

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

Mattia Soldan¹, Alejandro Pardo¹, Juan León Alcázar¹, Fabian Caba Heilbron², Chen Zhao¹, Silvio Giancola¹, Bernard Ghanem¹

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

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

Abstract

The recent and increasing interest in video-language research has driven the development of large-scale datasets that enable data-intensive machine learning techniques. In comparison, limited effort has been made at assessing the fitness of these datasets for the video-language grounding task. Recent works have begun to discover significant limitations in these datasets, suggesting that state-of-the-art techniques commonly overfit to hidden dataset biases. In this work, we present MAD (Movie Audio Descriptions), a novel benchmark that departs from the paradigm of augmenting existing video datasets with text annotations and focuses 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.

1. Introduction

Imagine you want to find the moment in time, in a movie, when your favorite actress is eating a Gumbo dish in a New Orleans restaurant. You could do so by manually scrubbing the film to ground the moment. However, such a process is tedious and labor-intensive. This task is known as natural language grounding [1, 3], and has gained significant momentum in the computer vision community. Beyond smart browsing of movies, the interest in this task stems from multiple real-world applications ranging from smart video search [23, 24] to helping patients with memory dysfunction [2, 28]. The importance of solving this task has resulted in novel approaches and large-scale deep-learning architec-

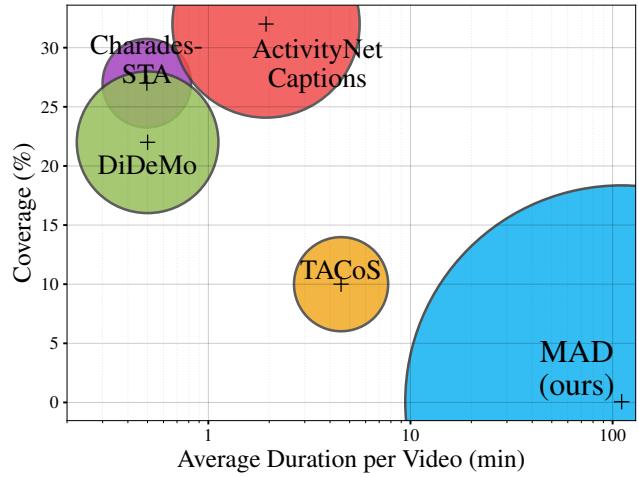


Figure 1. Comparison of video-language grounding datasets. The circle size measures the language vocabulary diversity. The videos in MAD are orders of magnitude longer in duration than previous datasets ($\sim 110\text{min}$), annotated with natural, highly descriptive, language grounding ($> 60\text{K}$ unique words) with very low coverage in video ($\sim 4.1\text{s}$).

tures that steadily push state-of-the-art performance.

Despite those advances, recent works [16, 30, 33] have diagnosed hidden biases in the most common video-language grounding datasets. Otani *et al.* [16] have highlighted that several grounding methods only learn location priors, to the point of disregarding the visual information and only using language cues for predictions. While recent methods [30, 33] have tried to circumvent these limitations by either proposing new metrics [31] or debiasing strategies [33, 36], it is still unclear if existing grounding datasets [1, 3, 6, 18] provide the right setup to evaluate progress in this important task. This is partly because datasets used by video-language grounding models were not originally collected to solve this task. These datasets provide valuable video-language pairs for captioning or re-

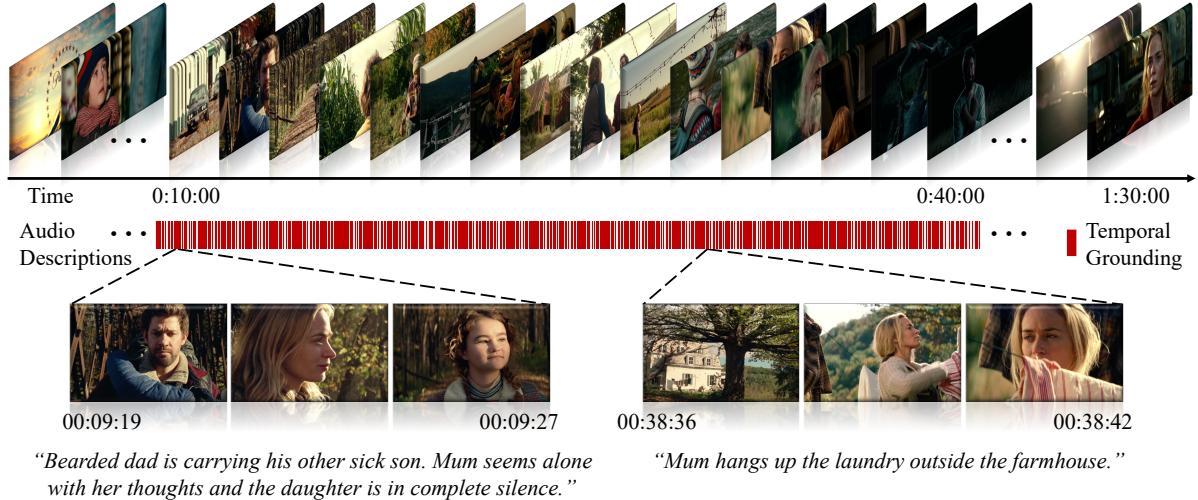


Figure 2. Example from our MAD dataset. We select the movie “A quiet place” as representative for our dataset. As shown in the figure, the movie contains a large amount of densely distributed temporally grounded sentences. The collected annotations can be very descriptive, mentioning people, actions, locations, and other additional information. Note that as per the movies plot, the characters are silent for the vast majority of the movie, rendering audio description essential for visually-impaired audience.

trieval, but the grounding task requires high-quality (and dense) temporal localization of the language. Figure 1 shows that the current datasets comprise relatively short videos, contain single structured scenes, and language descriptions that cover most of the video. Furthermore, the temporal anchors for the language are temporally biased in time (refer to Figure 3), leading to methods not learning from any visual features and eventually overfitting to temporal priors for specific actions, thus limiting their generalization capabilities [7, 16].

