Localizing Moments in Video with Natural Language

Lisa Anne Hendricks[1*], Oliver Wang[2], Eli Shechtman[2], Josef Sivic[2,3*], Trevor Darrell[1], Bryan Russell[2]
[1]UC Berkeley, [2]Adobe Research, [3]INRIA
https://people.eecs.berkeley.edu/~lisa_anne/didemo.html

This appendix includes the following material:

1. Qualitative examples illustrating when global video features and tef features improve performance.
2. Qualitative examples contrasting RGB and flow input modalities.
3. Additional qualitative examples using the full Moment Context Network. See https://www.youtube.com/watch?v=MRO7_4ouNWU for a video example.
4. Additional baselines.
5. Ablation of inter-intra negative loss.
6. Results when training without a language feature.
7. List of words used to generate numbers in Table 2 of the main paper.
8. Qualitative video retrieval experiment. See https://www.youtube.com/watch?v=fuz-UBvgapk for a video example.
9. Discussion on ambiguity of annotations and our metrics.
10. Histrogram showing the moments annotators mark in our dataset.
11. Example video showing our annotation tool (see https://www.youtube.com/watch?v=vAvT5Amp408 and https://www.youtube.com/watch?v=9WWgndeEjMU).

1. Impact of Global Video Features and TEF Features

In the main paper we quantitatively show that global video features and tef features improve model performance. Here, we highlight qualitative examples where the global video features and tef features lead to better localization.

Figure 1 shows examples in which including global context improves performance. Examples like "The car passes the closest to the camera" require context to identify the correct moment. This is sensible as the word "closest" is comparative in nature and determining when the car is closest requires viewing the entire video. Other moments which are correctly localized with context include "we first see the second baby" and "the dog reaches the top of the stairs".

Figure 2 shows examples in which including temporal endpoint features (tef) correctly localizes a video moment. For moments like "we first see the people" the model without tef retrieves a video moment with people, but fails to retrieve the moment when the people first appear. Without the tef, the model has no indication of *when* a moment occurs in a video. Thus, though the model can identify if there are people in a moment, the model is unable to determine when the people first appear. Likewise, for moments like "train begins to move", the model without tef retrieves a video moment in which the train is moving, but not a moment in which the train begins to move.

2. RGB and Flow Input Modalities

In the main paper, we demonstrate that RGB and optical flow inputs are complementary. Here we show a few examples which illustrate how RGB and flow input modalities complement each other. Figure 3 compares a model trained with RGB input and a model trained with optical flow input (both trained with global video features and tef). We expect the model trained with RGB to accurately localize moments which require understanding the appearance of objects and people in a scene, such as "child jumps into arms of man wearing yellow shirt" (Figure 3 top row). We expect the model trained with flow to better localize moments which require understanding of motion (including camera motion) such as "a dog looks at the camera and jumps at it" and "camera zooms in on a man playing the drums" (Figure 3 row 3 and 4). Frequently, both RGB and optical flow networks can correctly localize a moment (Figure 3 bottom row). However, for best results we take advantage of the complimentary nature of RGB and optical flow input modalities in our fusion model.

[*] Work done at Adobe Research during LAH's summer internship

3. Qualitative Results for MCN

Figure 4 shows four videos in which we evaluate with fine-grained temporal windows at test time. Observing the plots in Figure 4 provides insight into the exact point at which a moment occurs. For example, our model correctly localizes the phrase “the blue trashcan goes out of view” (Figure 4 bottom right). The finegrained temporal segments that align best with this phrase occur towards the end of the third segment (approximately 14s). Furthermore, Figure 4 provides insight into which parts of the video are most similar to the text query, and which parts are most dissimilar. For example, for the phrase “the blue trashcan goes out of view”, there are two peaks; the higher peak occurs when the blue trashcan goes out of view, and the other peak occurs when the blue trashcan comes back into view.

In the main paper, running a natural language object retrieval (NLOR) model on our data is a strong baseline. We expect this model to perform well on examples which require recognizing a specific object such as “a man in a brown shirt runs by the camera” (Figure5 top row), but not as well for queries which require better understanding of action or camera movement such as “man runs towards camera with baby” (row 2 and 4 in Figure 5). Though the Moment Context Network performs well on DiDeMo, there are a variety of difficult queries it fails to properly localize, such as “Mother holds up the green board for the third time” (Figure 5 last row).

Please see `https://www.youtube.com/watch?v=MRO7_4ouNWU` for examples of moments correctly retrieved by our model.

4. Additional Baselines

In the main paper we compare MCN to the natural language object retrieval model of [3]. Since the publication of [3], better natural language object retrieval models have been proposed (e.g., [2]). We evaluate [2] on our data, in a similar way to how we evaluated [3] on our data in the main paper (Table 3 Row 5 in the main paper). We extract frames at 10 fps on videos in our test set and use [2] to score each bounding box in an image for our description. The score for a frame is the max score of all bounding boxes in the frame, and the score for a moment is the average of all frames in the moment. We expect this model to do well when the moment descriptions can be well localized by localizing specific objects. Surprisingly, even though CMN outperforms [3] for natural language object retrieval, it does worse than [3] on our data (Table 1 row 6). One possible reason is that [2] relies on parsing subject, relationship, and object triplets in sentences. Sentences in DiDeMo may not fit this structure well, leading to a decrease in performance. Additionally, [2] is trained on MSCOCO [1] and [3] is trained on ReferIt [5]. Though MSCOCO is larger than ReferIt, it is possible that the images in ReferIt are more similar to ours and thus [3] transfers better to our task.

