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

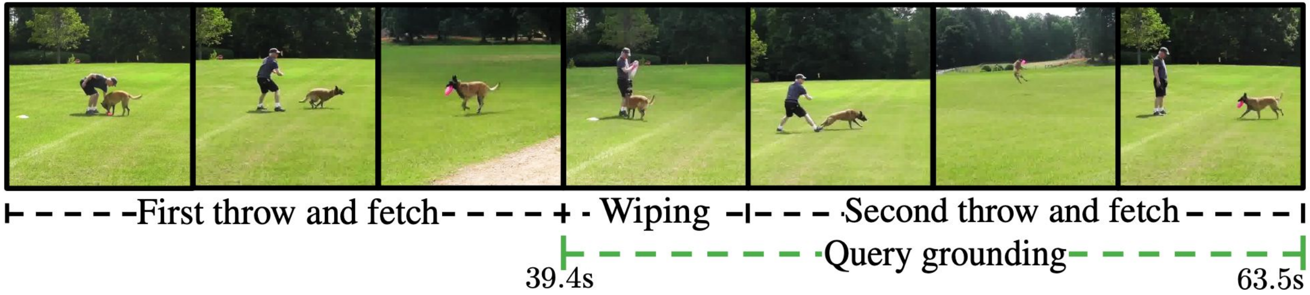
Mattia Soldan[1], Alejandro Pardo[1], Juan León Alcázar[1], Fabian Caba Heilbron[2], Chen Zhao[1], Silvio Giancola[1], Bernard Ghane[1]

[1]KAUST, [2]Adobe Research

Presented by Han and Mel

# Language Grounding in Videos

Query: ”The man wiped the frisbee and then threw it **again**,
and the dog caught and brought it back to the owner.”



# Previous Datasets and Their Limitations

## TACoS

ICCV 2013

### Limitations:

- Few videos (127)
- Static camera setup
- Domain specific (cooking only)

## DiDeMo

ICCV 2017

### Limitations:

- Coarse annotations in chunks of 5 seconds
- Only first 30 seconds of videos are annotated
- Constrains the grounding task to a simple classification among 21 possible video proposals.

# Previous Datasets and Their Limitations

Charades-STA

ICCV 2017

ActivityNet-Captions

ICCV 2017

Limitations:

- Strong priors
- Strong biases
- SOTA methods don’t use the visual information and only rely on the biases
- They drove the task development steering the research toward technical solution that made successful use of the inherent biases

# Dataset Creation

## Training Set

- Automatically collecting a large set of annotations from professional, grounded audio descriptions of movies for visually impaired audiences.
- These descriptions embody a rich narrative describing the most relevant visual information and adopt a highly descriptive and diverse language.

## Val / Test Set

- Reformat a subset of the LSMDC data, adapt it for the video grounding task.
- More than 104K grounded phrases coming from more than 160 movies.