In this work, we address these limitations with a novel large-scale dataset, called MAD (Movie Audio Descriptions). It comprises long-form videos, which bring new challenges to the video-language grounding task. We depart from the standard annotation pipelines that rely on crowd-sourced annotation platforms. Instead, we adopt a scalable data collection strategy that leverages professional, grounded audio descriptions of movies for visually impaired audiences. Our data collection approach consists of transcribing the audio description track of a movie, and automatically detecting and removing sentences associated with the actor’s speech, yielding an authentic “untrimmed video” setup where the highly descriptive sentences are grounded in long-form videos. Figures 1 and 3 illustrate the uniqueness of our dataset with respect to the current alternatives. As showcased in Figure 2, MAD contains 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. Video grounding in MAD requires a finer understanding of the video, since the average coverage of the sentences is much smaller than in current datasets.

The unique configuration of the MAD dataset introduces exciting challenges. First, the video grounding task is now mapped into the domain of long videos, preventing current methods to learn temporal location priors, instead requiring a more nuanced understanding of the video and language modalities. Second, having longer videos means producing a larger number of segment proposals, which will make the localization problem far more challenging. Last, these longer sequences emphasize the necessity for efficient methods in inference and training, mandatory for real-world applications such as long live streaming videos or moment retrieval in large video collections.

Contributions. In summary, they are threefold. **(1)** We propose Movie Audio Description (MAD), a novel large-scale dataset for video-language grounding, containing more than 384K natural language sentences anchored on more than 1.2K hours of video. **(2)** We design a scalable data collection pipeline that automatically extracts highly valuable video-language grounding annotations, leveraging speech-to-text translation on professionally generated audio descriptions. **(3)** We provide a comprehensive empirical study that highlights the benefits of our large-scale MAD dataset on video-language grounding as a benchmark, pointing out the difficulties faced by current video-language grounding baselines in long-form videos. To foster future research, we will publicly release the code to download our dataset and reproduce our baselines.

2. Related work

Video Grounding Benchmarks. Most of the current video grounding datasets were previously collected and tailored

for other computer vision tasks (*e.g.* Temporal Activity Localization [4, 18]) and purposes (*e.g.* Human Action Recognition), then annotated for the video grounding task. This adaptation limits the diversity of the video corpus towards a specific set of actions and objects, and its corresponding natural language sentences to specific sets of verbs and nouns [3, 6, 16, 19]. Currently, ActivityNet-Captions [6] and Charades-STA [3] are the most commonly used benchmarks for the task of video grounding. Both datasets have been collected atop pre-existing video datasets (ActivityNet [4] and Charades [22]) and have been diagnosed with severe biases by Otani *et al.* [16]. These findings show that the annotations contain distinct biases where language tokens are often coupled with specific temporal locations. Moreover, strong priors also affect the temporal endpoints, with a large portion of the annotations spanning the entire video. As a consequence, current methods seem to mainly rely on such biases to make predictions, often disregarding the visual input altogether. In comparison, the unique setup of MAD prevents these drawbacks as keywords are not associated with particular temporal regions, and annotation timestamps are much shorter than the video’s duration. Moreover, differently from Charades-STA, MAD defines an official validation set for hyper-parameter tuning.

Unlike Charades-STA and ActivityNet-Captions, TACoS [18] has not been diagnosed with annotation biases. However, its video corpus is small and limited to cooking actions recorded in a static-camera setting. Conversely, spanning over 22 genres across 90 years of cinema history, MAD covers a broad domain of actions, locations, and scenes. Moreover, MAD inherits a diverse set of visual and linguistic content from the broad movie genres, ranging from fiction to everyday life.

Furthermore, DiDeMo [1] was annotated atop Flickr videos with a discrete annotation scheme (*i.e.* in chunks of 5 seconds) for a maximum of 30 seconds, constraining the problem of video grounding to trimmed videos. Given these annotations, the grounding task can be simplified to choosing one out of 21 possible proposals for each video. Conversely, MAD provides a setup to explore solutions for grounding language in long-form videos, whose length can be up to 3 hours. In this scenario, naive sliding window techniques for proposal generation could produce hundreds of thousands of possible candidates. Therefore, developing efficient inference methods becomes a much more urgent requirement compared to previous benchmarks.

State-of-the-art video grounding methods [11, 12, 14, 25, 32, 34, 35] have relied on existing benchmarks to design novel modules (*e.g.* proposal generation, context modeling, and multi-modality fusion). However, most of these designs specifically target grounding in short videos and often rely on providing the entire video to the model when making a prediction. As the long-form setup introduced by MAD

prohibits this, new methods will have the opportunity to investigate and bridge previous ideas to these new challenging and real-world constraints.

Audio Descriptions. The pioneering works of Rohrbach *et al.* [20] and Torabi *et al.* [27] are the first to exploit audio descriptions to study the text-to-video retrieval task and its counterpart video-to-text. Rohrbach *et al.* [20] introduced the MPII-MD dataset, while Torabi *et al.* [27] presented M-VAD, both collected from audio descriptions in movies. Later, these datasets were fused to create the LSMDC dataset [21], which forms the core of the LSMDC annual challenge. Our annotation pipeline is similar in spirit to those adopted by these works. However, MAD seizes the potential of movies data source for the grounding task.

