

Supplementary Materials: Maskable Retentive Network for Video Moment Retrieval

Jingjing Hu

Hefei University of Technology,
School of Computer Science and
Information Engineering (School of
Artificial Intelligence)
xianhjj623@gmail.com

Dan Guo*

Hefei University of Technology,
Institute of Artificial Intelligence
(IAI), Hefei Comprehensive National
Science Center
guodan@hfut.edu.cn

Kun Li

CCAI,
Zhejiang University
kunli.hfut@gmail.com

Zhan Si

Anhui University, Department of
Chemistry and Centre for Atomic
Engineering of Advanced Materials
naa0528@stu.ahu.edu.cn

Xun Yang*

MoE Key Laboratory of
Brain-inspired Intelligent Perception
and Cognition, University of Science
and Technology of China
xyang21@ustc.edu.cn

Meng Wang*

Hefei University of Technology,
Institute of Artificial Intelligence
(IAI), Hefei Comprehensive National
Science Center
eric.mengwang@gmail.com

1 THEORETICAL ANALYSIS

Bi-recurrent and Parallel Equivalence. We state the theoretical rationale for how the **bidirectional retention mechanism** \mathcal{M}_{RD} (Eq.4 in main paper) we apply to visual branch has a bidirectional recursive reasoning property. In original Retention Network [6], given input $X \in \mathbb{R}^{|x| \times d}$ ($|x|$ is the length of word tokens, d is the feature dimension), define the current state as s_n and current output as o_n , apply a linear transform to encode sequence information recurrently with the weight matrix A and bias K , the unidirectional **recurrent** retention formulas are

$$\begin{aligned} s_n &= A s_{n-1} + K_n^\top v_n, \quad A \in \mathbb{R}^{d \times d}, K_n \in \mathbb{R}^{1 \times d}, \\ o_n &= Q_n s_n = \sum_{m=1}^n Q_n A^{n-m} K_m^\top v_m, \quad Q_n \in \mathbb{R}^{1 \times d}. \end{aligned} \quad (\text{Eq.1 in [6]})$$

In our paper, we consider applying the retention mechanism on visual sequence modeling, and propose a new bidirectional retention expressed as a square matrix \mathcal{M}_{RD} that decays exponentially along the diagonal to both sides controlled by the decay factor γ , allowing for **parallelized training**. Specifically, we apply the new BiRetention on visual branch with the input $X_v \in \mathbb{R}^{T \times d}$ (T is the number of video clip tokens):

$$\begin{aligned} \text{BiRetention}(V_v) &= (QK^\top \odot \mathcal{M}_{RD})V_v; \\ (\mathcal{M}_{RD})_{n,m} &= \gamma^{|n-m|}, \end{aligned} \quad (\text{Eq.3 \& Eq.4 revisited})$$

where $\mathcal{M}_{RD} \in \mathbb{R}^{T \times T}$. Here, we can denote the $Q = X_v W_Q \odot \Theta$, $K = X_v W_K \odot \bar{\Theta}$ with the addition of position embedding xPos [7] ($\Theta_n = e^{in\theta}$), where W is learnable parameter, Θ and $\bar{\Theta}$ are a conjugate complex pair. Then we expand the formula for BiRetention based on the definition of matrix multiplication as

$$\begin{aligned} o_n &= \sum_{m=1}^T \gamma^{|n-m|} (Q_n \Theta_n) (K_m \bar{\Theta}_m)^\top v_m; \\ &= \sum_{m=1}^T \gamma^{|n-m|} (Q_n e^{in\theta}) (K_m e^{im\theta})^\dagger v_m, \end{aligned} \quad (\text{A1})$$

where \dagger is the conjugate transpose, current output o_n (row n of the $\text{BiRetention}(X_v)$) depends on both its forward (the column index is greater than n) and backward (the column index is less than n) clip tokens. We then assume that a diagonalized matrix $A \in \mathbb{R}^{T \times T}$ has the following representation:

$$\begin{aligned} A &= P(\gamma e^{i\theta})P^{-1} \\ A^{|n-m|} &= P(\gamma e^{i\theta})^{|n-m|}P^{-1} \\ &= P\gamma^{|n-m|}e^{i(|n-m|)\theta}P^{-1}. \end{aligned} \quad (\text{A2})$$

