Localizing Moments in Long Video Via Multimodal Guidance (*Supplementary Material*)

Wayner Barrios[1], Mattia Soldan[2], Alberto Mario Ceballos-Arroyo[3], Fabian Caba Heilbron[4], Bernard Ghanem[2]

[1] Dartmouth [2]King Abdullah University of Science and Technology (KAUST) [3]Northeastern University [4]Adobe Research

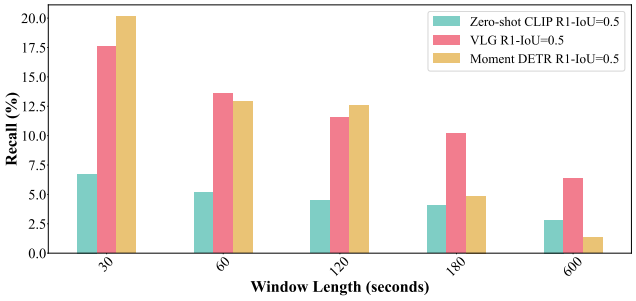
This document provides detailed information on the following topics: **(A)** Short-form video setup for the baseline models on MAD [8] dataset: Zero-shot CLIP [8], VLG-Net [9], and Moment-DETR [4]. **(B)** Comprehensive results using Moment-DETR [4] and VSL-Net [11] in Ego4D [2]. **(C)** In-depth results of each ablation study presented in the main paper for the Guidance Model. Lastly, **(D)** Evaluating the inference time of the Guidance Model.

A. Short-form video setup

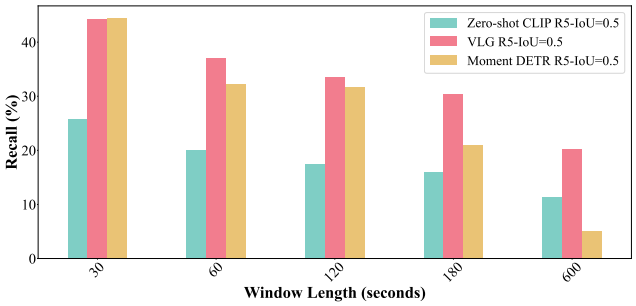
This section presents additional results for the short-form moment localization setup. To this end, we use the baseline grounding models reported in the main paper and evaluate them using several short-form video setups, following [8]. To be more specific, we first split a single (long) movie into non-overlapping windows (short videos). Then, we assign to each of these windows the ground-truth annotation with the highest temporal overlap. We evaluate five short-form video setups using 30, 60, 120, 180, and 600 second windows. Table 1 showcases the performance of the three baseline models in the (30 seconds) short-form video setup. Moment-DETR [4] achieves the highest scores for all metrics. This result indicates that the model is remarkably good for moment localization in short videos, however, it utterly fails to ground moments in video setups with much longer windows (see Table 1 in the main paper).

Model	IoU=0.1		IoU=0.3		IoU=0.5	
	R@1	R@5	R@1	R@5	R@1	R@5
Zero-shot CLIP [8]	37.83	69.53	17.01	47.63	6.75	25.81
VLG-Net [9]	44.70	73.99	23.91	61.81	17.59	44.30
Moment-DETR [4]	**46.30**	**93.17**	**34.00**	**68.88**	**20.12**	**44.48**

Table 1: **Short-form video setup: 30 seconds.** The table shows the performance of the three selected grounding baseline models in a short-video setup. In this setting, videos are chunked into 30-second, non-overlapping windows. Moment-DETR achieves the best performance in all the metrics for this setup, however, it falls behind zero-shot CLIP and VLG-Net in the long-form video setup (see Section 4 of the main paper).



(a) Baseline models performance in short-form video setup at R@1-IoU=0.5.



(b) Baseline models performance in short-form video setup at R@5-IoU=0.5.

Figure 1: **Performance trend across different short-form video setups.** Figure (a) illustrates the baseline performance in the short-form video setup at R@1 using tIoU=0.5. In contrast, Figure (b) displays the performance at R@5 with tIoU=0.5. The figure also demonstrates how the performance of all baseline models changes as the evaluation window lengthens. Notably, we observe that Moment-DETR is the most affected by long window sizes exceeding 120 seconds.