A concurrent work introduced a new benchmark based on audio descriptions in videos, called QuerYD [15]. It is a dataset for retrieval and event localization in videos crawled from YouTube. This benchmark focuses on the short-form video setup with videos of less than 5 minutes average duration. QuerYD also leverages audio descriptions, which have been outsourced to volunteer narrators. Similar to our takeaways, the authors noticed that audio descriptions are generally more visually grounded and descriptive than previously collected annotations.

3. Collecting the MAD Grounding Dataset

In this section, we outline MAD’s data collection pipeline. We follow two independent strategies for creating the training and testing set. For the former, we aim at automatically collecting a large set of annotations. While for the latter, we re-purpose the manually refined annotations in LSMDC. Finally, we provide detailed statistics of MAD’s annotations and compare them to existing datasets.

3.1. MAD Training set

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. Thus, they adopt a highly descriptive and diverse language. Audio descriptions are often available as an alternative audio track that can replace the original one. Professional narrators curate them, and significant effort is undertaken to describe a movie. The audio description process demands an average of 30 work hours to narrate a single hour of video [21]. In comparison, previous datasets that have used the Amazon Mechanical Turk service for video-language grounding estimate the annotation effort to be around 3 hours for each video hour [6].

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-

Dataset	Videos			Language Queries						
	Total Duration	Duration / Video	Duration / Moment	Total Queries	# Words / Query	Total Tokens	Vocabulary			
							Adj.	Nouns	Verbs	Total
TACoS [18]	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 [3]	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 [6]	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).

source and online repository¹. These audio files contain the original movie track mixed with the narrator’s voice, carefully placed when actors are not speaking. One potential problem is that the audio descriptions can be misaligned with the original movie. Such misalignment comes either from a delay in the recording of the audio description (concerning the original movie) or from audio descriptions being created from different versions of the movie (with deleted or trimmed scenes).

Alignment and Cleanup. 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. We define the original audio signal (f), the audio description signal (g), and the time delay (τ_{delay}) between the two signals. The maximum of the cross-correlation function (denoted with the operator \star) indicates the point in time where the signals exhibit the best alignment. As a result, the time delay τ_{delay} between f and g is defined as follows:

$$\tau_{\text{delay}} = \arg \max_t ((f \star g)(t)) \quad (1)$$

To verify that our single time delay τ_{delay} defines the best possible alignment between the audio descriptions and the original movies, we run our synchronization strategy over several temporal windows. In particular, we select 10 timestamps uniformly distributed along the video, and create 10 clips of length 1000 seconds and 10 clips of length 2000 seconds. We verify that the delays estimated for all 20 clips are consistent with each other, within a maximum range of ± 0.1 seconds w.r.t the median value of the distribution of the 20 samples. We discard the movies that do not satisfy this criterion to ensure that the full audio description track correctly aligns with the original movie’s visual content.

Audio Transcriptions and Verification. After the two audio tracks are aligned, we transcribe the audio description

file using Microsoft’s Azure Speech-to-Text service². Each recognized word is associated with a temporal timestamp. At this step in our pipeline, we have sentences temporally grounded to the original video stream. As the Speech-to-Text contains both the narrations and the actors’ speech, we set out to remove the latter and only retain the audio description in textual form. To do so, we resort to the movie’s original subtitles and use their timestamps as a surrogate for Speech Activity Detection (SAD). Particularly, we discard closed captions and retain subtitles associated with the actors’ speech or subtitles for songs. Then, we remove from the Speech-to-Text output every sentence overlapping with the SAD temporal locations, obtaining our target audio description sentences. We post-process the text for automatic punctuation refinement using the free tool Punctuator [26].

3.2. From LSMDC to MAD Val/Test

As outlined in Section 3.1, the test set of MAD also relies on audio descriptions for movies. Since the annotations in training are automatically generated, we decided to minimize 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 in the LSMDC dataset [21]. This dataset collected annotations from audio descriptions in movies targeting the video retrieval task. LSMDC manually refined the grammar and temporal boundaries of sentences. As a consequence, these annotations have very 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. Again, the full-length video data and the chunked video data provided by LSMDC might not be in sync. To align a video chunk from LSMDC with our full-length

¹<https://www.audiovault.net/>

²<https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/>

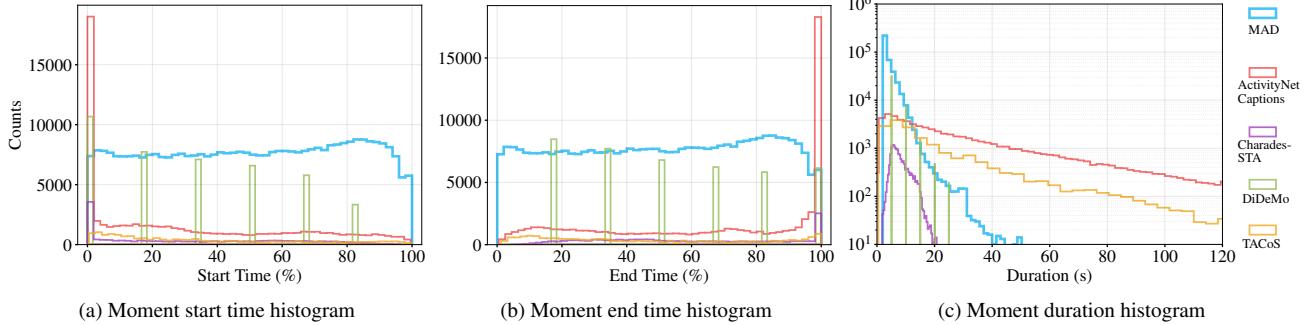


Figure 3. Histograms of moment start/end/duration in video-language grounding datasets. The plots represent the normalized (by video length) start/end histogram (a-b) and absolute duration distribution (c) for moments belonging to each of the five datasets. We notice severe biases in ActivityNet-Captions and Charades-STA, which show high peaks at the beginning and end of the videos. Conversely MAD does not show any particular preferred start/end temporal location.