In the above formula, P is a diagonal matrix, which can be absorbed into the Q_n and K_m , and the position embedding $e^{i(|n-m|)\theta}$ can be assigned to Q_n and K_m , respectively. So we update the BiRetention formula in Eq. A1, and represent it as a bidirectional recurrent formula with backward recurrent b_n and forward recurrent $f_{(T-n)}$:

$$\begin{aligned} o_n &= \sum_{m=1}^T Q_n A^{|n-m|} K_m^\top v_m \\ &= \sum_{m=1}^n Q_n A^{n-m} K_m^\top v_m + \sum_{m=n+1}^T Q_n A^{m-n} K_m^\top v_m \\ &= Q_n b_n + Q_{(T-n)} f_{(T-n)}, \quad (n = 1, 2, \dots, T). \end{aligned} \quad (\text{A3})$$

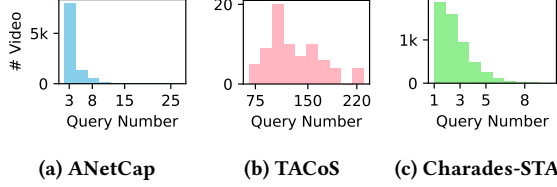
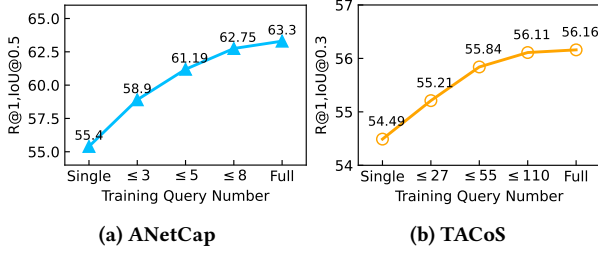
$$\begin{aligned} b_n &= A b_{n-1} + K_n^\top v_n, \quad (n = 1, 2, \dots), \\ f_{(T-n)} &= A f_{(T-n-1)} + K_{(T-n)}^\top v_{(T-n)}, \quad (n = T-1, T-2, \dots). \end{aligned} \quad (\text{A4})$$

Eq. A3 changes the unidirectional recurrent retentive reasoning formula into a bidirectional one, and in Eq. A4, we simplify the parameters in the two recurrent formulas to be represented by the same matrices A and K_n , which will be optimized in parallelized training. Compared Eq. A4 to Eq.1 in [6], we change the RNN formula from unidirectional to bidirectional in the BiRetention. Thus, theoretically, **the bidirectional retention mechanism \mathcal{M}_{RD} (Eq.4 in main paper) we apply to visual branch has the bidirectional sequential inference property like RNNs and the parallelizable training like Transformers.**

*Corresponding authors

Table A1: R@1 performance comparison of different retention decay factor γ in bidirectional retention mask \mathcal{M}_{RD} .

γ	ANetCap				Charades-STA			
	0.3	0.5	0.7	mIoU	0.3	0.5	0.7	mIoU
0.97	76.43	62.54	41.60	56.72	73.49	59.38	37.66	52.66
0.98	77.57	63.30	42.68	57.62	74.41	60.27	38.39	53.14
0.99	76.45	62.36	42.49	57.03	74.14	60.19	38.20	53.03

**Figure A1: Statistics on the query number per video. The detests can be divided into three categories: large query size (TACoS, most sizes are 110), middle query size (ANetCap, most sizes are 3), and small query size (Charades-STA, most sizes are 1, and the query description is often ambiguous, semantically insufficient and time overlapping as the video is too short for manually annotating events).****Figure A2: The impact of training query number m on ANetCap and TACoS dataset for NLMR task. “Full” denotes all queries for a video are entered simultaneously. When our MRNet is trained with the single-query mode (R@1, IoU@0.5 is 55.40 on ANetCap, R@1, IoU@0.3 is 54.49 on TACoS), we also achieve promising performance compared to the state-of-art methods listed in Tab. 2 of the main paper.**

2 ADDITIONAL EXPERIMENTS

We conduct additional experiments to explore some of the other important properties of our MRNet in order to get a fuller picture of the role of our model in the MR tasks. All additional experimental outlines are listed below:

- (1) **Hyperparameter Setting: Q1:** How to set the decay factor γ of the \mathcal{M}_{RD} to be optimal?
- (2) **About Training Mode: Q2:** Would it be beneficial for the model to refer to more query description information during training?
- (3) **About Cross-modal Attention: Q3:** Is a cross-modal base environment necessary for video sequence modeling?
- (4) **About Model Complexity: Q4:** How about the total parameter quantity and inference speed of the model?

Table A2: R@1 performance comparison of the model w/o cross-modal attention guidance.