Figure 1 presents the performance of the three baselines

Model	IoU=0.1					IoU=0.3					IoU=0.5				
	R@1	R@5	R@10	R@50	R@100	R@1	R@5	R@10	R@50	R@100	R@1	R@5	R@10	R@50	R@100
Moment-DETR [4]	12.87	37.35	41.32	42.93	–	5.21	17.23	20.73	21.14	–	2.09	4.47	9.50	9.94	–
VSL-Net [11, 5]	16.52	26.59	32.52	47.47	54.16	10.82	18.87	23.21	33.94	40.40	6.81	13.45	17.09	26.87	31.80
†Moment-DETR [4]	13.47	**39.55**	**43.10**	44.63	–	5.83	18.87	21.14	22.41	–	2.53	8.00	10.02	11.10	–
†VSL-Net [11, 5]	**17.11**	31.85	42.13	**66.99**	**67.20**	**11.25**	**23.95**	**30.25**	**53.99**	**54.20**	**7.20**	**16.37**	**27.55**	**42.80**	**43.30**

Table 2: **Benchmarking of grounding methods on the Ego4D dataset.** We present four models:Moment-DETR [4] and VSL-Net [11, 5] in rows 1 and 2, respectively, without Guidance Model. Rows 3 and 4 demonstrate the impact of our Guidance Model on Moment-DETR's performance and VSL-Net's performance. To denote the utilization of the Guidance Model, we use the symbol †. It's important to note that these results were obtained using the best Guidance Model under a query-dependent setup.

in the five short-form video grounding setups. Moment-DETR [4] has the best performance when grounding moments on 30 second windows and achieves comparable performance to VLG-Net [9] in the other short setups (i.e., 30, 60, 120 seconds). However, the performance gap between Moment-DETR and VLG-Net starts to grow for window lengths greater than 120 seconds. Furthermore, Moment-DETR is better than zero-shot CLIP in all short-form video setups except for the one with a 600 second window length, where Moment-DETR exhibits the worst performance. In any case, we must note that all models exhibit a significant drop in performance when attempting to process longer videos.

Takeaways: (i) Moment-DETR works well with short-form videos but fails for videos longer than 120 seconds, **(ii)** current state-of-the-art models cannot address the video grounding task in the long-form video setting, and **(iii)** using the Guidance Model along with Moment-DETR can boost the latter's performance allowing it to compete with state-of-the-art models on the MAD Dataset.

B. Ego4D Results

We also evaluated the effectiveness of the Guidance Model on the Ego4D [2] validation dataset. In Table 2, we present the performance of Moment-DETR [4] and VSL-Net [11, 5] and the boost in performance for the same baseline using the best configuration of our Guidance Model. By reporting Recall@K with IoU=θ, for $K\in\{1, 5, 10, 50, 100\}$ and $\theta\in\{0.1, 0.3, 0.5\}$, instead of mR@$K$, we are able to provide more detailed results. This allows us to obtain a more fine-grained understanding of the performance of the model and its ability to ground moments accurately by using natural language queries.

Generally speaking, VSL-Net with the Guidance model performs the best in most of the metrics. For instance, R@1-IoU=0.5 increases from 6.81 to 7.20, for R@5-IoU=0.3 increases from 18.87 to 23.95 and R@10-IoU=0.1 increases from 32.52 to 42.13. In contrast, Moment-DETR demonstrated notable performance enhancements, illustrated by a significant rise in R@5-IoU scores. Specifically, at the IoU threshold of 0.1, the metric increased from 26.59 to an impressive 39.55, while at the more stringent threshold of 0.5, it improved from 4.47 to 8.00.

Takeaway: The efficacy of the Guidance Model in enhancing the performance of the base grounding model demonstrates its versatility across datasets and its compatibility with a variety of vision-language grounding models.

C. Ablation Study

We focused our ablation studies on two key factors: **(i)** selecting the modalities (visual and/or audio), **(ii)** comparing query-agnostic versus query-dependent guidance performance using the MAD [8] dataset. Additionally, we extensively investigated actionless moments, which distinguishes our implementation from the temporal proposals method [1, 12, 10, 3, 6].

C.1. Modality Fusion in Guidance Model

As shown in Table 5, the Guidance Model can improve the baseline results, especially when combining all modalities (audio, video, and text). Instead of using mR@K as reported in Table 3 of the main paper, here we report Recall@K with IoU=θ, for $K\in\{1, 5, 10, 50, 100\}$ and $\theta\in\{0.1, 0.3, 0.5\}$, enabling us to carry out a more fine-grained analysis. For example, when leveraging all three modalities, the performance of Zero-shot CLIP grows from 6.65% to 9.05% for R@1-IoU=0.1 and 5.31% to 7.14% for R@5-IoU=0.5. VLG-Net [9] also achieves better performance when using either two or three modalities. For example, R@1-IoU=0.3 grows from 2.56% to 4.01% under the best configuration. On the other hand, the grounding model that is the most benefited by the Guidance Model is Moment-DETR [4]. As discussed in the previous section, this model performs poorly on the grounding task with long-form videos; however, by combining it with the Guidance Model, it can reach a performance that is competitive with state-of-the-art approaches. For instance, Moment-DETR [4] beats zero-shot CLIP [8] in the R@1-IoU=0.5 metric for the best configuration: the former achieves