	Baseline Comparison (Test Set)			
	Model	Rank@1	Rank@5	mIoU
1	Upper Bound	74.75	100.00	96.05
2	Chance	3.75	22.50	22.64
3	Prior (tef)	19.40	66.38	26.65
4	CCA	18.11	52.11	37.82
5	Natural Lang. Obj. Retrieval (SCRC [3])	16.20	43.94	27.18
6	Natural Lang. Obj. Retrieval (CMN [2])	12.59	38.52	22.50
7	Natural Lang. Obj. Retrieval (SCRC [3] re-trained)	15.57	48.32	30.55
8	Image Retrieval (DeFrag [4] re-trained)	10.61	33.00	28.08
9	MCN (ours)	**28.10**	**78.21**	**41.08**
	Ablations (Validation Set)			
10	MCN: Inter-Neg. Loss	25.58	74.13	39.77
11	MCN Intra-Neg. Loss	26.77	78.13	39.83
12	MCN	**27.57**	**79.69**	**41.70**

Table 1: MCN outperformes baselines (rows 1-8) on our test set. We show ablation studies for our inter-intra negative loss in rows 10-12.

Additionally, we train [4], which is designed for natural language image retrieval, using our data. [4] relies on first running a dependency parser to extract sentence fragments linked in a dependency tree (e.g., “black dog”, or “run fast”). It scores an image based on how well sentence fragments match a set of proposed bounding boxes. To train this model for our task, we also extract sentence fragments, but then score temporal regions based on how well sentence fragments match a ground truth temporal region. We train on our data (using a late fusion approach to combine RGB and optical flow), and find that this baseline performs similarly to other baselines (Table 1 row 8). In general, we believe our method works better than other baselines because it considers both positive and negative moments when learning to localize video moments and directly optimizes the R@1 metric.

5. Inter-Intra Negative Loss

In Table 1 we compare results when training with only an inter-negative loss, only an intra-negative loss, and our proposed inter-intra negative loss. Considering both types of negatives is important for best performance.

6. Importance of Language Feature

Because we ask annotators to mark any interesting moment and describe it, it is possible that annotators mark visually interesting moments which can be localized without text. We thus train a model with our temporal context features but no text query and observe that this model outperforms chance and the moment frequency prior, but does not perform as well as our full model (25.04, 75.23, and 36.12 on R@1, R@5, and mIoU metrics). This indicates that while understanding what constitutes a "describable" moment can be helpful for natural language moment retrieval, natural language is important to achieve best results on DiDeMo. Because the majority of videos include multiple distinct moments (86%), we believe the gap between model trained with and without language will improve with better video-language modelling.

7. Words Used to Construct Table 2

To construct Table 2 in the main paper, we used the following words:

- Camera words: camera, cameras, zoom, zooms, pan, pans, focus, focuses, frame, cameraman
- Temporal words: first, last, after, before, then, second, final, begin, again, return, third, ends
- Spatial words: left, right, top, bottom, background

Additionally, our vocab size is 7,785 words (which is large considering the total number of words in our dataset - 329,274).

8. Video Retrieval Experiment

We used our model to retrieve five moments closest to a specific text query in our shared embedding space from all videos in our test set (Figure 6). We find that retrieved moments are semantically similar to the provided text query. For example, the query "zoom in on baby" returns moments in which the camera zooms in on babies or young children. A similar query, "camera zooms in" returns example moments of the camera zooming, but the videos do not contain babies. Though the query "the white car passes by" does not always return moments with cars, it returns moments which include semantically similar objects (trains, busses and cars).

Please see `https://www.youtube.com/watch?v=fuz-UBvgapk` for an example of video retrieval results.

9. Annotation Ambiguity

Figure 7 shows an example in which the end point for specific moments are ambiguous. For the query "zoom in on man", three annotators mark the fourth segment in which the camera actively zooms in on the man. However, one annotator marks the segment in which the camera zooms in on the man and the following segment when the camera stays zoomed in on the man before zooming out.

This ambiguity informed how we chose our metrics. Based on the annotations for the query "zoom in on man", it is clear that the moment retrieved by our model should include the fourth segment. Though it is less clear if a moment retrieved by our model must include the fifth segment (which was only marked by one annotator to correspond to the phrase "zoom in on man"), it is clear that a model which retrieves both the fourth and fifth segment is more correct than a model which retrieves the third and fourth segment. When we compute a score for a specific example, we choose the maximum score when comparing the model's result to each four-choose-three combinations of human annotations. This results in scores which reflect the intuition outlined above; a model which retrieves only the fourth segment (and therefore agrees with most annotators) will get a higher score than a model which retrieves the fourth and fifth segment (which only agrees with one annotator). Additionally, a model which retrieves the fourth and fifth segment will receive a higher score than a model which retrieves the third and fourth segment.