Dataset	Cross-modal	IoU@0.3	IoU@0.5	IoU@0.7	mIoU
ANetCap	✗	62.18	45.11	26.35	44.82
	✓	77.57	63.30	42.68	57.62
TACoS	✗	45.09	33.27	19.17	31.61
	✓	71.98	41.31	22.27	39.45
Charades-STA	✗	74.65	60.30	38.20	53.27
	✓	71.88	55.91	34.44	50.61

Table A3: Analysis with respect to performance (R@1), model parameters ($\times 10^6$) and inference speed (s/query) on ANetCap dataset for NLMR task. “Mode” refers to the training mode: single-query training or multi-query training. * denotes the estimated number of parameters based on the model backbone network, due to lack of access to the source code.

Mode	Method	Venue	IoU@0.5	# Param.	Infer. Speed
Single	2D-TAN [12]	AAAI'20	44.51	84.94	0.061
	MS-2D-TAN [11]	TPAMI'21	46.16	479.46	0.141
	MGPN [5]	SIGIR'22	47.92	5.12	0.115
	BMRN [4]	CVPR'23	48.47	90.00*	-
	MRNet-S (Ours)		55.40	77.16	0.023
Multi	MMN [9]	AAAI'22	48.59	152.22	0.063
	PTRM [13]	AAAI'23	50.44	152.25	0.038
	DFM [8]	ACM MM'23	45.92	87.00*	-
	MRNet-M (Ours)		63.30	77.16	0.023

- (5) **About Case Visualization: Q5:** A more comprehensive presentation of moment retrieval results on five datasets.

2.1 Retention Decay Analysis (Q1)

In Bidirectional Retention Mask \mathcal{M}_{RD} in Eq.4, the value of decay factor γ will explicitly control the valid receptive field range of the current token, then the model’s temporally contextual learning ability will be affected. When setting the γ , we consider the current token to be able to view all tokens before and after it by a level of magnitude. From Tab. A1, $\gamma = 0.98$ is optimal.

2.2 More Analysis on Training Mode (Q2)

From the main paper (Tab. 2, Tab. 3 and Tab. 4), we can conclude that on datasets ANetCap and TACoS, when model refers to all query description information related to the same video during training, the model (MRNet-M) offers significant performance gains over the single-query-only model; actually, the single-query model MRNet-S has already achieved promising performance with the recent state-of-arts. This conclusion can be also observed in Tab. A3, e.g., the R@1, IoU@0.5 of MRNet-S is 55.40, which is much higher than the SOTA single-training BMRN [4] and multi-training PTRM [13].

Based on the statistics of the number of queries on the datasets in Fig. A1, we can simply divide the three datasets into two categories: 1) Multi-query: ANetCap and TACoS, and 2) Approximate single-query: Charades-STA. We further investigate the following question on the multi-query datasets: Whether it is more helpful for our MRNet to train a stronger moment retrieval model by **referring to more query language descriptions during training**? As the results in Fig. A2 indicate, when we progressively increase the number of queries trained at a time, the performance of the model

improves, suggesting that our model is able to efficiently merge the semantics of the language and the video in order to maximize the effect of moment retrieval. Ideally, the model performs optimally when it refers to all linguistic descriptions associated with a same video during training.

2.3 Cross-modal Attention Analysis (Q3)

Video Moment retrieval (MR) is an important cross-modal task, in our main paper, we focus on exploring the problem of **optimizing video sequence modeling in a cross-modal environment** and demonstrate the effectiveness of our proposed retention optimization approach. Here, we explore a more fundamental aspect of MRNet, that is, **the need for “multimodal guidance” in the MRT block**. We design two experimental setups: one with MRT block that includes cross-modal attention, *i.e.*, the model is able to attend to both language and video information, and the other with independent self-modal attention for video and query, *i.e.*, the model can only attend to one of the information alone.

The results on three different datasets are shown in Tab. A2, we can see that for datasets ANetCap and TACoS, the cross-modal environment significantly improves model performance, but for Charades-STA dataset, **purely self-modal modeling of video sequences is instead more effective**. Normally, cross-modal guidance between language and video promotes the model’s semantic alignment of the two modalities, which in turn improves performance [1, 3, 10], as indicated by the results on datasets ANetCap and TACoS in Tab. A2. However, for the Charades-STA dataset, there’s an interesting anomaly occurring, and we have found some explainable factors:

- **Semantic Consistency:** Most videos in Charades-STA dataset contains contradictory annotations with overlapping moments, *e.g.*, for video (id is “FSOFF”) with the total duration of 19.79s, there are two annotations “another person *walks by the takes off* their shoes.” and “person *proceeds to take* their shoes *off*.” that correspond to the same video moment 11.80s-17.97s, this causes semantic inconsistencies in queries and moments to appear in the cross-modal attention map, which affects the computation of attention and pulls down the performance of moment retrieval.
- **Query Length:** The number of words in a query is a direct reflection of the semantic richness of the language, however, the query length in the Charades-STA dataset averages 7 and contains more pronouns, this leads to ambiguous and semantically insufficient language description. Thus the attention interaction of language and video may instead impair the model’s comprehension of video information.

