Supplementary material for

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

A. MAD Detailed Statistics

This section provides additional statistics for the MAD dataset. First, we compare the automatically curated training set against the manually curated validation and test sets, highlighting similarities and differences. Second, we assess the presence of repetitive sentences which might be ambiguous for the language grounding in videos task. We follow by providing additional statistics about MAD's vocabulary and conclude the sections highlighting MAD's large visual and language diversity.

A.1. Data splits comparison

As described in Section 3 of the main paper, the training set is automatically collected and annotated, whereas the val/test sets of MAD were adapted from the LSMDC dataset [6]. Considering this difference, we analyze the key statistics and discrepancies between the training and the val/test sets in detail. We summarize these results in Table 1 and Figure 1.

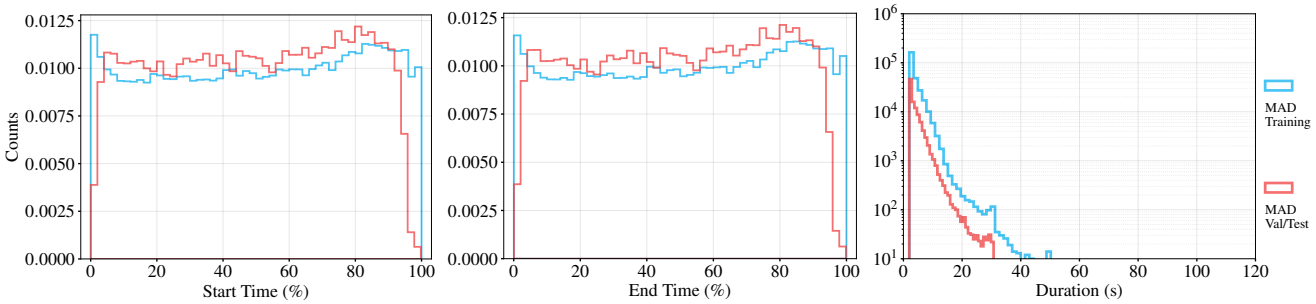
As shown in Table 1, the training set contains about $3/4$ of the total video hours and query sentences in MAD, val/test sets contain $1/4$. The average video duration in the two splits is similar, with training videos being, on average, only 6.2% shorter than those in val/test. Moreover, the average temporal span of the moments is very similar in the two splits, with a difference of only 0.1 seconds, on average. Regarding the language queries, the training set has slightly longer sentences than the val/test sets, with on average 2.9 extra words per sentence. We attribute this fact to the automatic annotations procedure of the training set. We observe that sometimes consecutive sentences that are annotated in a short temporal span can be joined together by our annotation pipeline. This does not happen for the val/test set, as sentences were manually refined.

Table 1 also highlights a significant difference between the two splits regarding the vocabulary size. The training vocabulary (57.6K tokens) is almost three times larger than the one of val/test (21.9K tokens). Note that the vocabulary size correlates with the diversity in the language queries. Thus, a more extensive vocabulary is a desirable feature in training, considering that real-world application scenarios might use a variety of words to express similar semantics. Finally, the overlap between val/test and training vocabularies is 83%, accounting for 18.1K unique words. There are 3.8K val/test tokens that do not overlap the training set vocabulary. However, these tokens only account for 0.69% of the total tokens in the val/test splits ($1.1M$). Moreover, there are 39.5K unique tokens in the training set that are not present in the val/test. Such unique tokens account for 6.6% of the total training tokens ($3.8M$). These features of the dataset will be valuable to evaluate the generalization capabilities of models developed in MAD.

Figure 1 shows the distribution of the relative start time of a moment (Fig. 1a) and the relative end time of a moment (Fig. 1b). Fig. 1c shows the distribution of segments by duration. We show MAD's training split in blue and val/test in red. We observe that the two splits have similar distributions in all three sub-figures. However, we notice that the training set has slightly more moments at the very beginning and at the very end of the videos (Fig. 1a and 1b). We attribute this discrepancy to the fact that we did not 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* removed them. We opt for including such annotations in our dataset. Overall, this design decision has little impact on the data distribution but saves manual effort and keeps our data collection method scalable. For the moment's duration (Fig. 1c), both splits exhibit a bias towards short instances and have a long tail distribution, with moments lasting up to 50 seconds for training and 30 seconds for val/test.

