# Moment Detection in Long Tutorial Videos - Supplementary Material

In this supplementary material we provide additional information and results for LongMoment-Detr and our datasets. We begin by showing a visual representation of our pipeline, then some additional ablations (Sec. 2), following with some additional clarifications on several components (Sec. 3) along with more dataset details (BMD in Sec. 4 and YTC in Sec. 5).

## 1. Visual representation

In Fig. 1 you can see an overview of our system. Initially, we source our videos from the Behance platform. Once obtained, these videos undergo an automatic transcription process powered by the Azure speech recognition system. This raw transcript is then segmented, and split into segments. Then, we use GPT-3 to generate queries for the segments by summarizing the transcript corresponding to the segment time-span and finally we train our model.

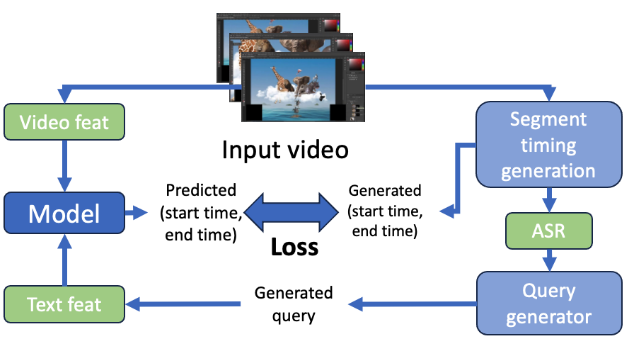


Figure 1. **System overview of LongMoment-Detr.** The process involves: 1) feature extraction, 2) Automatic transcript generation (ASR) using speech recognition, 3) Segment generation, 4) Summarization of transcripts for each segment, and 5) Training of model

## 2. Ablations

In this section, we provide additional ablations for our method, starting with extra results for the influence of different components.

### 2.1. Influence of different components

In Tab. 1, we present an overview of the influence of different components on the validation split for easier comparison (results on the testing split can be found in the main paper). We observe that both the segment timing generation and the query generation have a strong impact on performance (similar to what is presented in the main paper). We obtain the best results by combining the timing generation from OSG [5] with the GPT3 [1] query generation which are used by our method.

In Tab. 2, we present additional zero shot results of different models on the YTC dataset. We observed that both segment timing generation and query generation have an influence on the performance which is similarly to BMD. However, for YTC dataset the segment timing generation has a stronger influence.

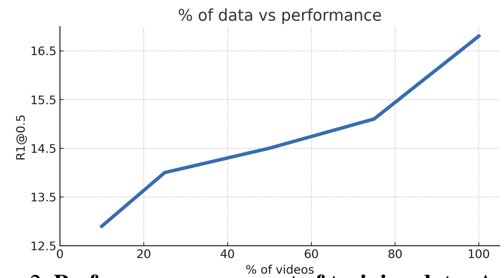


Figure 2. **Performance vs amount of training data.** As it can be seen, the performance increases with the amount of data.

| Component | Segments | Queries | $R1@0.5 \uparrow$ | $R1@0.7 \uparrow$ |
|---|---|---|---|---|
| Baseline [2] | No | No | $3.1_{\pm 0.4}$ | $0.5_{\pm 0.2}$ |
| Random seg | Random | No | $9.4_{\pm 1.6}$ | $2.6_{\pm 0.8}$ |
| ShotDetect | OSG | No | $13.4_{\pm 0.5}$ | $6.3_{\pm 0.3}$ |
| LongMoment-Detr | OSG | GPT3 | $16.8_{\pm 0.5}$ | $9.2_{\pm 1.0}$ |

Table 1. **Effect of different components on performance.** Both the segment timing generation and query generation have a strong impact on performance. Hence, in the final model, we use OSG and GPT3, thus obtaining our final model LongMoment-Detr. The results are presented on the validation split.

| Model | Segments | Queries | $R1@0.5 \uparrow$ | $R1@0.7 \uparrow$ |
|---|---|---|---|---|
| Baseline | No | No | $0.9_{\pm 0.3}$ | $0.4_{\pm 0.1}$ |
| Random sed | Random | No | $2.0_{\pm 0.1}$ | $0.5_{\pm 0.1}$ |
| Query gen | Random | GPT3 | $2.8_{\pm 0.9}$ | $0.9_{\pm 0.2}$ |
| ShotDetect | OSG | No | $4.8_{\pm 0.8}$ | $2.0_{\pm 0.4}$ |
| LongMoment-Detr | OSG | GPT3 | $5.0_{\pm 2.4}$ | $2.2_{\pm 1.5}$ |

Table 2. **Zero shot results on YTC.** For the YTC dataset we observe that the biggest influence on performance comes from using an automatic video segmentation method like OSG. However, by using the queries obtained from GPT3 the performance further increases.