# Movie Audio Description (MAD) Dataset

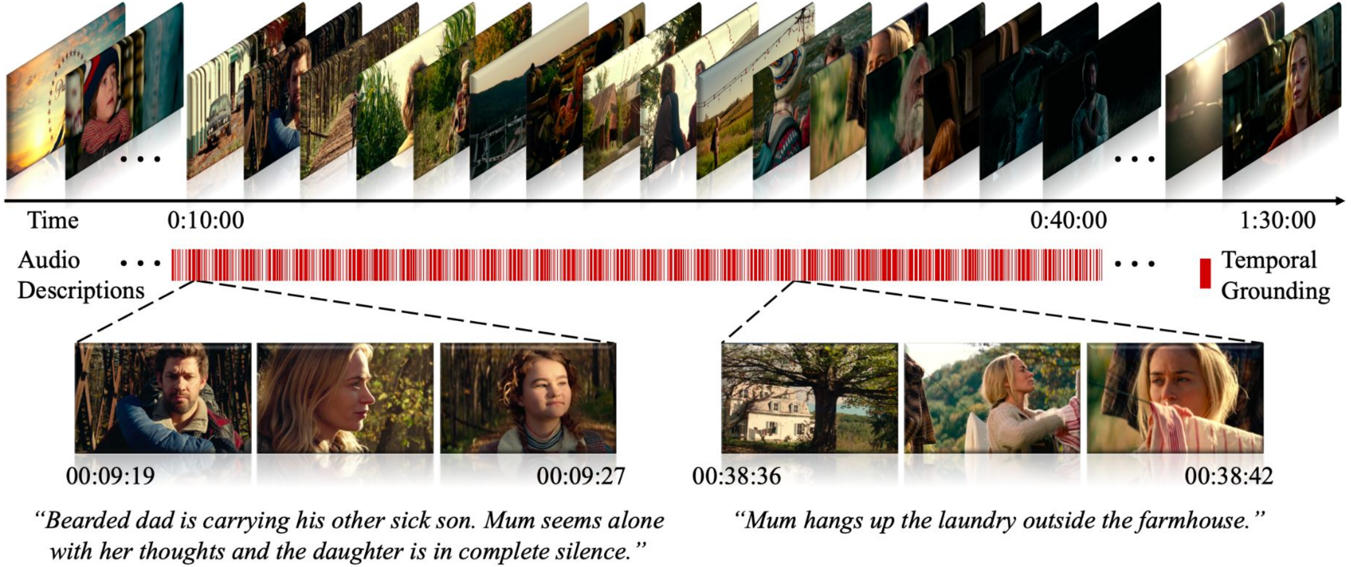
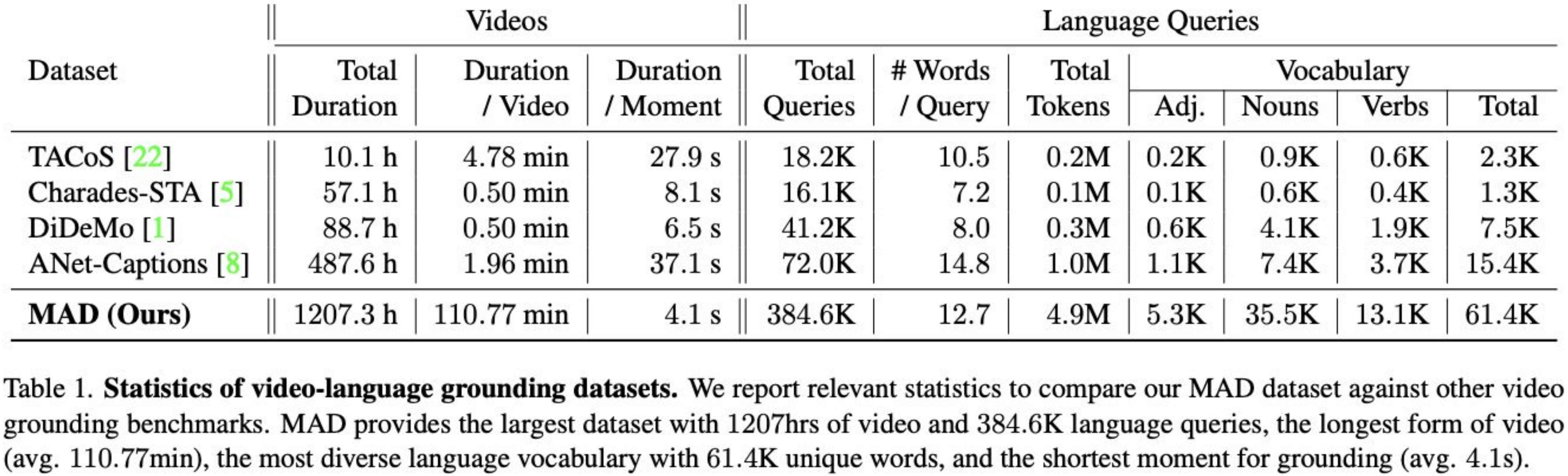

Time
0:10:00
0:40:00
1:30:00
Audio Descriptions
Temporal Grounding
00:09:19
00:09:27
00:38:36
00:38:42
"Bearded dad is carrying his other sick son. Mum seems alone with her thoughts and the daughter is in complete silence."
"Mum hangs up the laundry outside the farmhouse."

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 number 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.