Model	Query	IoU=0.1					IoU=0.3					IoU=0.5				
		R@1	R@5	R@10	R@50	R@100	R@1	R@5	R@10	R@50	R@100	R@1	R@5	R@10	R@50	R@100
Zero-shot CLIP [8]	✗	6.65	14.80	19.79	36.30	45.59	3.19	9.88	13.84	27.47	35.33	1.35	5.31	8.03	18.06	23.93
	AG	6.84	15.19	20.28	37.88	**45.80**	3.89	9.93	15.22	27.80	36.60	1.41	5.49	9.29	17.44	24.12
	DE	**9.05**	**18.00**	**22.76**	**37.24**	44.37	**4.52**	**12.60**	**16.83**	**30.13**	**37.02**	**2.01**	**7.14**	**10.65**	**21.75**	**27.76**
VLG-Net [9]	✗	3.43	11.32	17.19	37.80	48.63	2.56	8.92	13.92	32.38	42.34	1.49	5.54	9.01	22.93	31.34
	AG	3.89	14.31	18.69	39.37	49.87	2.82	9.30	15.24	35.22	44.19	1.70	7.12	9.87	24.00	34.16
	DE	**5.37**	**15.52**	**22.37**	**42.58**	**52.72**	**4.01**	**12.56**	**18.58**	**37.33**	**46.44**	**2.31**	**8.16**	**12.57**	**28.14**	**36.46**
Moment-DETR [4]	✗	0.28	1.44	2.62	10.41	18.21	0.20	1.07	1.87	7.61	13.18	0.12	0.62	1.05	4.47	7.93
	AG	0.65	2.11	4.54	18.44	32.34	0.59	1.78	2.83	14.22	22.97	0.36	1.02	1.45	8.32	12.54
	DE	**4.84**	**15.34**	**22.80**	**46.01**	**57.06**	**3.69**	**11.95**	**17.99**	**37.10**	**46.19**	**2.17**	**7.25**	**10.93**	**22.62**	**28.09**

Table 3: **Describable windows (full metrics).** In this table, we report Recall@K on the validation partition of MAD for the baseline models under three settings: without any guidance (rows 1, 4, and 7), with query agnostic guidance (AG), and with query dependent guidance (DE). All three baselines benefit the most from using query-dependent guidance, and modest performance improvement by query-agnostic setup. Our findings suggest that the query-dependent approach yields superior performance despite its high computational cost. However, adopting a query-agnostic setup can also enhance performance at a lower computational cost.

2.17%, while the latter gets 2.01%. Moreover, the R@K gap between the two models grows significantly with values of K greater than 10.

Takeaway: We observe that using audio alone, without visual input, results in only modest performance improvements for Zero-shot CLIP, VLG-Net, and Moment-DETR. On the other hand, incorporating visual cues can enhance the performance of all baselines. Nevertheless, the best overall performance is achieved by combining audio, visual and text cues.

C.2. Describable Windows

Although the first ablation study suggests that combining audio and video inputs as a representation is powerful, the current Guidance Model used is query-dependent and relies on textual queries to identify irrelevant windows. Our method could be made more efficient by identifying windows that are non-describable regardless of the input query, allowing the Guidance Model to process the video/audio streams only once. To investigate an efficient approach, we devise a query-agnostic Guidance Model that does not process any textual query as part of its input.

Model	IoU=0.1			IoU=0.3			IoU=0.5		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Zero-shot CLIP [8]	5.43	11.87	15.98	2.69	7.88	11.06	1.15	4.54	6.86
VLG-Net [9]	2.15	7.65	12.00	1.50	5.84	9.56	0.95	3.83	6.43
Moment-DETR [4]	0.24	0.97	1.78	0.13	0.76	0.85	0.08	0.21	0.09
†Zero-shot CLIP [8]	**7.39**	**14.90**	**19.09**	**3.76**	**10.32**	**13.95**	1.77	5.93	9.00
†VLG-Net [9]	3.31	10.17	16.06	2.21	7.83	13.09	1.45	**6.20**	**10.24**
†Moment-DETR [4]	3.71	11.37	17.94	2.84	8.53	13.62	**2.04**	5.41	8.98

Table 4: **Describable windows beyond actions.** We compare the performance of our baseline models with the guidance module (†) and without guidance on actionless queries only. For this purpose, we extract actionless queries from MAD [8] test set and evaluate using mR@K for $K \in \{1, 5, 10\}$. The results showcase that our guidance method is not solely learning action-based concepts.