According to experimental results and our analyses, it can be inferred that when the quality of the annotations is relatively low, or there are semantic inconsistencies between query annotations and video moments, we can consider omitting cross-modal guidance and instead focus solely on video sequence modeling. As shown in Tab. A2, with the purely self-modal modeling of video sequence, our MRNet report new state-of-the-art performances (*i.e.*, R@1, IoU@0.3 is 74.65 as shown in main paper) on Charades-STA dataset.

2.4 More Analysis on Model Efficiency (Q4)

We have presented the model parameters and the performance of our model compared to recent MR methods in Fig. 2 of the main paper. And our MRNet demonstrates the optimal trade-off between model size and accuracy. In this section, we give a more detailed efficiency comparison in Tab. A3. All experiments are conducted on an RTX 2080Ti GPU. Noting that the parameters of MRT block in our MRNet are only 1.98 M as shown in Tab. ??, and the total parameters of whole back-end of MMN [9] (a typical proposal-based method) are 74.59 M, which indicates that **our proposed MRT block occupies a very lightweight proportion of the entire model parameters**. It can be found that our MRNet achieves a balance of model size and performance in both the single-query training moment retrieval methods [4, 5, 11, 12] and the multi-query training methods [8, 9, 13]. Moreover, our MRNet achieves faster inference compared to the MR methods in Tab. A3, the average inference time per query is 0.023 s.

2.5 More Visualization Cases (Q5)

To make it clearer how well our MRNet works on the five benchmarks for the three MR tasks, we provide additional visualizations to complement the analysis. We show the distribution of GT moment samples in the test (or validation) set for all datasets on the right side of Figs. A3~A5 and randomly select a video sample from among them for specific visualization and analysis.

2.5.1 Natural Language Moment Retrieval.

The Fig. A3 shows the visualisation samples for most popular video moment retrieval task of NLMR (*Natural Language Moment Retrieval*). And there are three widely used datasets ActivityNet Captions (ANetCap), Caharades-STA and TACoS. The data distribution of these three datasets exhibits different characteristics:

- **Different moment distributions** as shown in the Fig. A3 (right, **green** contour maps). For ANetCap dataset, there are **four peak regions** of target moments: the beginning of the video, the middle of the video, the end of the video, and the moment from across the entire video. For Charades-STA dataset, two peaks at the beginning and end of the video, the entire distribution is close to the diagonal, containing **a large number of short moments**. For TACoS dataset, a lot of the moments are focused on the beginning of the video.
- **Different query information length (avg.):** 15 words in ANetCap, 10 words in TACoS and 7 words in Charades-STA.
- **Different video duration (avg.):** 117.61s in ANetCap, 287.14s in TACoS and 30.59s in Charades-STA.
- **Different Video Scene:** Open-world activities in ANetCap, Cooking in TACoS and indoors in Charades-STA.

From the visualization results, our model shows more accurate moment retrieval compared to the MMN [9] **across all three NLMR datasets**, this indicates our added cross-modal guidance and video sequence modeling approach effectively corrects the moment prediction bias of MMN. For example, in Fig. A3 (b), the challenge with this sample is that the back-and-forth continuous action makes the **event boundaries difficult to capture**, both the beginning and ending boundaries of the MMN prediction contain some background clips, while our MRNet predicts **nearly no**