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 1. **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, moment 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 $2.7\times$ larger than the val/test set.



(a) Moment start time histogram (b) Moment end time histogram (c) Moment duration histogram

Figure 1. **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.2. Sentences uniqueness

Repeating sentences within a movie can be a source of ambiguity for the video-language grounding task. Our automated annotation pipeline, does not enforce individual sentences to be semantically or grammatically different. To quantify this phenomenon, we compute the METEOR similarity score between each pair of sentences within a movie. The METEOR metric is a widely used metric in NLP [2, 5] which correlates well with human judgment on sentence similarity. We use the implementation provided by the NLTK library [1] and empirically observe that the scores are bounded between $[0, 1]$. Given these boundaries, we consider a sentence to be unique if its METEOR score with **every other sentence in the movie** is below $th{=}0.99$. Following this threshold, 99.7% of sentences can be considered unique. If we lower the threshold to $th{=}0.9$, the uniqueness decreases slightly to 99.2%. This suggests that only a few sentences repeat in each movie. We emphasize that this estimation cannot directly assess the semantic similarity between sentences, which is a much harder matching problem and requires further research, but remains a good approximation.

A.3. Additional language statistics

The MAD dataset contains about 384K query sentences. The average sentence length is 12.7 tokens (see Tab 1 in the main paper) with a standard deviation of 8.1 tokens. We show in Figure 2 the distribution of the number of tokens per sentence which showcases the variability in query length in the entire dataset. It is known in the field of computational linguistics that natural language usually follows a long-tailed distribution. We find that it is also the case in the textual annotations of MAD. We compute the frequency distribution of the vocabulary words and find that only 471 unique tokens out of 61.4K repeat more than 1000 times. In comparison, that number increases to 6.3K if we relax the frequency threshold to only 50 repetitions. This means that 90% of the tokens in the vocabulary (55.1K) appear less than 50 times in the entire queries corpus.

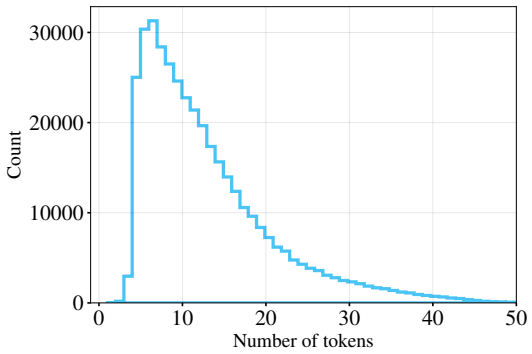


Figure 2. **Sentences length distribution.** Queries length is measured in number of tokens.

A.4. Diversity

Figure 3 shows the distributions of genres and years of MAD movies. We can see that MAD has a wide range in the years the movies are produced (from the 1940s to the last decade) and a large variety of genres. A movie's production year is closely related to its picture quality, filming, edition techniques [3], character's attire, apparel, action types, etc. The movie genre characterizes how people behave and talk, storytelling techniques, the overall scenes setup, and how fast-paced is the information displayed. These diversities are contained in MAD's videos and descriptions, thus endowing our dataset with a large diversity in video content and related query sentences.

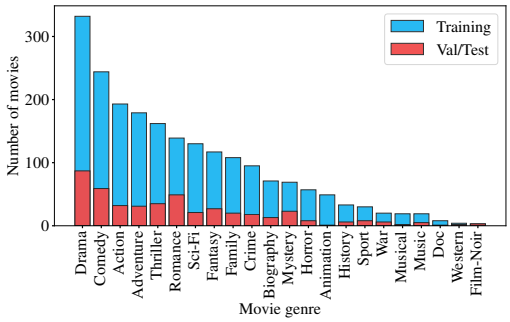
B. VLG-Net Long-Form Adaptation

In the paper, we select VLG-Net [7] as a representative model of the state-of-the-art architectures for the natural language grounding in videos task. The challenging long-form nature of the MAD dataset requires some technical changes in the architecture. We detail below the three main upgrades made to this baseline to enable the training and inference over very long videos.