In Table 3, we report the full results from which we derived Table 4 of the main paper. Using audio-visual cues alone (that is, a query-agnostic setup) already leads to improvements for VLG-Net [9] and Moment-DETR [4]. On the other hand, employing query-dependent guidance plays a vital role in boosting the performance of all three baseline models, as it helps reduce the search space for the video grounding task. For instance, VLG-Net [9] goes from 11.32% to 14.31% R@5-IoU=0.1 when employing query agnostic guidance, and from 11.32% to 15.52% R@5-IoU=0.1 when employing query dependent guidance. Similarly, the performance of Moment-DETR [4] for most metrics also grows: it goes from 2.62% to 4.54% R@10-IoU=0.1 with query agnostic guidance, and from 2.62% to 22.80% with query dependent guidance. Lastly, we observe that zero-shot CLIP only displays a noticeable performance boost when leveraging query-dependent guidance. For example, R@5-IoU=0.3 increases from 9.88% to 12.60% in the query-dependent setting, and from 9.88% to 9.93% in the query-agnostic setting.

Takeaway: Our findings suggest that the query-dependent setup delivers better performance, but at the expense of increased computational cost as the number of queries increases. Conversely, the query-agnostic setup is computationally efficient since the Guidance Model processes the video and audio only once, making it ideal for real-time or low-resource scenarios.

C.3. Describable Windows beyond actions.

We proposed a new evaluation setup that highlights the differences between our guidance model and proposal mod-

Model	Modalities		IoU=0.1					IoU=0.3					IoU=0.5				
	Audio	Visual	R@1	R@5	R@10	R@50	R@100	R@1	R@5	R@10	R@50	R@100	R@1	R@5	R@10	R@50	R@100
Zero-shot CLIP [8]	✗	✗	6.65	14.80	19.79	36.30	**45.59**	3.19	9.88	13.84	27.47	35.33	1.35	5.31	8.03	18.06	23.93
	✓	✗	6.75	16.08	21.67	36.51	45.40	3.50	10.85	15.72	28.45	35.49	1.63	6.23	9.28	19.31	25.07
	✗	✓	8.33	17.42	22.16	36.91	44.10	4.21	11.93	16.16	29.51	36.47	1.89	6.69	10.01	21.23	27.32
	✓	✓	**9.05**	**18.00**	**22.76**	**37.24**	44.37	**4.52**	**12.60**	**16.83**	**30.13**	**37.02**	**2.01**	**7.14**	**10.65**	**21.75**	**27.76**
VLG-Net [9]	✗	✗	3.43	11.32	17.19	37.80	48.63	2.56	8.92	13.92	32.38	42.34	1.49	5.54	9.01	22.93	31.34
	✓	✗	4.43	13.21	19.20	38.70	48.93	3.34	10.67	15.80	33.20	42.18	2.03	6.81	10.28	23.96	31.70
	✗	✓	4.85	14.56	21.15	41.73	51.70	3.60	11.55	17.45	36.43	45.55	2.04	7.30	11.62	26.86	35.18
	✓	✓	**5.37**	**15.52**	**22.37**	**42.58**	**52.72**	**4.01**	**12.56**	**18.58**	**37.33**	**46.44**	**2.31**	**8.16**	**12.57**	**28.14**	**36.46**
Moment-DETR [4]	✗	✗	0.28	1.44	2.62	10.41	18.21	0.20	1.07	1.87	7.61	13.18	0.12	0.62	1.05	4.47	7.93
	✓	✗	1.41	5.72	9.67	27.59	39.10	1.05	4.31	7.32	21.66	31.02	0.62	2.56	4.35	12.91	18.49
	✗	✓	4.24	14.41	21.74	44.77	56.11	3.13	11.07	16.84	35.99	45.26	1.86	6.58	9.98	21.64	27.28
	✓	✓	**4.84**	**15.34**	**22.80**	**46.01**	**57.06**	**3.69**	**11.95**	**17.99**	**37.10**	**46.19**	**2.17**	**7.25**	**10.93**	**22.62**	**28.09**

Table 5: **Modality comparison for the Guidance Model (full metrics)**. This table contains an ablation study on the modalities to be used in the Guidance Model. In the first row of each box (rows 1, 5, and 9), we report the baseline performances without score fusion. The Guidance Model uses three different combinations of modalities: (i) audio and text; (ii) video and text; and (iii) audio, video, and text. We evaluate all model configurations on the validation partition of the MAD dataset, using Recall@K with IoU=θ for $K \in \{1, 5, 10, 50, 100\}$ and $\theta \in \{0.1, 0.3, 0.5\}$. Generally speaking, the Guidance Model can boost the baseline results under all configurations, and it achieves the best overall performance when combining all modalities.