movies, we follow a similar procedure as the one described in Section 3.1 for the audio alignment, but using visual information. We use CLIP [17] to extract five frame-level features per second for both the video chunks from LSMDC and our full-length movie. Then, we use the maximum cross-correlation score to estimate the delay between the two. We run the alignment on 10 different window lengths and take the median value as the delay of the chunk.

Once the audiovisual data is aligned, we use the text annotations and re-define their original timestamps according to the calculated delays. This process creates full-length movies with curated grounding data for MAD’s validation and test sets. In doing so, we obtain large and clean validation and test sets, since LSMDC was curated by humans and has been used for years by the community. Our val/test sets evaluate video-grounding methods with more than 104K grounded phrases coming from more than 160 movies.

3.3. MAD Dataset Analysis

We now present the most relevant statistics of MAD and compare it with other datasets for video grounding. Table 1 summarizes the most notable aspects of MAD, including total video duration, average moment length, total language queries, and a vocabulary breakdown.

Scale and Scalability. MAD is the largest dataset in video hours and number of sentences. As shown in Table 1, our training, validation, and test sets consist of 488, 50, and 112 movies, 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. Overall, MAD splits contain a combined 50 days of continuous video. We highlight that our test set alone is already larger than any other video grounding dataset. We also emphasize that each movie in MAD is long and composed of several diverse scenes, making it a rich source for long-form video analysis. Also, as the cinema industry is ever-growing, we expect to

periodically expand MAD with subsequent releases to fuel further innovation in this research direction.

Vocabulary Size. Besides having the largest video corpus, MAD also contains the largest and most diverse query set of any dataset. In Table 1, we show that 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. The number of sentences in MAD training, validation, and test is 280.5K, 32.1K, and 72.0K, respectively, one order of magnitude larger than the equivalent set in any other dataset. Overall, MAD contains 61.4K unique words, almost 4 times more than the 15.4K of ActivityNet-Captions [6] (the highest among the other benchmarks). Finally, the average length per sentence is 12.7 words, which is similar to the other datasets.

Bias Analysis. Figure 3 plots the histograms for start/end timestamps of moments in all grounding datasets. We notice clear biases in current datasets: Charades-STA [3], DiDeMo [1], and ActivityNet-Captions [6]. Charades-STA and ActivityNet-Captions are characterized by tall peaks at the beginning (Figure 3a) and end (Figure 3b) of the video, meaning that most temporal annotations start at the video’s start and finish at the video’s end. This bias is easily learned by algorithms, resulting in trivial groundings of moments that span a full video. The smaller dataset TACoS also exhibits a similar bias, although it is less pronounced. DiDeMo is limited by its annotation strategy, where chunks of 5 seconds are labeled up to a maximum of 30 seconds. This favors structured responses that roughly approximate the start and end points of a moment. In contrast, MAD has an almost uniform histogram. This means that moments of interest can start and end at any point in the video. We only observe a minor imbalance where the end of the movie has slightly more descriptions than the beginning. This is related to the standard structure of a film, where the main plot elements are resolved towards the end, thus creating more situations worth describing. Figure 3c plots the histograms

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 [17]	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 [25]	3.50	11.74	18.32	38.41	49.65	2.63	9.49	15.20	33.68	43.95	1.61	6.23	10.18	25.33	34.18

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. The first two validate the choice of proposals by computing the upper bound to the performance and the random performance. CLIP and VLG-Net use visual and language features to score and rank proposals. For all experiments, we adopt the same proposal scheme as in VLG-Net [25], and use CLIP [17] features for video and language embeddings.

for moment duration. MAD is characterized by shorter moments on average, having a long tail distribution with moments that last up to one minute.

4. Experiments

We now proceed with the experimental assessment of video-language grounding on the MAD dataset. We first describe the video grounding task in the MAD dataset along with its evaluation metrics and then report the performance of four selected baselines.

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, 3].

Metric. Following the grounding literature [1, 3], we adopt Recall@ K for IoU=θ (R@ K -IoU=θ). Given a ranked set of video proposals, this metric measures if any of the top K ranked moments have an IoU larger than θ with the ground truth temporal endpoints. Results are averaged across all test samples. Given the long-form nature of our videos and the large amount of possible proposals, we investigate Recall@ K for IoU=θ with $K \in \{1, 5, 10, 50, 100\}$ and $\theta \in \{0.1, 0.3, 0.5\}$. This allows us to evaluate for loose alignment (*i.e.* IoU=0.1) and approximate ranking (*i.e.*, $K = 100$), as well as tight predictions (*i.e.* IoU=0.5) and accurate retrieval (*i.e.* $K = 1$).