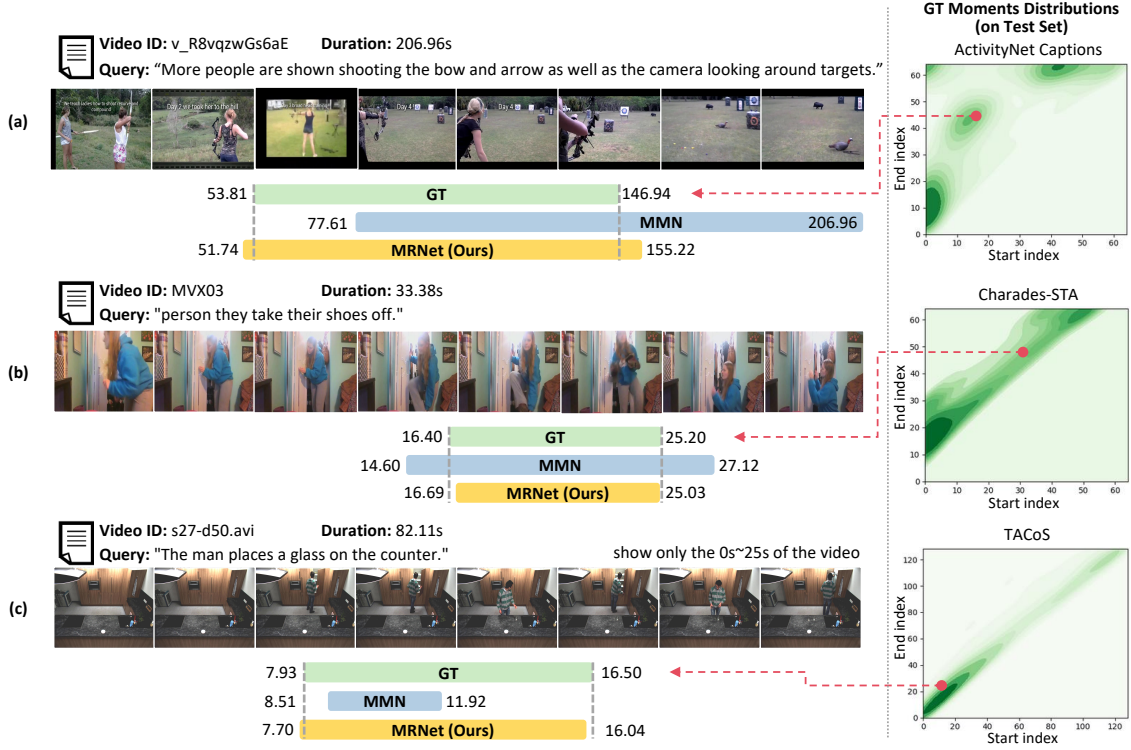


Figure A3: More visualization samples for NLMR (Natural Language Moment Retrieval) task, on three widely used datasets: (a) ActivityNet Captions, (b) Caharades-STA and (c) TACoS. On the right is the distribution of target moments for the test set corresponding to each dataset, we randomly select a sample case in the GT distribution to visualize. Compared with MMN [9], our MRNet can locate more accurate temporal regions.

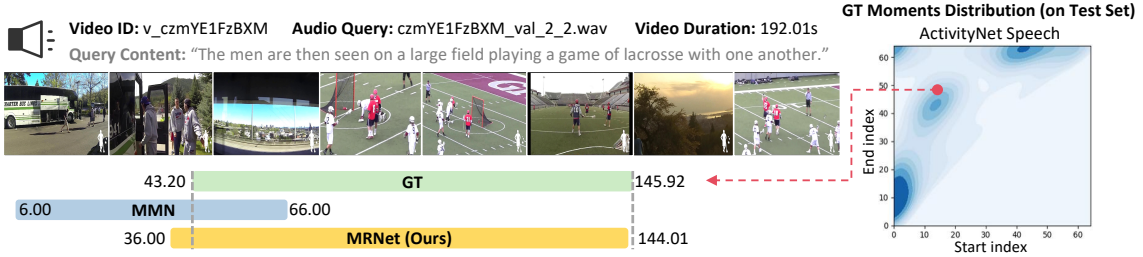


Figure A4: More visualization samples for SLMR (Spoken Language Moment Retrieval) task, on recent dataset ActivityNet Speech. The results show that our MRNet is also more effective than MMN for SLMR task.

gap boundaries with GT. This underscores the effectiveness of our proposed MRNet, which achieves background de-redundancy learning and contextual temporal correlation learning from videos.

2.5.2 Spoken Language Moment Retrieval.

We show additional visualization example on the recently proposed ActivityNet Speech dataset in Fig. A4. When using the audio as the query input, the MMN method shows some semantic bias in the prediction of the test set sample "v_czmYE1FzBXM", essentially recognizing no video clips about "large field playing a game of lacrosse with one another". Our MRNet, on the other hand, predicts a **more precise range of video semantics**, this again demonstrates the

validity of our proposed MRT block, which additionally optimizes video sequence modeling in a cross-modal environment, enhancing contextual learning of video semantics of the model.