Note that if two annotators had marked both the fourth and fifth segment, no retrieved moment would perfectly align with any four choose three combination of annotations. Thus, for some examples, it is impossible for any model to achieve a perfect score. In all our qualitative examples where we mark the "ground truth" moment in green, at least three annotators perfectly agree on the start and end point.

10. Distribution of Annotated Moments

Figure 8 shows the distribution of annotated start and end points in DiDeMo. Moments marked by annotators tend to occur at the beginning of the videos and are short. Though a "prior baseline" which retrieves moments which correspond to the most common start and end points in the dataset does much better than chance, our model significantly outperforms a "prior baseline".

References

[1] X. Chen, T.-Y. L. Hao Fang, R. Vedantam, S. Gupta, P. Dollr, and C. L. Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arxiv:1504.00325*, 2015.

[2] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko. Modeling relationships in referential expressions with compositional modular networks. In *CVPR*, 2017.

[3] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. In *CVPR*, 2016.

[4] A. Karpathy, A. Joulin, and F. F. F. Li. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*, 2014.

[5] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014.

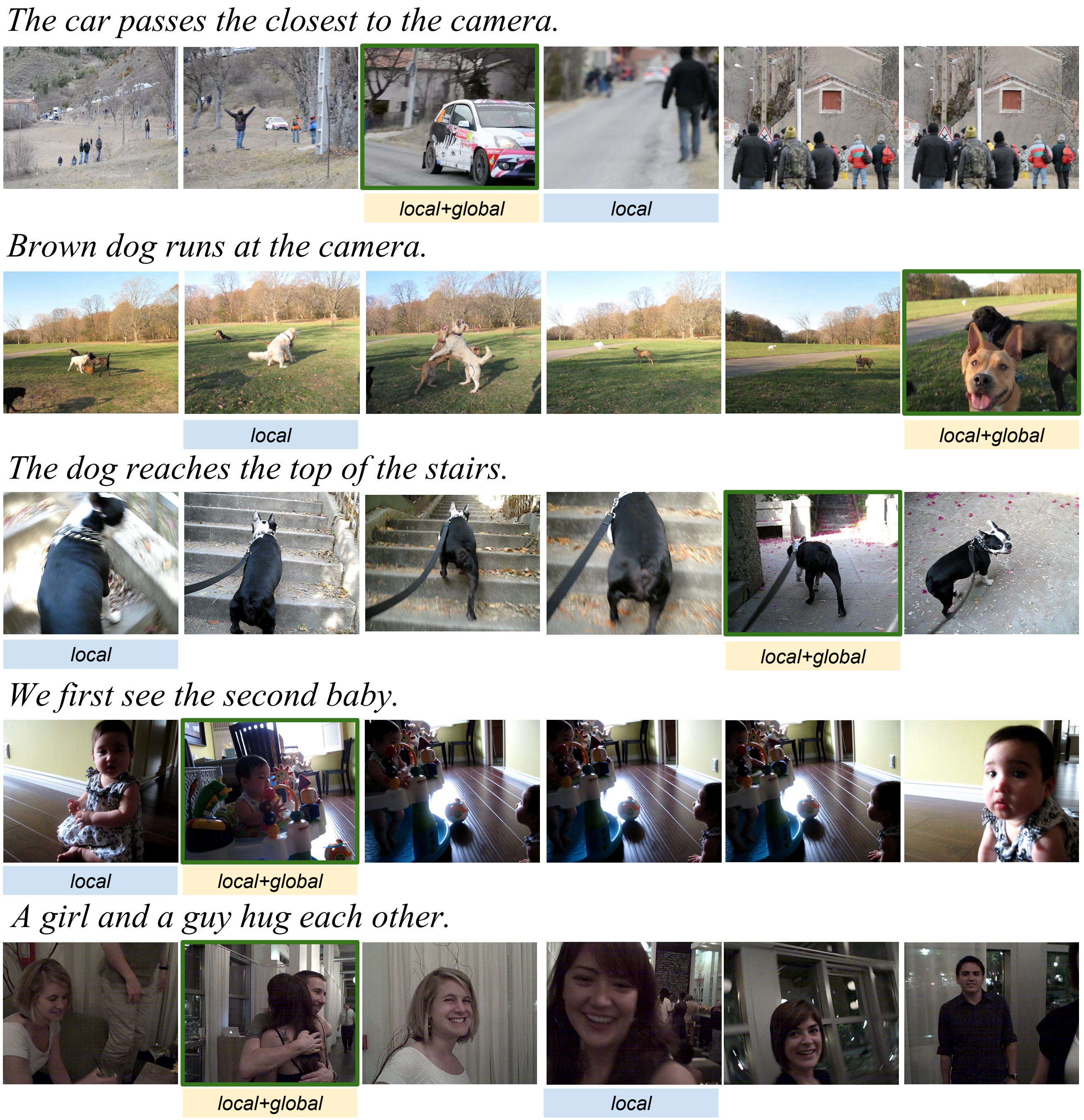

The car passes the closest to the camera.
local+global
local
Brown dog runs at the camera.
local
local+global
The dog reaches the top of the stairs.
local
local+global
We first see the second baby.
local
local+global
A girl and a guy hug each other.
local+global
local

Figure 1: Comparison of moments which are correctly retrieved when including global context, but not when only using local video features. Ground truth moments are outlined in green. Global video features improve results for a variety of moments. For moments like "the car passes the closest to the camera", it is not enough to identify a car but to understand when the car is closer to the camera than in any other moment. For moments like "brown dog runs at the camera", the model must not only identify when the brown dog is running, but when it runs towards the camera.

