MAD: A Scalable Dataset for Language Grounding in Videos from Movie Audio Descriptions

Mattia Soldan[1], Alejandro Pardo[1], Juan León Alcázar[1], Fabian Caba Heilbron[2], Chen Zhao[1], Silvio Giancola[1], Bernard Ghanem[1]

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

{mattia.soldan, alejandro.pardo, juancarlo.alcazar, chen.zhao, silvio.giancola, bernard.ghanem}@kaust.edu.sa caba@adobe.com

Abstract

We present MAD (Movie Audio Descriptions), a novel benchmark for the video-language grounding task. We depart from the paradigm of augmenting existing video datasets with text annotations and focus on crawling and aligning available audio descriptions of mainstream movies. MAD contains over $384,000$ *natural language sentences grounded in over* $1,200$ *hours of video and exhibits a significant reduction in the currently diagnosed biases for video-language grounding datasets. MAD's collection strategy enables a novel and more challenging version of video-language grounding, where short temporal moments (typically seconds long) must be accurately grounded in diverse long-form videos that can last up to three hours. Find out more at* https://github.com/Soldelli/MAD.

1. Introduction

The natural language grounding task [1, 4] has gained significant momentum in the computer vision community due to the multiple potential real-world applications [2, 3, 11, 12, 14]. The importance of solving this task has resulted in novel approaches and large-scale deep-learning architectures that steadily push state-of-the-art performance. Despite those advances, recent works [7, 15, 17] have diagnosed hidden biases in the most common video-language grounding datasets. In detail, the temporal anchors for the language are temporally biased in time, leading to methods overfitting to temporal priors, thus limiting their generalization capabilities [6, 7] (Fig. 2). The community has tried to circumvent these limitations by either proposing new metrics [15, 16] or debiasing strategies [17, 18]. However, it is still unclear if existing grounding datasets [1, 4, 5, 9] provide the right setup to evaluate progress in this relevant task.

In this work, we present a novel large-scale dataset called MAD (Movie Audio Descriptions). MAD builds atop (and includes part of) the LSMDC dataset [10], which is a pioneering work in leveraging audio descriptions to enable the investigation of a closely related task: text-to-video retrieval. Similar to LSMDC, we depart from the standard annotation pipelines that rely on crowd-sourced annotation platforms. Instead, MAD relies on audio descriptions professionally created to make movies accessible to visually-impaired audiences. These descriptions embody a rich narrative describing the most relevant visual information, adopting a highly descriptive and diverse language.

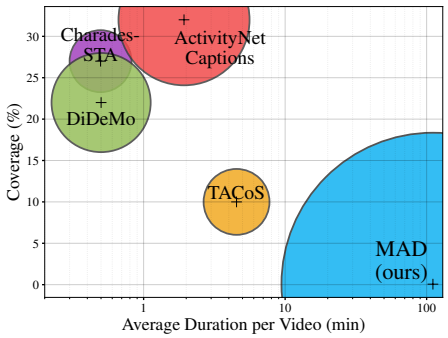

Figure 1. **Comparison of video-language grounding datasets.** The circle size measures the language vocabulary diversity. The videos in MAD are orders of magnitude longer in duration than previous datasets (∼110min), annotated with natural, highly descriptive, language grounding (>60K unique words) with very low coverage in video (∼4.1s). Coverage is defined as the average % duration of moments with respect to the total video duration.

Fig. 1 shows that the current datasets comprise short videos containing single structured scenes and language descriptions that cover most of the video. Conversely, MAD contains long-form videos that, on average, span over 110 minutes, as well as grounded annotations covering short time segments, which are uniformly distributed in the video, and maintain the largest diversity in vocabulary.

The unique configuration of the MAD dataset introduces exciting challenges. *(i)* The video grounding task is now mapped into the unexplored domain of long-form videos.