Baselines. We benchmark MAD using four different grounding strategies, namely: *Oracle*, *Random Chance*, *CLIP* [17], and *VLG-Net* [25]. The first two provide upper bounds and random performance for the recall metric given a predefined set of proposals. *Oracle* chooses the proposal with the highest IoU with the ground-truth annotation, while *Random Chance* chooses a random proposal with uniform probability. We also use *CLIP*, the pre-trained image-text architecture from [17], to extract frame-level and sentence-level features. 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* [25] as a representative, state-of-the-art method for the grounding task. *VLG-Net* leverages recent progress in Graph Con-

volution Networks (GCNs) [8–10, 29] to model individual modalities (video and language in the case of grounding), while also enabling the aggregation of non-local and cross-modal context through graph convolutions. We use *VLG-Net* and adapt it to work with MAD’s long-form videos. See Appendix A.2 for additional details.

Implementation Details. For the lower bound estimation using *Random Chance*, we average the performance over 100 independent runs. To favor a fair comparison against the *CLIP* baseline, we train *VLG-Net* using *CLIP* features for both modalities. We use the official *VLG-Net*’s implementation with a clip size of 128 input frames spanning 25.6 seconds (frames are extracted at 5 fps). We train using the Adam optimizer [5] with learning rate of 10^{-4} . For inference over an entire movie, we adopt a sliding window approach, where we stride the input window by 64 frames and discard highly redundant proposals through Non Maximum Suppression (NMS) with a threshold of 0.3.

4.1. Grounding Performance on MAD

Table 2 summarizes the baseline performance on MAD. 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 reach almost 50%, a 5× relative improvement. An even larger gap is present for the most strict metric, R@1-IoU=0.5, with a relative improvement of two orders of magnitude.

The *CLIP* [17] 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 in a zero-shot fashion, it results in a strong baseline for long-form grounding, achieving the best R@K for the least strict IoU=0.1 at $K = \{1, 5, 10\}$. Although IoU=0.1 corresponds to very loose grounding, this result is nonetheless valuable given a large number of negatives in the long-form

Model	IoU=0.1		IoU=0.3		IoU=0.5	
	R@1	R@5	R@1	R@5	R@1	R@5
Oracle	100.00	—	99.89	—	99.41	—
Random Chance	3.36	15.58	1.43	6.97	0.51	2.52
CLIP [17]	20.58	43.89	9.56	28.45	3.87	15.16
VLG-Net [25]	20.09	45.10	14.64	37.29	8.51	25.91

Table 3. Short video setup. The table showcases the performance of the selected baselines in a short-video setup, where movies are chunked into three minutes (non-overlapping windows). VLG-Net achieves the best grounding performance in most metrics, which fall short of CLIP in the long-form setup. We can conclude that a new generation of deep learning architectures will have to be investigated to tackle the specific properties of the MAD dataset.

setup, and the fact that MAD is characterized by containing short moments (4.1s on average). Although VLG-Net is trained for the task at hand, it achieves comparable or better performance with respect to CLIP only when a strict IoU (IoU=0.5) is considered. However, it lags behind CLIP for most other metrics. 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 must compare the sentence against all segments of the movie and then aggregate the predictions in a post-processing step. *(ii)* VLG-Net training procedure defines low IoU moments as negatives, thus favoring high performance only for higher IoUs.

4.2. The Challenges of Long-form Video Grounding

This section presents an in-depth analysis of the performance of the selected baselines in the long-form setup. We first investigate how the performance changes when the evaluation is constrained over segments of the movie, whose length is comparable to current datasets. Then explore how methods behave as the size of the movie chunks changes. 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.

Short-video Setup. In Table 3, we set the short-video window length to three minutes. This duration is a candidate representative for short videos. The upper bound performance *Oracle* slightly decreases to 99.41% for IoU=0.5. This is a consequence of the division into short videos, which occasionally breaks down a few ground truth moments. The *Random Chance* baseline reports increased performance as the number of proposals generated is reduced. In particular for R@5-IoU=0.1, the performance increases from 0.44% (Table 2) to 16.00% (Table 3), demonstrating that the short-video setup is less challenging compared to MAD’s original long-form configuration. In a similar trend, performance substantially increases for both CLIP

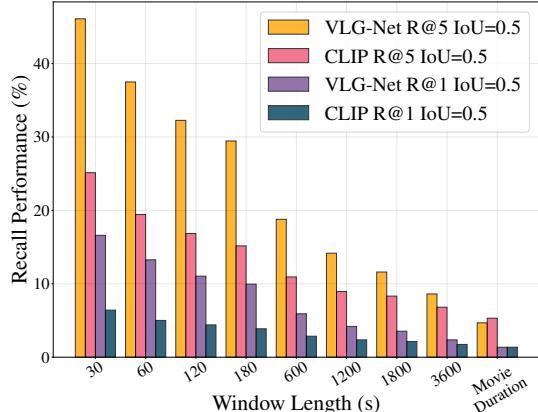


Figure 4. Performance trend across different windows lengths. We observe from the graph the decrease in performance for both CLIP and VLG-Net, as the evaluation window length increases. This demonstrates that current grounding methods cannot tackle the task in the long-form video setting.

and VLG-Net baselines, with the latter now obtaining the best performances, in most cases.

From Short- to Long-form Results. Figure 4 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 graphs display how the performance steadily drops as the window length increases, showing the challenging setup of long-form grounding enabled by MAD.

Takeaway. This set of experiments verify that VLG-Net could successfully outperform the zero-shot CLIP baseline when evaluated in a short-video setup. We can conclude that current state-of-the-art grounding methods are not ready to tackle the long-form setting proposed by MAD. This opens the door to opportunities for the community to leverage previously developed techniques in a more challenging setting and potentially incorporate new constraints when designing deep learning architectures for this task.