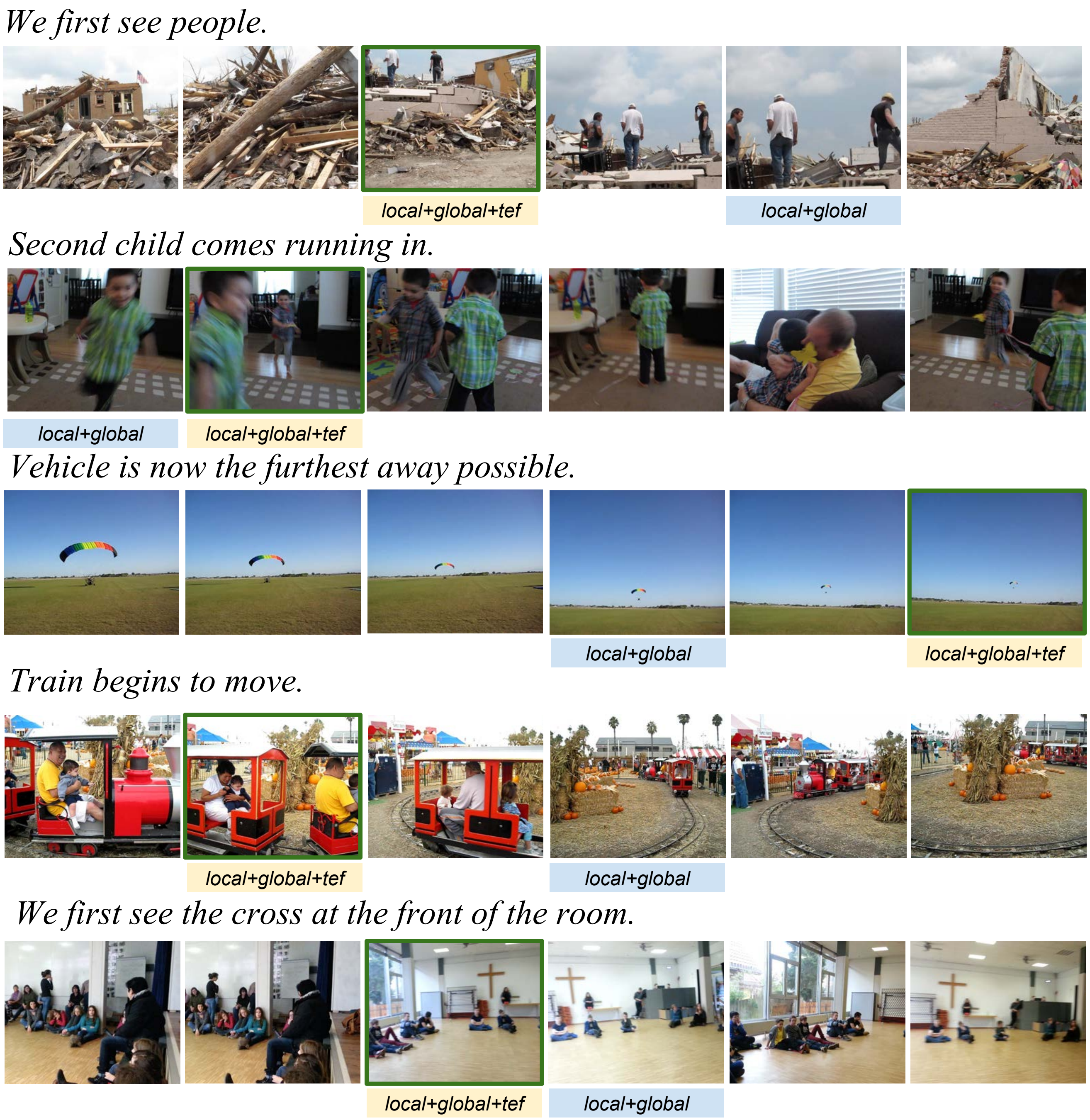


Figure 2: Comparison of moments which are correctly retrieved when including the temporal endpoint feature (tef), but not when only using local and global video features. Ground truth moments are outlined in green. For moments like "we first see the people" the model without tef retrieves a video moment with people, but fails to retrieve the moment when the people first appear. Likewise, for moments like "train begins to move", the model without tef retrieves a video moment in which the train is moving, but not a moment in which the train begins to move.