Dataset	Videos			Language Queries						
	Total Duration	Duration / Video	Duration / Moment	Total Queries	# Words / Query	Total Tokens	Vocabulary			
							Adj.	Nouns	Verbs	Total
TACoS [9]	10.1 h	4.78 min	27.9 s	18.2K	10.5	0.2M	0.2K	0.9K	0.6K	2.3K
Charades-STA [4]	57.1 h	0.50 min	8.1 s	16.1K	7.2	0.1M	0.1K	0.6K	0.4K	1.3K
DiDeMo [1]	88.7 h	0.50 min	6.5 s	41.2K	8.0	0.3M	0.6K	4.1K	1.9K	7.5K
ANet-Captions [5]	487.6 h	1.96 min	37.1 s	72.0K	14.8	1.0M	1.1K	7.4K	3.7K	15.4K
MAD (Ours)	1207.3 h	110.77 min	4.1 s	384.6K	12.7	4.9M	5.3K	35.5K	13.1K	61.4K

Table 1. **Statistics of video-language grounding datasets.** We report relevant statistics to compare our MAD dataset against other video grounding benchmarks. MAD provides the largest dataset with 1207hrs of video and 384.6K language queries, the longest form of video (avg. 110.77min), the most diverse language vocabulary with 61.4K unique words, and the shortest moment for grounding (avg. 4.1s).

(ii) Longer videos will make the localization problem far more challenging. *(iii)* The longer sequences emphasize the necessity for efficient methods in inference and training, mandatory for real-world applications.

Contributions (1) We propose MAD, a novel large-scale dataset for video-language grounding. **(2)** We design a scalable data collection pipeline that automatically extracts annotations. **(3)** We provide a empirical study that highlights the benefits of our large-scale MAD dataset on the video-language grounding task.

2. Collecting the MAD Dataset

We follow independent strategies for the training and testing set. For the former, we aim at automatically collecting a large set of annotations. For the latter, we re-purpose the manually refined annotations from the LSMDC dataset.

2.1. MAD Training set

Data Crawling Not every commercially available movie is released with audio descriptions. However, we can obtain these audio descriptions from 3rd party creators. In particular, we crawl our audio descriptions from a large open-source and online repository (www.audiovault.net).

Alignment One problem is that the audio descriptions can be misaligned with the original movie. Since the audio description track also contains the movie's original audio, we can circumvent this misalignment by maximizing the cross-correlation between overlapping segments of the original audio track and the audio description track, obtaining in this way the time delay τ_{delay}. We discard the movies that do not reach a consensus on the delay estimation.

Audio Transcriptions We transcribe the audio description file using Microsoft's Azure Speech-to-Text service. To remove the actors' speech from the transcription, we resort to the movie's subtitles and use their timestamps as a surrogate for Voice Activity Detection (VAD). We remove from the Speech-to-Text output every sentence overlapping with the VAD temporal locations, obtaining our target annotations.

From LSMDC to MAD Val/Test Since the annotations in training are automatically generated, we decided to minimize the noise in the validation and test splits. Hence, we avoid the automatic collection of data for these sets and resort to the data made available by the LSMDC dataset [10]. LSMDC manually refined the grammar and temporal boundaries of sentences. As a consequence, these annotations have clean language and precise temporal boundaries. We reformat a subset of the LSMDC data, adapt it for the video grounding task, and cast it as MAD's validation and test sets. LSMDC data for retrieval is made available only as video chunks, not full movies. To create data suitable for long-form video language grounding, we collect 162 out of the 182 videos in LSMDC and their respective audio descriptions. To align a video chunk from LSMDC with our full-length movies, we follow a similar procedure as the one described for the audio alignment, but using visual information.

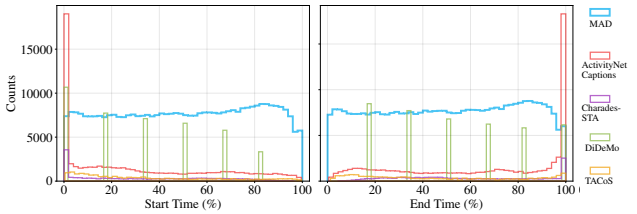

Figure 2. **Histograms of normalized (by video length) moment start/end in video-language grounding datasets.** We notice severe biases in ActivityNet-Captions and Charades-STA, with high peaks at the beginning and end of the videos. Conversely, MAD does not show any particular preferred start/end temporal location.