5. Ablation Study

This section focuses on the empirical assessment of the quality of MAD training data. To help the reader navigate the following experiments, Table 4 defines a naming convention associated with different splits of the data. LSMDC16 refers to the original data collected by [20, 21, 27] for the task of text-to-video retrieval. LSMDC-G is our adaptation of this dataset for the grounding task, as described in Section 3.2. We remind the reader that we could not retrieve every movie (only 162/182) from LSMDC16. For these two datasets, we follow the original train/val/test partitions. Finally, we include MAD’s split details.

Improving Grounding Performance with MAD Data.

We are interested in evaluating the contribution our data can

Dataset Name	Task	Videos		Annotations	
		Train / Val / Test			
LSMDC16 [21]	Retrieval	155 / 12 / 17	101.1 K	7.4 K	10.1 K
LSMDC-G	Grounding	138 / 11 / 13	89.7 K	6.7 K	7.6 K
MAD	Grounding	488 / 50 / 112	280.5 K	32.1 K	72.0 K

Table 4. **Data split cheat-sheet.** This table clarifies the data splits used in the following experiments (Table 5 and Table 6). LSMDC16 [21] is the original data collected for retrieval. LSMDC-G is our adaptation to the grounding task. MAD is our proposed dataset. Note that LSMDC-G videos and annotations constitute the validation and test splits of MAD.

bring to the grounding task. We investigate the performance of VLG-Net in the long-form grounding setup, when the training data changes. All trained models are evaluated on the same test split: LSMDC-G test.

The first row of Table 5 shows the performance when VLG-Net is exclusively trained on the LSMDC-G training split. This set only contains data that was manually curated in [21]. The second row is trained with 32% of MAD-training data, which is equivalent to the size of LSMDC-G training split. We observe a drop in performance, which can be associated with the presence of noise introduced by the automatic annotation process in MAD. In the third row, we use the full MAD training set. Here, MAD’s scale allows us to overcome the performance issues associated with noisy data, and to actually improve performance with respect to only using the clean LSMDC-G for training. Using 100% of MAD data yields a relative improvement of 20% for R@5-IoU=0.5. Then, we investigate whether the performance of VLG-Net saturates given the amount of data available. To this end, we use 100% of LSMDC-G training and gradually augment it by adding MAD training samples. In these three experiments (rows 4-6), the performance steadily increases. These results suggest that current models for video ground-

Training Set		Testing Set		IoU=0.5		
% LSMDC-G	% MAD	LSMDC-G		R@1	R@5	R@10
100%	0%	Test		1.36	5.18	8.82
0%	32%	Test		0.60	2.60	5.11
0%	100%	Test		1.51	6.23	10.18
100%	32%	Test		2.18	6.63	10.73
100%	64%	Test		2.23	7.79	11.74
100%	100%	Test		2.82	8.74	13.36

Table 5. **Grounding performance with varying training data.** We investigate VLG-Net [25] grounding performance on LSMDC-G test, when different data is used for training. This highlights the quality and usefulness of our automatically collected data (MAD training) against the manually curated one provided by LSMDC-G. We conclude that expensive and time-consuming manual curation can be avoided if volume of data is large.

% LSMDC16	% MAD	Training Set	Testing Set	R@1 R@5 R@10		
				LSMDC16		
100%	0%	Test	Test	20.9	39.4	48.5
0%	36%	Test	Test	19.2	35.5	44.8
0%	100%	Test	Test	20.5	38.8	48.7
100%	36%	Test	Test	23.3	40.3	48.8
100%	72%	Test	Test	23.6	41.4	49.3
100%	100%	Test	Test	24.8	40.5	50.0

Table 6. **Retrieval performance on LSMDC16 with model CLIP4Clip [13].** This experiment showcases how MAD data can be valuable for a related task, beyond grounding.

ing do benefit from larger scale datasets, even though the automatically collected training data may be noisy. Moreover, designing scalable strategies for automatic dataset collection is crucial, as the drawbacks of noisy data can be offset, and even overcome, by mining more training data.

Improving Retrieval Performance with MAD Data. Adopting the same ablation procedure, we evaluate the possible contribution of our data in a related task, namely text-to-video retrieval. Here, we format our MAD data in the same way as LSMDC16, where short-videos are trimmed around the annotated timestamps. For this experiment, we use CLIP4Clip [13], a state-of-the-art architecture for the retrieval task, as baseline. Table 6 reports the performance when different amounts of data is used for training. Again, we see that training with the whole LSMDC16 or MAD leads to a very similar performance. Moreover, our previous argument also holds true in this task. Pouring more data into the task boosts the performance, motivating the benefit of having a scalable dataset like MAD.

Takeaway. The MAD dataset is able to boost performance in two closely related tasks, video grounding and text-to-video retrieval, where we demonstrate scale can compensate for potential noise present due to automatic annotation.

6. 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. Our methodology comes with two main hypotheses and limitations: (i) Noise cannot be avoided but can be dealt with through scale. (ii) Due to copyright constraints, MAD’s videos will not be publicly released. However, we will provide all necessary features for our experiments’ reproducibility and promote future research in this direction.