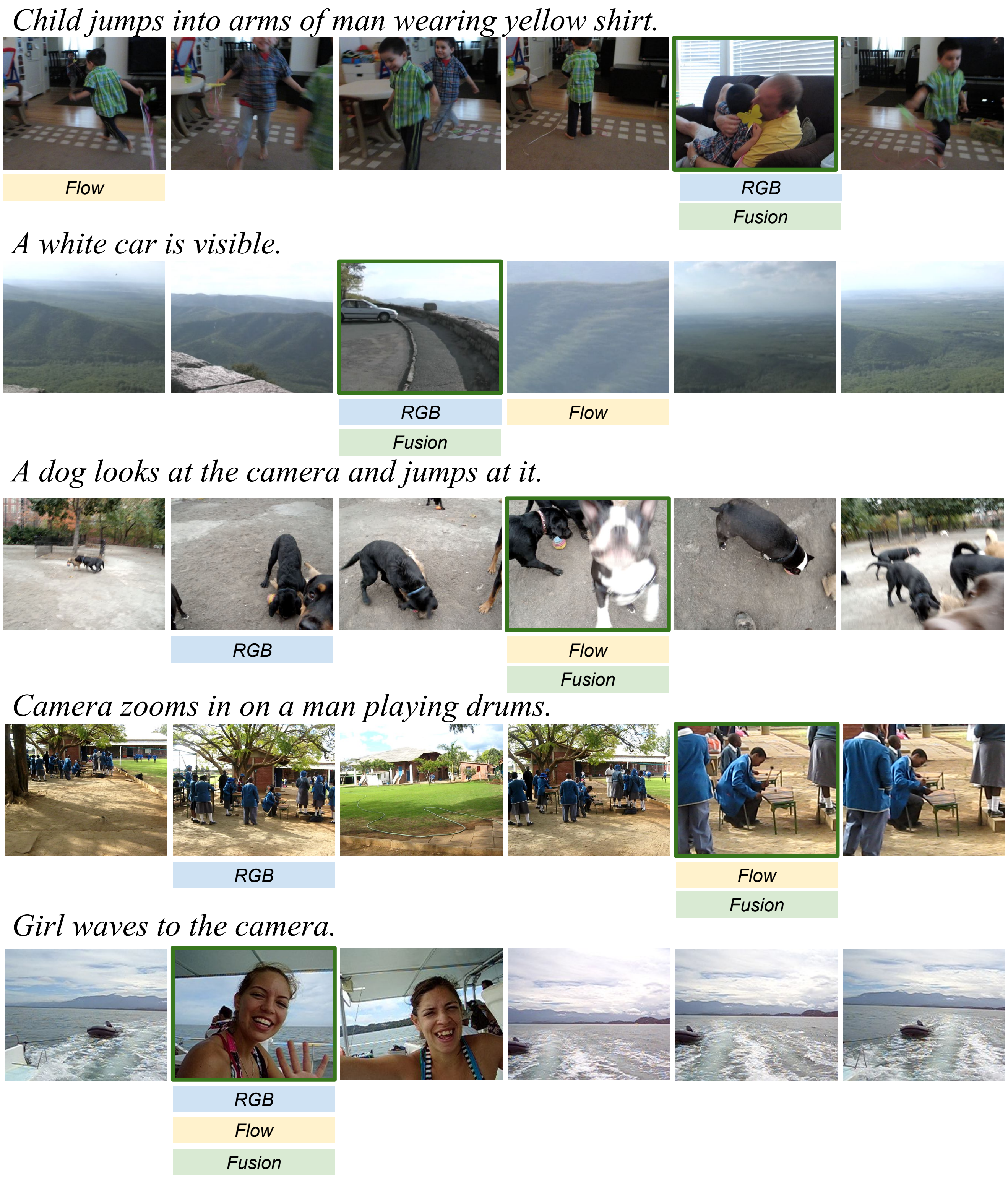
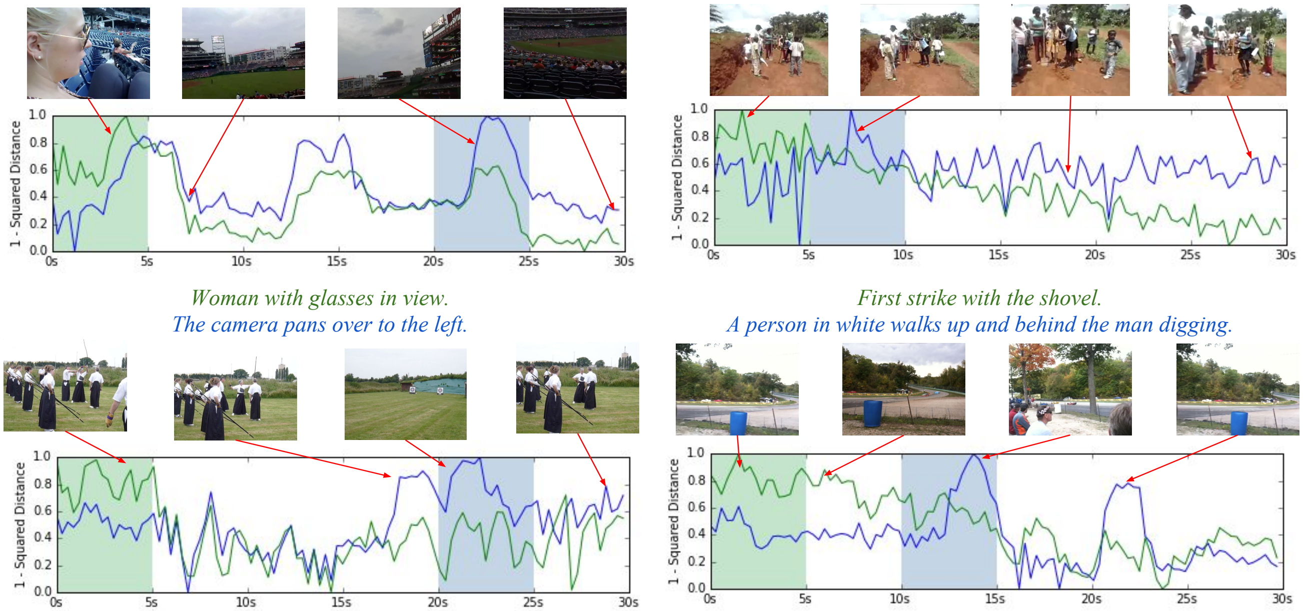