2.2. MAD Dataset Analysis

Table 1 summarizes the most notable aspects of grounding datasets and compares MAD with legacy datasets. MAD is the largest dataset in video hours and number of sentences. The training, validation, and test sets consist of 488, 50, and 112 movies with 280.5K, 32.1K, and 72.0K queries, respectively. Although other datasets have a larger number of clips, MAD's videos are full movies which last 2 hours on average. In comparison, the average clip from other datasets spans just a few minutes. Moreover, MAD contains the largest set of adjectives, nouns, and verbs among all available benchmarks. In almost every case, it is an order of magnitude larger. Overall, MAD contains 61.4K unique words, almost 4 times more than the 15.4K of ActivityNet-Captions [5] (the highest among the related benchmarks). The average length per sentence

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
Oracle	100.00	–	–	–	–	100.00	–	–	–	–	99.99	–	–	–	–
Random Chance	0.09	0.44	0.88	4.33	8.47	0.04	0.19	0.39	1.92	3.80	0.01	0.07	0.14	0.71	1.40
CLIP [8]	**6.57**	**15.05**	**20.26**	37.92	47.73	**3.13**	**9.85**	14.13	28.71	36.98	1.39	5.44	8.38	18.80	24.99
VLG-Net [13]	3.64	11.66	17.89	**39.78**	**51.24**	2.76	9.31	**14.65**	**34.27**	**44.87**	**1.65**	**5.99**	**9.77**	**24.93**	**33.95**

Table 2. **Benchmarking of grounding baselines on the MAD dataset.** We report the performance of four baselines: *Oracle*, *Random Chance*, *CLIP*, *VLG-Net*, on the test split. For all experiments, we adopt the same proposal scheme as in VLG-Net [13]. For *CLIP* and *VLG-Net* we use CLIP [8] features for video frames (extracted at 5 FPS) and language embeddings.

is 12.7 words, which is similar to the other datasets. Finally, note how the average moment duration (4.1 seconds) is shorter with respect to previous dataset, yielding a very low coverage and making the task challenging.

3. Experiments

Task Given an untrimmed video and a language query, the video-language grounding task aims to localize a temporal moment (τ_s, τ_e) in the video that matches the query [1, 4].
Metric Following the grounding literature [1, 4], we adopt Recall@K for IoU=θ (R@K-IoU=θ). Given the long-form nature of our videos and the large amount of proposals, we chose $K \in \{1, 5, 10, 50, 100\}$ and $\theta \in \{0.1, 0.3, 0.5\}$.
Baselines We adopt four grounding strategies, namely: *Oracle*, *Random Chance*, *CLIP* [8], and *VLG-Net* [13]. *Oracle* measures the upper bound on the performance by choosing the proposal with the highest IoU with the ground-truth annotation. *Random Chance* chooses a random proposal with uniform probability. *CLIP* [8] is used in a zero-shot setting. The frame-level features for each proposal are combined using mean pooling, then we score each proposal using cosine similarity between the visual and the text features. Finally, we adopt VLG-Net [13] as a representative, state-of-the-art method for the grounding task.
Grounding Performance on MAD As shown in Table 2, the *Oracle* evaluation achieves a perfect score across all metrics except for IoU=0.5. Only a negligible portion of the annotated moments cannot be correctly retrieved at a high IoU (0.5), this result showcases the suitability of the proposal scheme. The low performance of the *Random Chance* baseline reflects the difficulty of the task, given the vast pool of proposals extracted over a single video. For the least strict metric (R@100-IoU=0.1), this baseline only achieves 8.47%, while CLIP and VLG-Net baselines are close to 50%, a $\sim 6\times$ relative improvement. The CLIP [8] baseline is pre-trained for the task of text-to-image retrieval, and we do not fine-tune this model on the MAD dataset. Nevertheless, when evaluated with a zero-shot setting, it results in a strong baseline achieving the best R@K for the least strict IoU=0.1 at $K=\{1, 5, 10\}$. Conversely, VLG-Net is trained for the task at hand, but achieves comparable-to-better performance only when a strict IoU (IoU=0.5) is considered. We believe the shortcomings of VLG-Net are due to two factors. *(i)* This architecture was developed to ground sentences in short videos, where the entire frame-set can be compared against a sentence in a single forward pass. Thus, it struggles in the long-form setup where we compare the sentence against segments of the movie and then aggregate the predictions. *(ii)* VLG-Net training procedure defines low IoU moments as negatives, thus favoring high performance only for higher IoUs.

