

---

# Appendix for

## QVHIGHLIGHTS: Detecting Moments and Highlights in Videos via Natural Language Queries

---

**Jie Lei Tamara L. Berg Mohit Bansal**  
Department of Computer Science  
University of North Carolina at Chapel Hill  
{jielei, tlberg, mbansal}@cs.unc.edu

### A Additional Results

**Performance breakdown by video category.** In Table 1, we show model performance breakdown on the 3 major video categories: daily vlog, travel vlog and news.

Table 1: Performance breakdown by video category, on QVHIGHLIGHTS *test* split. We highlight the best score in each column in **bold**, and the second best score with underline. All models are trained from scratch.

Method	Moment Retrieval   R1 IoU=0.5			Highlight Detection   HIT@1		
	daily (46.5%)	travel (43.1%)	news (10.4%)	daily (46.5%)	travel (43.1%)	news (10.4%)
BeautyThumb [5]	-	-	-	24.13	17.44	20.62
DVSE [4]	-	-	-	21.90	21.50	22.50
MCN [2]	8.23	14.44	13.12	-	-	-
CAL [1]	24.83	26.92	22.50	-	-	-
XML [3]	45.05	40.45	33.12	<u>58.58</u>	53.08	<u>49.38</u>
XML+	49.37	46.62	35.00	57.18	54.44	48.12
Moment-DETR	<u>51.80</u>	<u>56.57</u>	<u>42.50</u>	56.15	<u>56.93</u>	47.62
Moment-DETR w/ PT	<b>63.22</b>	<b>59.08</b>	<b>48.63</b>	<b>60.27</b>	<b>61.95</b>	<b>51.75</b>

**Ablations on #moment queries.** In Table 2, we show the effect of using different #moment queries. As can be seen from the table, this hyper-parameter has a large impact on moment retrieval task where a reasonably smaller value (e.g., 10) gives better performance. For highlight detection, the performance does not change much in terms of mAP, but HIT@1 favors smaller number of moment queries as well. Considering performance of both tasks, our best model use 10 moment queries.

Table 2: Ablations on #moment queries on QVHIGHLIGHTS *val* split.

#Moment Queries	Moment Retrieval			Highlight Detection (>=Very Good)	
	R1@0.5	R1@0.7	mAP avg	mAP	Hit@1
5	<b>54.90</b>	<u>34.06</u>	<u>31.08</u>	<u>36.04</u>	<b>57.03</b>
10	<u>53.94</u>	<b>34.84</b>	<b>32.20</b>	<u>35.65</u>	55.55
20	47.94	29.10	24.81	<b>36.34</b>	<u>55.94</u>
50	39.81	21.16	18.47	34.96	53.48
100	41.16	21.68	19.51	34.52	51.87

**Saliency loss ablations.** As described in main text Equation 3, Moment-DETR’s saliency loss consists of two terms, one term that distinguishes between high and low score clips (i.e.,  $t_{high}$ ,  $t_{low}$ ),

another term distinguishes between clips in and outside the ground-truth moments (i.e.,  $t_{in}$ ,  $t_{out}$ ). In Table 3, we study the effect of using the two terms. We notice that adding one of them improves the model performance across all metrics, while the term  $(t_{in}, t_{out})$  typically works better. Overall, the best performance is achieved by using both terms.

Table 3: Ablations on saliency loss on QVHIGHLIGHTS *val* split.

Saliency Loss Type	Moment Retrieval			Highlight Detection ( $\geq$ Very Good)	
	R1@0.5	R1@0.7	mAP avg	mAP	HIT@1
None	44.84	25.87	25.05	17.84	20.19
$(t_{in}, t_{out})$	<u>52.90</u>	<b>36.32</b>	31.46	<u>35.62</u>	<u>52.58</u>
$(t_{high}, t_{low})$	52.52	33.16	30.35	29.32	40.77
$(t_{in}, t_{out}) + (t_{high}, t_{low})$	<b>53.94</b>	<u>34.84</u>	<b>32.20</b>	<b>35.65</b>	<b>55.55</b>

**More prediction examples.** We show more correct predictions and failure cases from our Moment-DETR model in Figure 1 and Figure 2.