Child jumps into arms of man wearing yellow shirt.
Flow
RGB
Fusion
A white car is visible.
RGB
Fusion
Flow
A dog looks at the camera and jumps at it.
RGB
Flow
Fusion
Camera zooms in on a man playing drums.
RGB
Flow
Fusion
Girl waves to the camera.
RGB
Flow
Fusion

Figure 3: Comparison of moments retrieved using different input modalities (ground truth marked in green). For queries like "A white car is visible" which require recognizing an object, a network trained with RGB performs better whereas for queries like "Camera zooms in on a man playing drums" which require understanding motion, a network trained with optical flow performs better. For some queries, networks trained with either RGB or optical flow retrieve the correct moment.



Woman with glasses in view.
The camera pans over to the left.

First strike with the shovel.
A person in white walks up and behind the man digging.

Person raises bow above head.
The target is shown, then the people.

First car to make it around the curve.
The blue trashcan goes out of view.

Figure 4: Comparison of similarity between text queries and finegrained temporal segments. Though ground truth annotations correspond to five second segments, evaluation with more finegrained segments at test time can provide better insight about where a moment occurs within a specific segment and also provide insight into which other parts of a video are similar to a given text query.

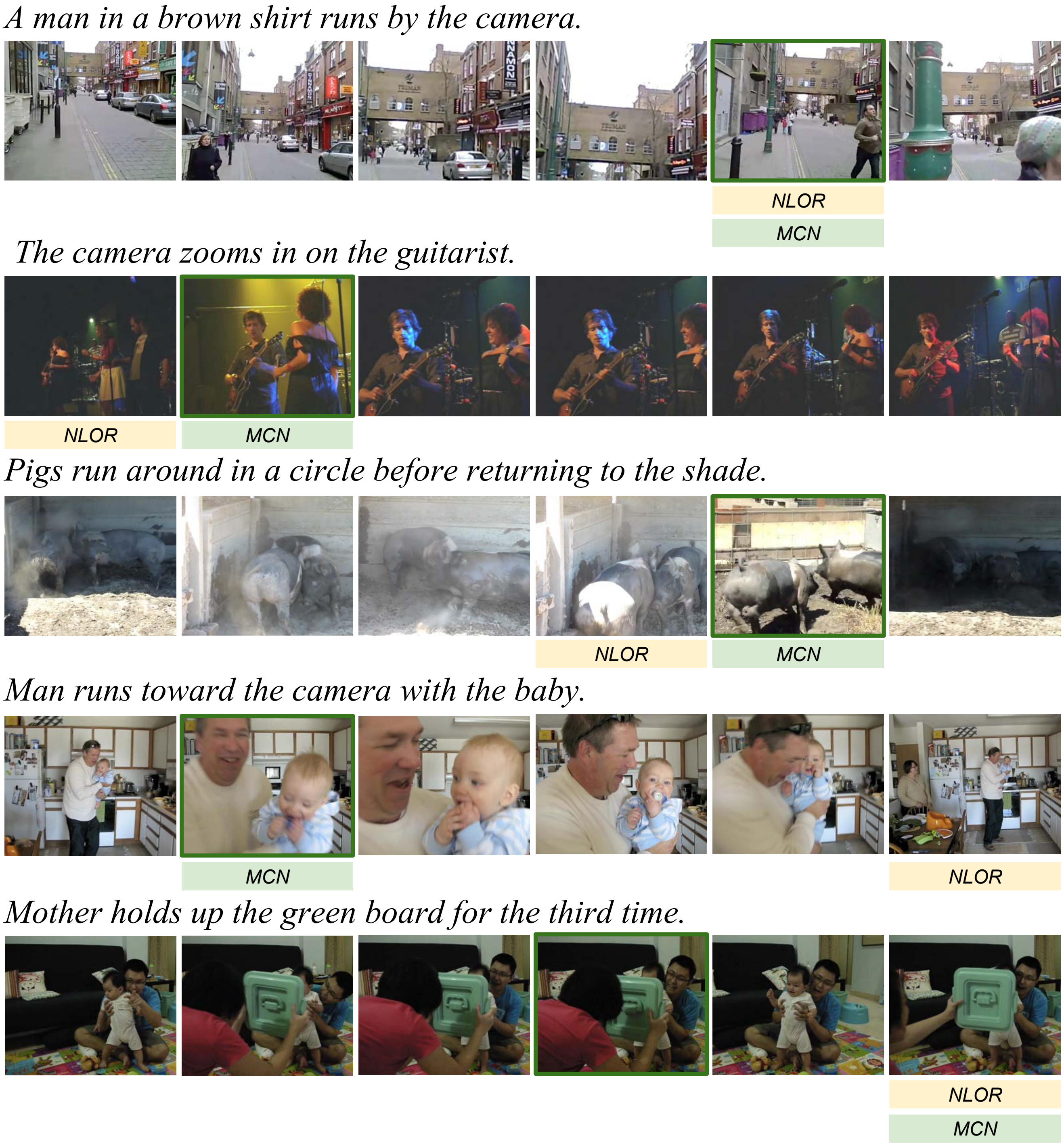

A man in a brown shirt runs by the camera.
NLOR
MCN
The camera zooms in on the guitarist.
NLOR
MCN
Pigs run around in a circle before returning to the shade.
NLOR
MCN
Man runs toward the camera with the baby.
MCN
NLOR
Mother holds up the green board for the third time.
NLOR
MCN

Figure 5: We compare our Moment Context Network (MCN) model to a model trained for natural language object retrieval (NLOR). We expect a model trained for natural language object retrieval to perform well when localizing a query relies on locating a specific object (e.g, a man in a brown shirt). However, in general, the MCN model is able to retrieve correct moments more frequently than a model trained for natural language object retrieval. DiDeMo is a difficult dataset and some queries, such as “mother holds up green board for third time” are not correctly localized by the MCN.