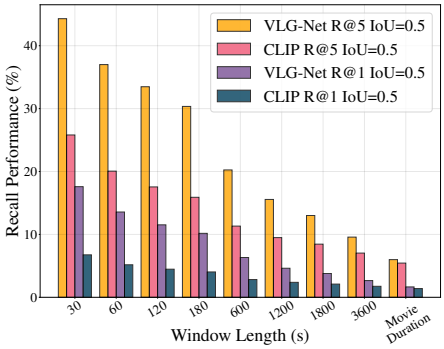

Figure 3. **Performance trend across different windows lengths**. Performance of deep learning based methods steadily decrease as the window evaluation window length increases.

The Challenges of Long-form Video Grounding We investigate how the performance changes when the evaluation is constrained over segments of the movie, and vary the size of the segment. To this end, we split each video into non-overlapping windows (short videos), and assign the annotations to the short-video with the highest temporal overlap. Figure 3 showcases the performance trend for the metrics R@$\{1, 5\}$-IoU=0.5, when the window length is changed from a small value (30 seconds) to the entire movie duration (average duration is 2hrs). The graph displays how the performance steadily drops as the window length increases, showing the challenging setup of long-form grounding enabled by MAD.

4. Conclusion

The paper presents a new video grounding benchmark called MAD, which builds on high-quality audio descriptions in movies. MAD alleviates the shortcomings of previous grounding datasets. Our automatic annotation pipeline allowed us to collect the largest grounding dataset to date. The experimental section provides baselines for the task solution and highlights the challenging nature of the long-form grounding task introduced by MAD.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing Moments in Video With Natural Language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2, 3

[2] Andrew E Budson and Bruce H Price. Memory dysfunction. *New England Journal of Medicine*, 352(7):692–699, 2005. 1

[3] Victor Escorcia, Mattia Soldan, Josef Sivic, Bernard Ghanem, and Bryan Russell. Temporal localization of moments in video collections with natural language. *arXiv preprint arXiv:1907.12763*, 2019. 1

[4] Gao Jiyang, Sun Chen, Yang Zhenheng, Nevatia, Ram. TALL: Temporal Activity Localization via Language Query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2, 3

[5] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-Captioning Events in Videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2

[6] Xiaohan Lan, Yitian Yuan, Xin Wang, Zhi Wang, and Wenwu Zhu. A survey on temporal sentence grounding in videos. *arXiv preprint arXiv:2109.08039*, 2021. 1

[7] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. Uncovering hidden challenges in query-based video moment retrieval. In *The British Machine Vision Conference (BMVC)*, 2020. 1

[8] 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. 3

[9] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding Action Descriptions in Videos. *Transactions of the Association for Computational Linguistics (ACL)*, 2013. 1, 2

[10] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 123(1):94–120, 2017. 1, 2

[11] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, IEEE International Conference on*, volume 3, pages 1470–1470. IEEE Computer Society, 2003. 1

[12] Cees Snoek, Kvd Sande, OD Rooij, Bouke Huurnink, J Uijlings, M van Liempt, M Bugalhoy, I Trancosoy, F Yan, M Tahir, et al. The mediamill trecvid 2009 semantic video search engine. In *TRECVID workshop*. University of Surrey, 2009. 1

[13] 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. 3

[14] Takumi Toyama and Daniel Sonntag. Towards episodic memory support for dementia patients by recognizing objects, faces and text in eye gaze. In *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*, pages 316–323. Springer, 2015. 1

[15] Yitian Yuan, Xiaohan Lan, Long Chen, Wei Liu, Xin Wang, and Wenwu Zhu. A closer look at temporal sentence grounding in videos: Datasets and metrics. *CoRR*, abs/2101.09028, 2021. 1

[16] Yitian Yuan, Xiaohan Lan, Long Chen, Wei Liu, Xin Wang, and Wenwu Zhu. A closer look at temporal sentence grounding in videos: Datasets and metrics. *arXiv preprint arXiv:2101.09028*, 2021. 1

[17] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Towards debiasing temporal sentence grounding in video. *arXiv preprint arXiv:2111.04321*, 2021. 1

[18] Hao Zhou, Chongyang Zhang, Yan Luo, Yanjun Chen, and Chuanping Hu. Embracing uncertainty: Decoupling and de-bias for robust temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8445–8454, 2021. 1