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, 3, 4, 5, 6
- [2] Andrew E Budson and Bruce H Price. Memory dysfunction. *New England Journal of Medicine*, 352(7):692–699, 2005. 1
- [3] 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, 3, 4, 5, 6
- [4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, 2015. 3
- [5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [6] 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, 3, 4, 5
- [7] 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. 2
- [8] Guohao Li, Matthias Müller, Guocheng Qian, Itzel Carolina Delgadillo Perez, Abdulellah Abualshour, Ali Kassem Thabet, and Bernard Ghanem. Deepgencs: Making gcns go as deep as cnns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 6
- [9] Guohao Li, Matthias Müller, Ali Thabet, and Bernard Ghanem. Deepgencs: Can gcns go as deep as cnns? In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 6
- [10] Guohao Li, Chenxin Xiong, Ali Thabet, and Bernard Ghanem. Deepergcn: All you need to train deeper gcns, 2020. 6
- [11] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11235–11244, June 2021. 3
- [12] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Jointly cross-and self-modal graph attention network for query-based moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4070–4078, 2020. 3
- [13] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 8
- [14] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-Global Video-Text Interactions for Temporal Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3
- [15] Andreea-Maria Oncescu, João F. Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie. Queryd: A video dataset with high-quality text and audio narrations. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2265–2269, 2021. 3
- [16] 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, 2, 3
- [17] 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. 5, 6, 7, 12
- [18] 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, 3, 4
- [19] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. In *German conference on pattern recognition*, pages 184–195. Springer, 2014. 3
- [20] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015. 3, 7
- [21] 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. 3, 4, 7, 8
- [22] Gunnar A. Sigurdsson, Gülsün Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 510–526, Cham, 2016. Springer International Publishing. 3
- [23] 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
- [24] Cees Snoek, Kvd Sande, OD Rooij, Bouke Huurnink, J Uijlings, M van Liempt, M Bugalhoj, I Trancosoy, F Yan, M Tahir, et al. The mediaville trecvid 2009 semantic video search engine. In *TRECVID workshop*. University of Surrey, 2009. 1
- [25] 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, 6, 7, 8, 12
- [26] Ottokar Tilk and Tanel Alumäe. Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *Interspeech 2016*, 2016. 4

- [27] Atousa Torabi, Christopher Pal, Hugo Larochelle, and Aaron Courville. Using descriptive video services to create a large data source for video annotation research. *arXiv preprint arXiv:1503.01070*, 2015. 3, 7
- [28] 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
- [29] Mengmeng Xu, Chen Zhao, David S. Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 6
- [30] 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
- [31] 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
- [32] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense Regression Network for Video Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3
- [33] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Towards debiasing temporal sentence grounding in video. *arXiv preprint arXiv:2111.04321*, 2021. 1
- [34] Zhang Songyang, Peng Houwen, Fu Jianlong, Luo, Jiebo. Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 3
- [35] Yang Zhao, Zhou Zhao, Zhu Zhang, and Zhijie Lin. Cascaded prediction network via segment tree for temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4197–4206, June 2021. 3
- [36] 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

Split	Videos			Language Queries						
	Total Duration	Duration / Video	Duration / Moment	Total Queries	# Words / Query	Total Tokens	Vocabulary			
							Adj.	Nouns	Verbs	Total
MAD (Train)	891.8 h	109.65 min	4.0 s	280.5K	13.5	3.8.M	4.8K	33.5K	12.2K	57.6K
MAD (Val/Test)	315.5 h	116.85 min	4.1 s	104.1K	10.6	1.1M	2.2K	11.6K	5.8K	21.9K

Table 7. **Statistics comparison between MAD training and MAD val/test splits.** We verify that the two splits follow similar distributions. We assess that the average video duration, moments length, and sentence length have similar values. Moreover, we highlight how 2/3 of the video content is reserved for the training split. The size of the training split is also reflected in the total number of queries, with the training set being 2.7× larger than the val/test set.

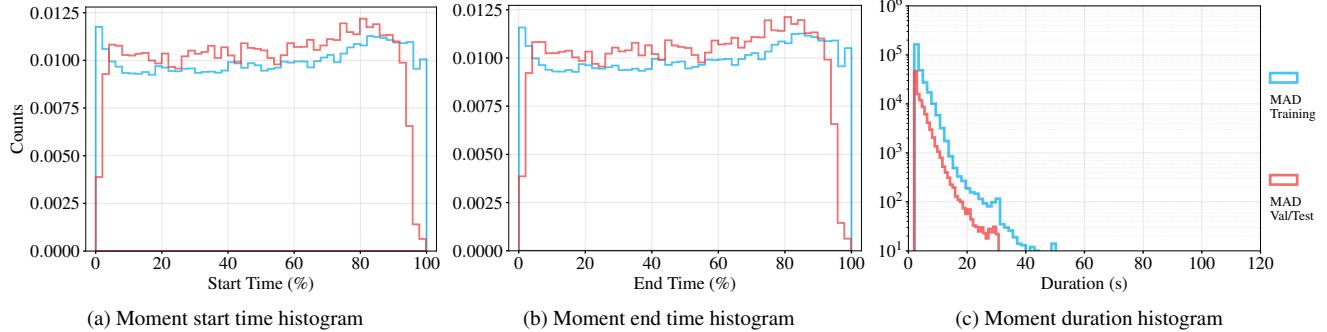


Figure 5. **Histograms of moment start/end/duration in MAD splits.** The plots represent the normalized (by video length) start/end distributions (a-b), and absolute duration distribution (c) for moments belonging to the training and val/test splits of MAD. The figure showcases that both training and val/test splits follow the same distributions with minor differences between them.