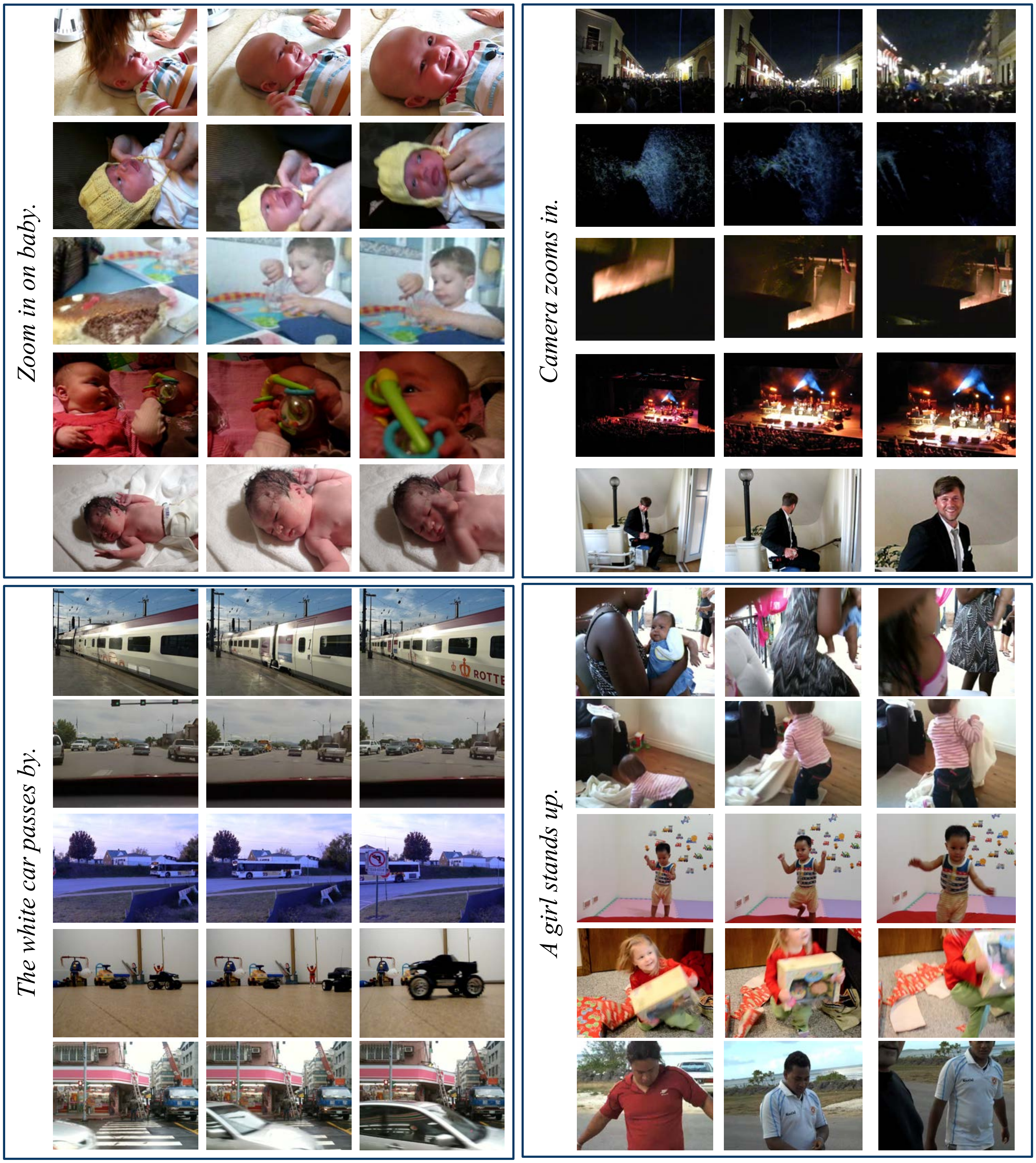


Figure 6: We use our model to retrieve the top moments which correspond to a specific query from the entire test set. Though MCN was not trained to retrieve specific moments from a set of different videos, it is able to retrieve semantically meaningful results. Above we show the top five moments retrieved for four separate text queries. A video showing retrieved momenents can be found here: `https://www.youtube.com/watch?v=fuz-UBvgapk`.

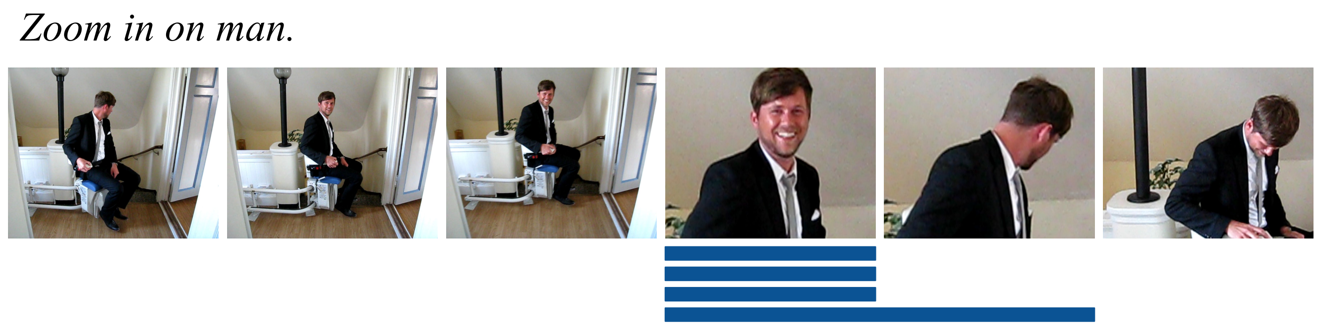


Figure 7: Humans do not always perfectly agree on start and end points for a moment. In the above example we show annotations (denoted as blue lines) from four separate crowd-sourced