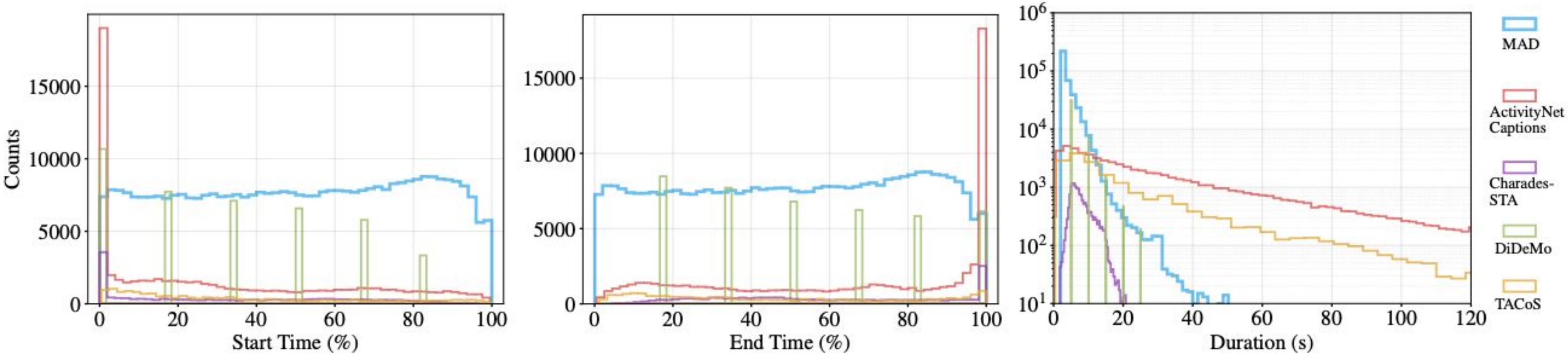
# Dataset Analysis - Scale and Vocabulary size



|  | Videos |  |  | Language Queries |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | Total Duration | Duration / Video | Duration / Moment | Total Queries | # Words / Query | Total Tokens | Vocabulary |  |  |  |
|  |  |  |  |  |  |  | Adj. | Nouns | Verbs | Total |
| TACoS [22] | 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 [5] | 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 [8] | 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).

# Dataset Analysis - Bias Analysis



(a) Moment start time histogram

(b) Moment end time histogram

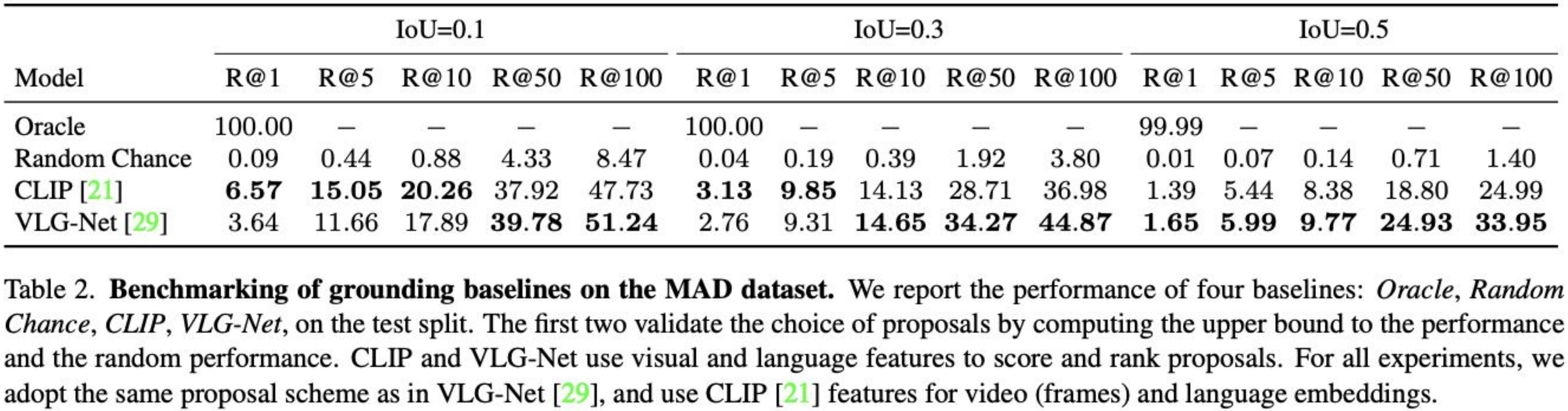
(c) Moment duration histogram

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.