A. Appendix

A.1. MAD Detailed Statistics

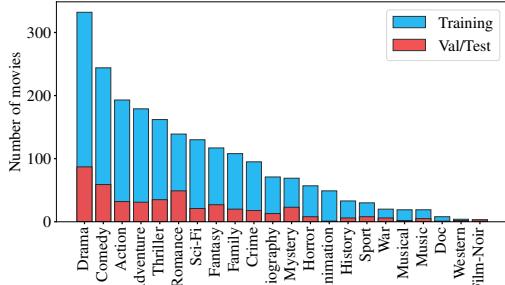
As described in Section 3 in the main paper, the training set is automatically collected and annotated, whereas the val/test sets of MAD were adapted from the LSMDC dataset. Considering this difference in the splits of MAD, we analyze in detail the key statistics and discrepancies between the training and the val/test sets. We summarize these results in Table 7 and Figure 5.

As shown in Table 7, the training set contains about 2/3 of the total MAD video hours and query sentences, val/test sets contain 1/3. For the video data, the average duration is quite similar between the two splits. Train videos are, on average, only 6.2% shorter than those in val/test, the mean span of a moment is almost equal with only 0.1 seconds difference. Regarding the language queries, the training set has slightly longer sentences than the val/test sets, with an average of 2.9 extra words per sentence. We notice a significant difference between the two splits in the vocabulary size, for training is almost two times larger than val/test. The size of the vocabulary correlates with the diversity in the language queries, a larger vocabulary is a desirable feature in training, considering that real-word scenarios might contain different words to express similar semantics. Finally, the intersection between the vocabulary of val/test and

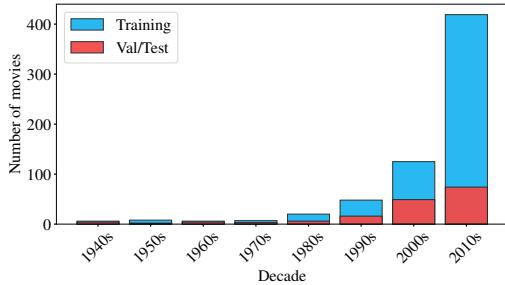
training is 83%. While the majority of tokens are present in training, there are some words in val/test which were never seen in training, such an arrangement could be important to evaluate the generalization capability of models developed in MAD.

Figure 5 shows the distributions of the relative moment’s start time (a), relative end time (b). Subfigure (c) shows the number of segments by its duration. We show MAD’s training split in light blue and val/test in red. We observe that the two splits have similar distributions in all three sub-figures. Although, we notice that the training set has slightly more moments at the very beginning and the very end of the videos. We attribute this discrepancy to the fact that we didn’t remove the audio descriptions from the movie’s opening and credits, as there is not an automatic and reliable way to drop them; LSMDC *manually* removes those. We opt for including such annotations in our data. Overall, this design decision has little impact on the data distribution, but saves manual effort and keeps our collection method highly-scalable. For the moment’s duration, both splits exhibit a bias towards short instances and have a long tail up to 294 seconds for train and 176 seconds for the val/text.

Figure 6 shows the distributions of genres and years of MAD movies. We can see that MAD has a wide range in years the movies are produced (1940s to last decade) along with a large variety in genres. The production year of a



(a) Movies genres.



(b) Movies release year.

Figure 6. **Diversity.** The Figure depicts the wide diversity contained in the dataset. Spanning 22 different genres and 90 years of cinema history, MAD presents a highly diverse dataset for the video grounding task.

movie is closely related to its picture quality, filming techniques, people’s clothing and apparel, action types, etc. The movie genre characterizes how people behave and talk, how the story is told, the overall scenes set, how dense information is displayed, etc. These diversities are contained in MAD’s videos and descriptions, thus providing our dataset with a large diversity in video content and its related query sentences.

A.2. VLG-Net Long-Form Adaptation

In the paper, we selected VLG-Net [25] as a representative model of the state-of-the-art architectures for natural language grounding in videos. However, the challenging long-form nature of the MAD dataset requires some technical changes in the architecture.

(i) Input. VLG-Net default inputs are frame/snippet level features for the entire video. As videos are of different duration, VLG-Net interpolates or extrapolates the features to a predefined number before passing them to the architecture. We change this modeling strategy for the following; We input a constant number of frame features (i.e., 128) extracted at a constant frame rate (i.e., 5 FPS). During training, we randomly select a window of frames that contains each annotation and ground the moment accordingly. Such window changes at each epoch. This strategy can be interpreted as a regularization technique that prevents the model from

leveraging biases in the input representation and promotes the model to correctly understand the multi-modal input to predict the best temporal extent for each language query.

During inference, we adopt a sliding window technique, matching each annotation against each possible window in a movie. We select a window stride of 64 frames. We aggregate all windows predictions for each sentence for the ground truth movie the sentence belongs to and make a global grounding prediction.

(ii) Negatives. VLG-Net does not use negative samples during training. This means that only positive video-language pairs are used. Following the change in the input modeling, VLG-Net has only access to a local portion of the video when making a prediction. Therefore, it is necessary for us to train using negatives/unpaired video-language pairs. We define a negative as a video window (128 frames) that has $\text{IoU}=0$ with the ground truth for each sentence. At training time, for each sentence, we randomly select a negative within the same video. We do not consider cross-movie negatives. Empirical evidence shows that selecting a negative 70% of the times yields the best performance.

(iii) Modules. In Section 4 we described how we adopted CLIP [17] features for both the visual and language stream. This promotes the fairest comparison against the baseline CLIP. However, it poses a technical challenge. Following the requirement to change the original Glove-based embeddings for CLIP ones, we are required to remove the syntactic graphs in the language branch. We could not retrieve such information from the CLIP tokenizer. Instead, we opted to remove such layers and only retain the LSTM layers.