2.5.3 Moment Retrieval + Highlight Detection.

An additional visualization example on recent MR-related multitasking dataset QVHighlights is in Fig. A5. We also present the highlihtness scores of the clip-by-clip predictions on the validation set video sample "izeyQalOwGg_60.0_210.0". In fact, GT's real-time annotations are 6s-10s, 12s-16s and 18s-36s and we simplify its representation in Fig. A5 as 6s-36s. With the Moment Retrieval task as the main focus, we can see the prediction of the highlighted scores

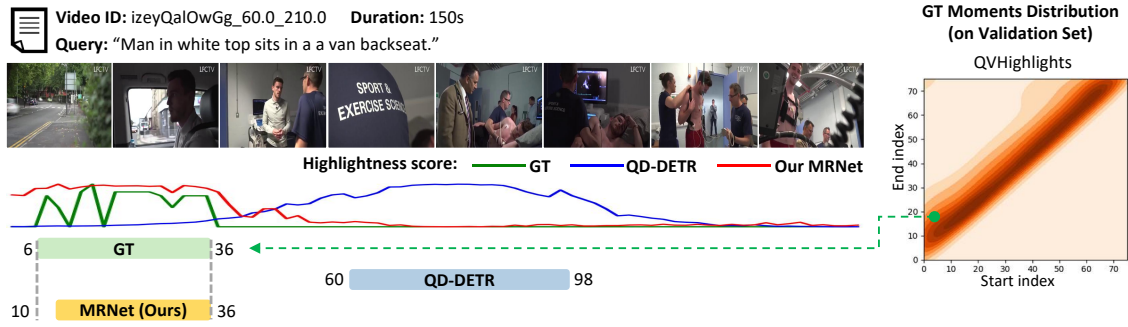


Figure A5: More visualization samples for MR+HD (Moment retrieval and Highlight Detection) task, on recent dataset QVHighlights. Compared with QD-DETR [2], our MRNet can predict more accurate temporal regions and highlightness scores.

as an aid to this task. From the visualization, compared to QD-DETR, our approach to this multi-task learning has **the advantage of more accurate semantic understanding of video**.

REFERENCES

- [1] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. 2020. Jointly cross-and self-modal graph attention network for query-based moment localization. In *ACM MM*. 4070–4078.
- [2] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. 2023. Query-dependent video representation for moment retrieval and highlight detection. In *CVPR*. 23023–23033.
- [3] Jonghwan Mun, Minsu Cho, and Bohyung Han. 2020. Local-global video-text interactions for temporal grounding. In *CVPR*. 10810–10819.
- [4] Muah Seol, Jonghee Kim, and Jinyoung Moon. 2023. BMRN: Boundary Matching and Refinement Network for Temporal Moment Localization with Natural Language. In *CVPRW*. 5571–5579.
- [5] Xin Sun, Xuan Wang, Jialin Gao, Qiong Liu, and Xi Zhou. 2022. You need to read again: Multi-granularity perception network for moment retrieval in videos. In *SIGIR*. 1022–1032.
- [6] Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. 2023. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621* (2023).
- [7] Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhami, Vishrav Chaudhary, Xia Song, and Furu Wei. 2022. A length-extrapolatable transformer. *arXiv preprint arXiv:2212.10554* (2022).
- [8] Xin Wang, Zihao Wu, Hong Chen, Xiaohan Lan, and Wenwu Zhu. 2023. Mixup-Augmented Temporally Debiased Video Grounding with Content-Location Disentanglement. In *ACM MM*. 4450–4459.
- [9] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. 2022. Negative sample matters: A renaissance of metric learning for temporal grounding. In *AAAI*, Vol. 36. 2613–2623.
- [10] Mingxing Zhang, Yang Yang, Xinghan Chen, Yanli Ji, Xing Xu, Jingjing Li, and Heng Tao Shen. 2021. Multi-stage aggregated transformer network for temporal language localization in videos. In *CVPR*. 12669–12678.
- [11] Songyang Zhang, Houwen Peng, Jianlong Fu, Yijuan Lu, and Jiebo Luo. 2021. Multi-scale 2d temporal adjacency networks for moment localization with natural language. *TPAMI* 44, 12 (2021), 9073–9087.
- [12] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, Vol. 34. 12870–12877.
- [13] Minghang Zheng, Sizhe Li, Qingchao Chen, Yuxin Peng, and Yang Liu. 2023. Phrase-level Temporal Relationship Mining for Temporal Sentence Localization. In *AAAI*. 3669–3677.