames is beneficial during the retrieval.



(a) Low temperature enables our sparsifier to approximate a categorical distribution over the relevancy of frames and achieves better performance.



(b) Increasing the number of decoder queries up to 5 enhances performance, while further increments start to deteriorate the results.



(c) While decreasing the FPS improves retrieval accuracy, it deteriorates the grounding performance due to the loss of temporal information.

Fig. 8: Ablation Studies on the sparsifier, transformer decoder, and input FPS.

Transformer Decoder: RGNet utilizes learnable decoder queries [27] to localize the moments. The number of queries equates to the number of predicted moments from each retrieved proposal. With more queries, the decoder can detect moments in different temporal locations of various widths. However, for long videos, the decoder runs on multiple retrieved clips, and increasing the queries beyond 5 leads to an increasing number of predicted moments, which decreases the LVTG performance (ref. Fig. 8b). Consequently, we set the number of queries in our decoder to 5.

Frame Rate: Manual reduction of frame rate (FPS) yields enhanced retrieval accuracy but significantly degrades grounding performance, as shown in Fig. 8c. Lowering the FPS serves as a heuristic sparsification strategy that aids retrieval but results in temporal information loss, leading to diminished overall performance. In contrast, our learned sparsifier enhances retrieval accuracy without compromising grounding performance, presenting a superior alternative to manual sparsification.

Method	NaQ	R1$_{.3}$	R5$_{.3}$	R1$_{.5}$	R5$_{.5}$
VSLNet	✗	5.45	10.74	3.12	6.63
EgoVLP	✗	10.84	18.84	6.81	13.45
ReLER	✗	14.66	17.84	8.67	11.54
RGNet (Ours)	✗	**18.28**	**34.02**	**12.04**	**22.89**
VSLNet	✓	10.26	19.01	5.81	12.67
EgoVLP	✓	15.90	26.38	9.46	17.80
ReLER	✓	19.31	23.62	11.59	15.75
RGNet (Ours)	✓	**20.63**	**41.67**	**12.47**	**25.08**

Table 5: Impact of NaQ. We compare RGNet on the Ego4D-NLQ dataset with and w/o NaQ annotations. RGNet achieves the best performance in both cases.

NaQ Pretraining: Similar to the prior state-of-the-art model [39], we employ NaQ annotations to pre-train RGNet on the Ego4D dataset. The grounding annotations of NaQ are automatically generated from the ground truth narrations of the Ego4D dataset. With this pretraining, we improve $R@1_{.3}$ and $R@1_{.5}$ by **1.32%** and **18.05%** (refer to Table 5). Importantly, without the extra NaQ annotations, RGNet demonstrates a larger improvement of **3.62%** for $R@1_{.3}$. These results highlight RGNet's superior performance, even in scenarios with limited data. The evaluation of the Ego4D test set is exclusively available on their official server, which is presently closed. Therefore, in alignment with recent publications [37,52] in ICCV'23, we present our performance of Ego4D-NLQ on the validation set.

A.3 Qualitative Results

We visualize the relevance score of all the video clips and clip frames for a single video in Fig. 10. For both the clip and frame levels, our model scores higher on the ground truth regions than the baseline. Then, we visualize some successful moment localization on both MAD and Ego4D datasets in Tab. 9 and 11.

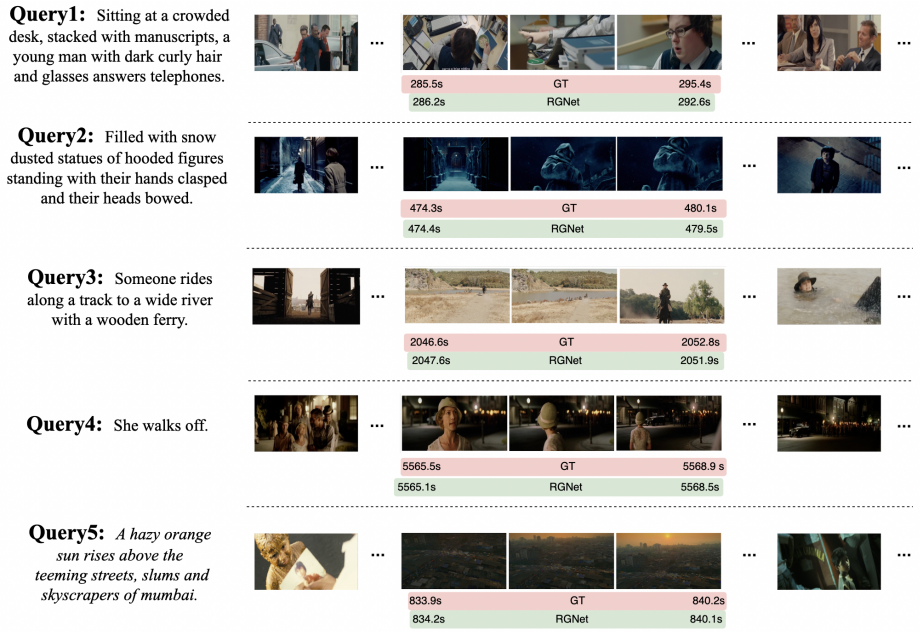


Fig. 9: Qualitative results on MAD. RGNet successfully localizes moments from hour-long movies by parallelly processing them in clip and frame level granularity.

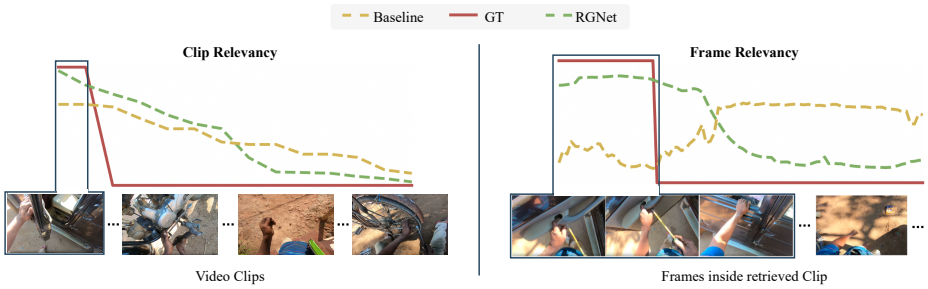


Fig. 10: Clip and Frame Relevancy. We visualize the relevancy score of video clips **(left)** and proposal frames **(right)** for the query "*Did I leave the car door open*?" from Ego4D-NLQ. We present the score for both RGNet and the disjoint baseline model in green and yellow, respectively. RGNet approximates the ground truth clip and frames better than the baseline in both stages.

Fig. 11: Qualitative results of Ego4D. RGNet can localize fine-grained events in long videos across various scenes and scenarios.