Method	Inference Time	Avg. Temp. Windows per movie	# Queries×movies	Number of Params (M)
Grounding: VLG-Net	18 Hours	545	72044×112	7.5
Grounding: Moment-DETR	4.2 Hours	545	72044×112	4.8
Guidance: Query-agnostic	∼2 Minutes	1091	$- \times 112$	4.8
Guidance: Query-dependent	8.3 Hours	1091	72044×112	5.0

Table 6: **Inference time on the MAD [8] dataset.** Query-agnostic needs a forward pass per movie, while dependent is proportional to the no. of queries/temporal windows. We don't include CLIP [7, 8] as our approach leverages the model's own pre-computed features making the comparison unfair.

els. Specifically, we consider MAD [8] queries that do not contain verbs and therefore do not refer to actions. We acknowledge that interpreting this type of query is a challenging task, distinguishing our approach from other computer vision methods like temporal proposals [1, 12, 10, 3, 6], which do not deal with moments lacking any action element. Thus, to confirm our hypothesis, we present in Table 4 the performance improvement that our Guidance Model offers for actionless queries. The Guidance Model demonstrated a boost in performance in all of the baselines. For example, R@1-IoU0.1 for Zero-shot CLIP improved from 5.43% to 7.39%, VLG-Net at R@10-IoU0.3 improved from 9.56% to 13.09%, and Moment-DETR from 0.21% to 5.41% in R@5-IoU0.5.

Takeaway: The higher performance of the Guidance Model for queries without actions is clear evidence that it is not solely focused on actions, but can also handle queries that involve describing the environment and its characteristics (adjectives, nouns). The above statement sets our approach apart from proposal-based methods, since proposals methods [1, 12, 10, 3, 6] focus only on actions.

D. Inference Time.

Table 6 displays inference time for the baseline video grounding models and the Guidance Models. The query-agnostic configuration is intended for situations with restricted computational capabilities. It only requires one sliding-window forward pass per movie, covering 112 movies in total. On the other hand, the query-dependent configuration requires more computational power, as the number of forward passes is directly linked to the total number of queries (72,000) and temporal windows.

A potential direction for future exploration in this paper involves improving inference time efficiency by strategically employing the Guidance Model to pre-filter windows before grounding. This concept revolves around the careful prioritization of recall or precision, depending on the specific needs of the given application. By using the abilities of the Guidance Model, we can choose to handle windows that are most important, thereby reducing the total time for inference while maintaining desired performance standards. This potential also offers an opportunity to tailor the processing pipeline for different applications' unique traits, finding a balanced compromise between computational efficiency and outcome accuracy.

References

[1] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. Sst: Single-stream temporal action proposals. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6373–6382, 2017.

[2] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Q. Chavis, Antonino Furnari, Rohit Girdhar, Jack-

son Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh K. Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Z. Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina González, James M. Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Yu Heng Khoo, Jáchym Kolár, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran K. Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbeláez, David J. Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard A. Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18973–18990, 2021.

[3] Tingting Han, Sicheng Zhao, Xiaoshuai Sun, and Jun Yu. Modeling long-term video semantic distribution for temporal action proposal generation. *Neurocomputing*, 490:217–225, 2022.

[4] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 11846–11858. Curran Associates, Inc., 2021.

[5] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z XU, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022.

[6] Zhiwu Qing, Haisheng Su, Weihao Gan, Dongliang Wang, Wei Wu, Xiang Wang, Yu Qiao, Junjie Yan, Changxin Gao, and Nong Sang. Temporal context aggregation network for temporal action proposal refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 485–494, 2021.

[7] 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. *arXiv preprint arXiv:2103.00020*, 2021.

[8] Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5026–5035, June 2022.

[9] Mattia Soldan, Mengmeng Xu, Sisi Qu, Jesper Tegner, and Bernard Ghanem. Vlg-net: Video-language graph matching network for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3224–3234, 2021.

[10] Haosen Yang, Wenhao Wu, Lining Wang, Sheng Jin, Boyang Xia, Hongxun Yao, and Hujie Huang. Temporal action proposal generation with background constraint. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 3054–3062. AAAI Press, 2022.

[11] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6543–6554, Online, July 2020. Association for Computational Linguistics.

[12] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*, 2017.