## B Additional Data Analysis and Collection Details

**Distribution of saliency scores.** In Table 4, we show the distribution of annotated saliency scores. We noticed 94.41% of the annotated clips are rated by two or more users as ‘Fair’ or better (i.e.,  $\geq 3$ , meaning they may be less saliency, but still relevant, see supplementary file Figure 6). Only 0.96% of the clips have two or more users rated as ‘Very Bad’. This result is consistent with our earlier moment verification experiments.

Table 4: Distribution of annotated saliency scores on QVHIGHLIGHTS *train* split. Since we have scores from 3 users, we show the percentage as two or more users agree on a certain setting, e.g., at least two users agree that 5.59% of the clips should be rated with a score lower than or equal to ‘Bad’.

Score	=1 (Very Bad)	$\leq 2$ (Bad)	$\leq 3$ (Fair)	$\leq 4$ (Good)	$\leq 5$ (Very Good)
Percentage of Clips	0.96	5.59	23.44	62.10	100.00

**Annotation Instructions and Interfaces.** To ensure data quality, we require workers to pass our qualification test before participating in our annotation task. We show an example question from our qualification test in Figure 3. Our data collection process consists of two stages: (1) query and moment annotation, we show its instructions and annotation interface in Figure 4 and Figure 5, respectively; (2) saliency score annotation, we show its instructions and interface in Figure 6.

## C Content, License and Usage.

Our data<sup>1</sup> and code<sup>2</sup> are publicly available at [https://github.com/jayleicn/moment\\_detr](https://github.com/jayleicn/moment_detr). Additionally, this dataset should be used for research purposes only and not be used for any purpose (e.g., surveillance) that may violate human rights. The videos in the dataset are collected from a curated list of non-offensive topics such as vlogs and news. We use these YouTube videos under the Fair Use.<sup>3</sup> Our study was conducted on Amazon Mechanical Turk (AMT), based on an IRB application approved by our university IRB officials. The collected data via AMT does not contain any personally identifiable information.

## References

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

<sup>1</sup>CC BY-NC-SA 4.0 license <https://creativecommons.org/licenses/by-nc-sa/4.0/>

<sup>2</sup>MIT license <https://opensource.org/licenses/MIT>

<sup>3</sup><https://www.copyright.gov/fair-use/>

- [2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017. [1](#)
- [3] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *ECCV*, 2020. [1](#)
- [4] Wu Liu, Tao Mei, Yongdong Zhang, Cherry Che, and Jiebo Luo. Multi-task deep visual-semantic embedding for video thumbnail selection. In *CVPR*, 2015. [1](#)
- [5] Yale Song, Miriam Redi, Jordi Vallmitjana, and Alejandro Jaimes. To click or not to click: Automatic selection of beautiful thumbnails from videos. In *CIKM*, 2016. [1](#)



Figure 1: Correct predictions from Moment-DETR. Predictions are shown in **solid red boxes or lines**, ground-truth are indicated by **dashed green lines**.



Figure 2: Wrong predictions from Moment-DETR. Predictions are shown in **solid red boxes or lines**, ground-truth are indicated by **dashed green lines**.



(a) (b) (c)  
Query: Mother and daughter sit on the side of the road.

Question: Based on the query, which one is the best as a highlight for the video?

A (a)

B (b)

C (c)

Figure 3: Example question from our qualification test.

## Write a query and locate it in the video.

[Detailed Instructions & Examples \(click to show/hide\)](#)

You will be showing a 2-3 minutes long video. Your task is to find and write a sentence that describes a salient visual event in the video, and select all video segments/clips related to this event.

### Steps:

- Click to watch the video, then write a sentence (aka. **event query**) to describe one of the main events in the video.
  - An event can be:
    - activities of people or animal, for example, **A man in blue top is surfing.**
    - or anything else that are **visually salient and important** in the video. While describing what you hear from the video is also acceptable, we encourage you to always **describe events that can be seen.**
  - It should be relevant to  $\geq 10\%$  of the video segments. After hitting the submit button, our system will automatically notify you if it is too short. Note that the whole video can be relevant to your query, in which you need to select all the segments.
  - The description should be a **single sentence** written in standard English, and contains at least 5 words.
  - Be specific, avoid general and boring ones** like "Two people are talking" (X), "people vlog their day" (X), "adventure to a hotel" (X).
  - Write different events for different videos**, repetitions should be avoid.
  - Please describe events located at different parts of the videos, that is, **do not always describe events at the beginning.**
  - Some videos are in **foreign languages**, you can describe what is happening from what you see.
- Select **all** the video segments below that are **relevant to this event**.
  - The long video is split into 2-seconds long segments and are shown in the selection area below.
  - Select all query-relevant video segments.** They can be consecutive or non-consecutive.