annotators. Though three annotators agree that the moment corresponds to the fourth segment, a fourth annotator believes the moment corresponds to both the fourth and fifth segment. Our metrics reflect this ambiguity; a model which retrieves only the fourth segment will receive a high score. A model which retrieves both the fourth and fifth segment will receive a lower score, but it will receive a higher score than a model which retrieves the third and fourth segments (which no annotators marked as the correct start and end point).

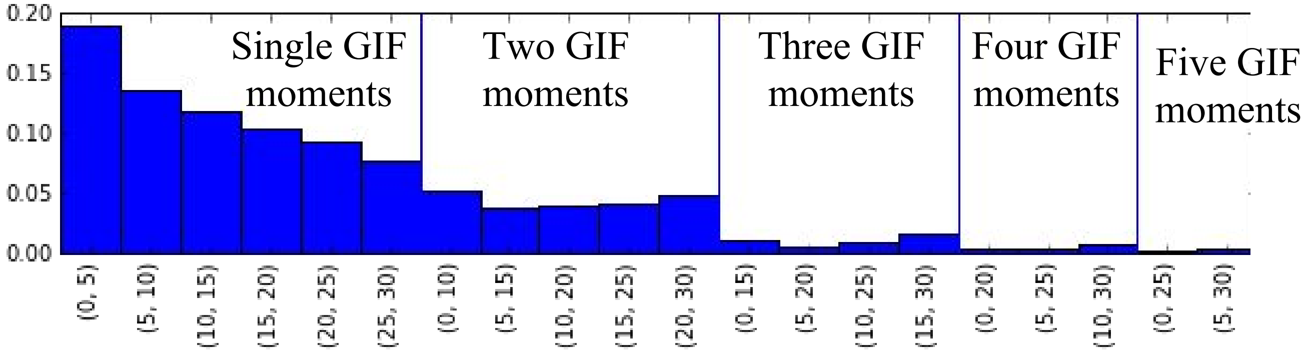


Figure 8: Distribution of segments marked in DiDeMo. Moments tend to be short and occur towards the beginning of videos.