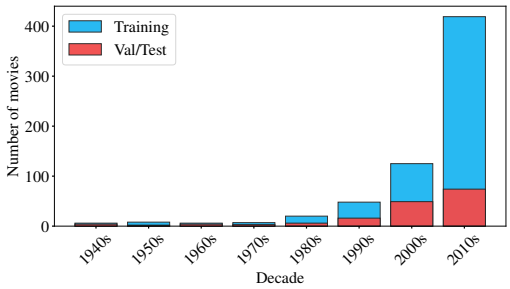
(i) Input. VLG-Net's default inputs are either frames or snippet-level features that span an entire video. As videos are of different durations, VLG-Net interpolates or extrapolates the features to a predefined length before feeding them to the remaining of the architecture. We change such modeling strategy with the following one: we consider a window of consecutive frames features (*i.e.*, 128) and input each window independently to the model instead of an entire video. Frames are sampled at a constant frame rate (*i.e.*, 5 frames per second).

During training, for a given sentence and corresponding grounding timestamps, we randomly select a window that contains the annotation's temporal extent. Let us draw an example to understand this approach better. Given a clip's frameset $V = \{v_i\}_{i=1}^{n_v}$ and an associated sentence S. We can map the grounding timestamps from the time domain to the frame-index one which we regard as (t_s, t_e) such that $t_s \geq 1$ and $t_e \leq n_v$. At training time, we sample a starting index (t_s^*) in the interval $[t_e - W, t_s]$ and construct our training window as the sequence of frames $\{v_i\}_{i=t_s^*}^{t_s^*+W}$ with $W = 128$. Note that $t_e - W \leq ts$.

This process can be seen as a temporal jittering process. Thanks to this jittering process, the window enclosing the ground-truth segment changes at every epoch (as t_s^* changes at every epoch) for a given sentence. This strategy can be interpreted as a regularization technique that prevents the model from leveraging intra-window biases in the input representation. Moreover, it promotes the model to understand the multi-modal input better and predict the best temporal extent for each language query.



(a) Movies genres.



(b) Movies release year.

Figure 3. **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.

At inference time, we adopt a sliding window technique, which strides a fixed window over the entire movie. The window size is kept fixed to 128 frames, and we use a stride of 64 frames. For each window, VLG-Net produces a set of proposals with an associated confidence score. The window-level predictions are then collected and sorted according to the confidence score. The recall metric is measured at the video-level.

(ii) Negative samples. The original VLG-Net implementation does not make use of negative samples during training. This means that only positive video-language pairs are used. Following the change in the input modeling, the model now only has access to a local portion of the video when making a prediction. Therefore, it is deemed necessary to train the VLG-Net architecture using negatives/unpaired video-language pairs. This teaches the model to predict low confidence scores for windows that do not contain visual information relevant to the query being grounded (which are the majority during inference).

Negative samples are defined as a video window (128 frames) with IoU is equal to 0 with the ground truth temporal span of a given sentence. With respect to the previous example, a negative video sample is considered as a sequence of consecutive frames of size W which starting index (t_s^*) is sampled outside of the interval $[t_s - W, t_e]$.

At training time, for each sentence, we randomly select a negative sample within the same movie with a probability

p or a positive sample (*i.e.*, window containing the ground truth) with probability $1 - p$. Our experiments show that selecting a negative 70% of the times yields the best performance. We do not consider cross-movie negative samples.

(iii) Modules. In Section 4, we described how, to promote a fair comparison against the CLIP baseline [4], we adopted CLIP's visual and language features as inputs for the VLG-Net baseline. Notably, the language feature extraction strategy poses a technical challenge. The original sentence tokenizer used by VLG-Net has the capability of extracting syntactic dependencies that are represented as edges in the SyntacGCN module. Because CLIP uses a different tokenizer, we could not retrieve such syntactic dependencies; hence we remove the SyntacGCN module and only retaining the LSTM layers for the language branch.

References

[1] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[2] 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.

[3] Alejandro Pardo, Fabian Caba Heilbron, Juan León Alcázar, Ali Thabet, and Bernard Ghanem. Moviecuts: A new dataset and benchmark for cut type recognition. *arXiv preprint arXiv:2109.05569*, 2021.

[4] 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] 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.

[6] 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.

[7] 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.