Examples with explanations (**Read/Understand all of them helps you to get a higher approval rate!**):



Query-relevant clips/segments are circled by **red-boxes**. Other symbols are used in the explanations.

**Event Query: Mother tries to help daughter walk down the sidewalk.**

Event Explanation:	<b>correct</b>	Good! It describes a visually salient event with some details.
Relevant Explanation:	<b>correct</b>	Good! The selected clips (by red boxes) all match the event query, and no irrelevant clips are selected. Note that the few clips circled by the thick yellow boxes should not be selected, as they have stopped walking and started playing with dogs.

You may need to scroll inside the text boxes above to see all the text.

Last example **1 / 4** Next example

Figure 4: Annotation instructions (with some examples) for collecting queries and moments.

## Write a query and locate it in the video.

[Detailed Instructions & Examples \(click to show/hide\)](#)



**Steps:** Make sure to use your mouse over  to see tips or additional requirements.

1.  Click to watch the long video on the left, then write a sentence (aka. **event query**) to describe one of the main events in the video. We encourage **visually salient events** that you can see instead of events in the dialogue. Some videos are in foreign languages. 

Event Query:

2. Select **all** the video segments in the selection area below that are **relevant to this event**. Please select all clips that are relevant, and do not select irrelevant clips. 

For **news videos**, please do **not** focus on scenes where reporters or hosts are reporting. For example, 'A car drive through a flooded road in a heavy rain' is much better than 'A reporter is reporting in the rain'.

Please use the latest Google Chrome browser for this task.

**Selection Area**

**Step 2. Relevant Selection**  → Select all **query-relevant** video segments (**>= 3 segments**). Click, hold and drag allows you to select/de-select multiple segments.

Event Query:



Figure 5: Annotation interface for collecting queries and moments. A short version of the instruction is also shown, the complete instructions (shown in Figure 4) can be viewed by clicking the top green button *Detailed Instructions & Examples (click to show/hide)*.

## Rate the clips based on query

**Examples:**

Event Query: Whale sharks swim in the ocean.



**Good**  
(event clearly showing, but not aesthetically very good)



**Fair**  
(relevant but whale shark too small)



**Very Bad**  
(completely irrelevant)

Event Query: Mother tries to help daughter walk down the sidewalk.



**Very Good**  
(event clearly showing, aesthetically good)



**Fair**  
(relevant but largely occluded)



**Bad**  
(somewhat relevant, mother, daughter, sidewalk are showing, but sitting not walking down)

**Instruction:**  
In this HIT, we will present you a sentence query that describes one or more events, and multiple clips from a video (these clips might be non-consecutive). For each clip, given the query, your task is to decide whether it is good to be used as a cover/highlight for the video. See examples on the left. There are 5 options, from *Very Good* to *Very Bad*:

Is the clip good as a cover or highlight for the video?

Very Good

Good

Fair

Bad

Very Bad

Relevant, eye-catching, salient, clearly visible...

Relevant, small, occluded, less interesting...

Irrelevant to query

(1) A **Very Good** clip should be relevant to the query, clearly showing the query events, objects, people. It should also look attractive and beautiful to people, as we are gonna use it as a cover. (2) A **Fair** clip is still relevant to the query, but the query event or object/people are small, occluded, the video quality is lower compared to a "Very Good" clip. (3) A **Very Bad** clip means it is completely irrelevant to the query.

Please use the latest Google Chrome browser since we only tested on it.

---

Event Query:






Is this clip good as a cover or highlight?

**Very Good**

Good

**Fair**

Bad

**Very Bad**

<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>




Is this clip good as a cover or highlight?

**Very Good**

Good

**Fair**

Bad

**Very Bad**

<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Submit

Figure 6: Annotation instructions and interface for saliency score annotation.