# Experiments

- Goal: Given an untrimmed video and a language query, localize a temporal moment (τs, τe) in the video that matches the query
- Metric:
  - IoU - Measures overlap between prediction and ground truth
  - Recall@K for IoU = $\Theta$ → Measures if any of the top K ranked moments have an IoU larger than $\Theta$ with the ground truth temporal endpoints.
  - $K \in \{1, 5, 10, 50, 100\}$
  - $\Theta \in \{0.1, 0.3, 0.5\}$

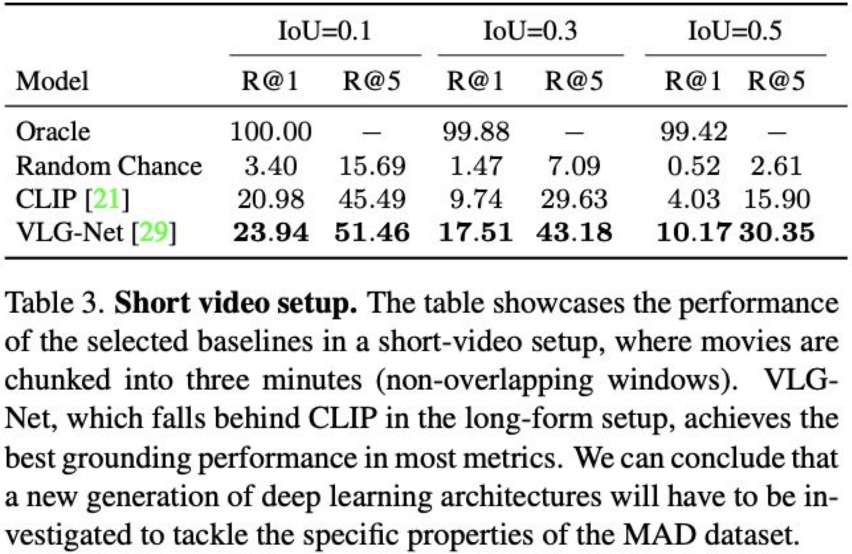
# Experiments - Baseline performance



| | IoU=0.1 | | | | | IoU=0.3 | | | | | IoU=0.5 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | 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 [21] | **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 [29] | 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. 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 [29], and use CLIP [21] features for video (frames) and language embeddings.

# Experiments - In long-form setup



| | IoU=0.1 | | IoU=0.3 | | IoU=0.5 | |
|---|---|---|---|---|---|---|
| Model | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 |
| Oracle | 100.00 | – | 99.88 | – | 99.42 | – |
| Random Chance | 3.40 | 15.69 | 1.47 | 7.09 | 0.52 | 2.61 |
| CLIP [21] | 20.98 | 45.49 | 9.74 | 29.63 | 4.03 | 15.90 |
| VLG-Net [29] | **23.94** | **51.46** | **17.51** | **43.18** | **10.17** | **30.35** |

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, which falls behind CLIP in the long-form setup, achieves the best grounding performance in most metrics. 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.

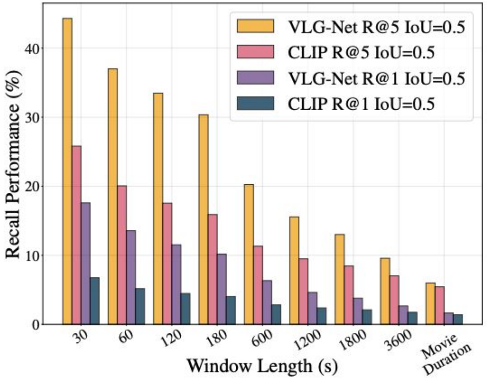
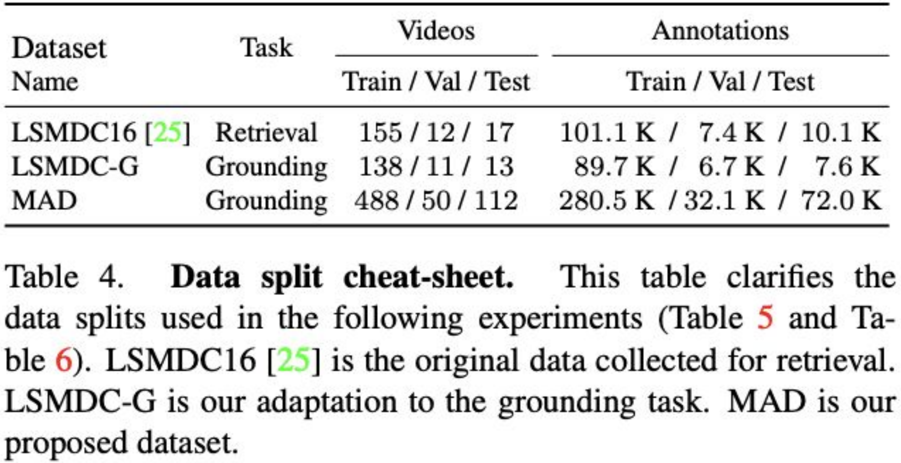


Figure 4. **Performance trend across different windows lengths.** We observe from the graph the decrease in performance for both CLIP and VLG-Net, as the evaluation window length increases. This demonstrates that current grounding methods cannot tackle the task in the long-form video setting.

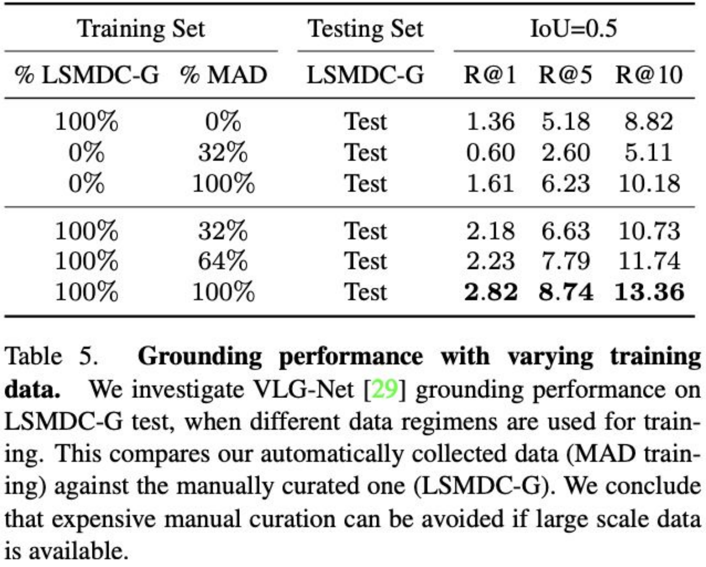
→ **Current state-of-the-art grounding methods are not ready to tackle the long-form setting proposed by MAD**

# Ablation Studies



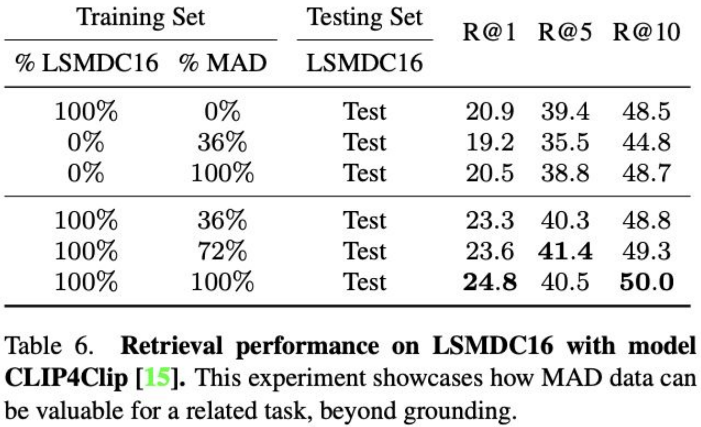
| Dataset Name | Task | Videos | Annotations |
|---|---|---|---|
| | | Train / Val / Test | Train / Val / Test |
| LSMDC16 [25] | Retrieval | 155 / 12 / 17 | 101.1 K / 7.4 K / 10.1 K |
| LSMDC-G | Grounding | 138 / 11 / 13 | 89.7 K / 6.7 K / 7.6 K |
| MAD | Grounding | 488 / 50 / 112 | 280.5 K / 32.1 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 [25] is the original data collected for retrieval. LSMDC-G is our adaptation to the grounding task. MAD is our proposed dataset.



| 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 | 1.61 | 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 [29] grounding performance on LSMDC-G test, when different data regimens are used for training. This compares our automatically collected data (MAD training) against the manually curated one (LSMDC-G). We conclude that expensive manual curation can be avoided if large scale data is available.



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

Table 6. **Retrieval performance on LSMDC16 with model CLIP4Clip [15].** This experiment showcases how MAD data can be valuable for a related task, beyond grounding.

# Conclusion

- Introduce a new video grounding benchmark, MAD
  - Over 384,000 natural language sentences in > 1,200 hours of video content
  - 650 movies spanning 20 genres and 90 years of cinema history
  - Address hidden biases in most common video-language grounding datasets
  - Introduce a new problem → Long-form grounding