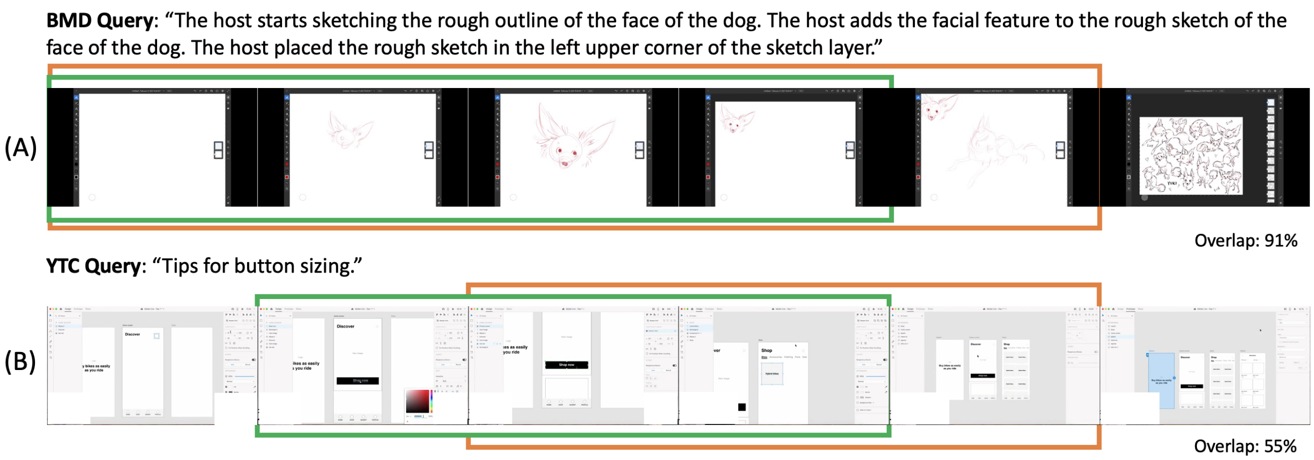


Figure 3. **Qualitative examples for BMD (A) and YTC (B).** Along with the query, we show several video frames, the prediction results in orange and the ground truth in green. We also specify the overlap between the prediction and ground truth segments.

As the person stands, sits on their shoulders. And if the person is big. Let's not. That's going to be not going to balance very well. So let's say the bird has a slightly longer neck. Or if yeah, maybe I could get a bird with a longer neck. But this is though these are wings, so the person sits here. In the Heights that I have to work with would be. The length of the neck. Which is not the majority. Of the. Hurt us. Body mass, which makes them very. I guess this could be. It's like it's like a horse, but higher. Even this. Is making the bird look very? Morpholine I think the fruit has to be bigger to comfortable Leafly something like this. At least by looks, I'm sure there's no way a bird can fly like this. I'll cut this out into a new. Layer so that I can't draw over it. Hello welcome. The legs have to be decently long worlds. This bird cancer will have a really short and pathetic gate. When it's not flying. But not too long that that person is super high open. Unreachable. Sticks out from the top. Of the. The the top of the headline. If everyone else it would be like up here. It makes some really easy to shoot. OK, I'm going to do something that I don't want to. Which is looking at a old. Each. Other night roll the birds that they arrived in mattress and in. We need to see if they are usable. Or like I need to see. The sizes are believable. OK, these birds are much bigger. It makes sense their long distance they should be. Able to glide. They should fly or stay airborne with ease, but not so much. Might be able to maneuver. Well or. Passe. Aline hit box. Whatever that means. So OK, I don't need to reference this. Again, just make a new bird for for him. OK. I duplicated this layer so I am Oh my God. I forgot to turn on the Tilt Tool timeline the whole time. Latest special than never. Duplicate the layer I'm going to draw this into a silhouette and see. What kind of silhouettes I can come up with? Before I decide to. Settle on a shape. Attack bird. It should probably be a Hawk, right? If it's a. If it's something that I'm considering. For. Left 4 not just. Transportation. They can walk around pretty well. Their legs are pretty. Robust. They can step on anything. They can fly around pretty. On pretty well. Let's see Hawks. Need to know the shapes of 1. So yeah, he's kind of like. Speak legs may have little bit of fluff behind there. Behind their drumsticks, which is nice. I wonder wings or whole digits. Not very nice. Or it could be not as neat as a kitchen. It's like a mess here. Their heads are a little short. I also stand pretty upright. I don't know how that changes held comfortably. Person can sit on them. Thanks. OK, what can they wear? They shouldn't wear anything on their heads because it would block the person's view. We shouldn't have a Crown either. Maybe a little tip like this. Maybe they can have those secretary bird. Whiskers they're so pretty. But it would just block their way as well if they want to like reach in front or something. Beardsman OK, I'm going to play this game of how big can I make the person without it feeling like it's going to tax the bird? Too much. Wait, I can draw a horse silhouette? And then try to compare. Like you have these, if this person is walking among horses. I would want the person to be. About the same head Heights as everyone else. Don't make this a different color. Out my eyes. OK, not bad. I'll make the bird about the same size as the horse, and I'll use that to gauge. Yeah, the horse will have to be smaller or the bird has to be bigger then. A horse. So I think this is a decent size like. From shoulder down to the is the length of the neck of the burden, and I'd rather rest based on the proportion. So the head can stick out from. On top of the head of the birds should be able to see everything. Around them. Impressively dedication to believability invention but homeless. Oh man, I want some. I. I want to say I don't really care for believability. A lot of times. There are very basic things from my character design. How these people survive with such long hair and such long scarves. And I'm just like, yeah? It's fantasy, but I I guess when I can. I do care about it. But I. I'm just as easily willing to throw it out if it means that I can. Make something look more beautiful. I guess things like say, them being able to see over the bird's head is important because this is a comic and it will become really evidenced in like two or three panels of different angles, where if I draw from the front of them and there the big bird phase just completely covers of the persons. Head. It creates problems really fast, so those things I try to. I try to. The rain and the fantastical elements. So a little bit of armor on the bird's legs. OK, nice. Wouldn't it be nice if I could get a 3D model of this bird and I don't have to struggle with the. But the perspectives and everything. OK, I can't have these spikes in the back 'cause when it flies. To step its own but. What the birds do when they fly with their legs. They do something like this, right? Flying chicken from. From Mulan so I can't put armor here. But I can lose. Most plates in front serve the bird. I remember when I drew the ferry. Within like a couple pages. Of me drawing her with all the long and claws and stuff I was like Oh no. She can't form a fist and. Every page where I had her grasping on to something I was just like trying really hard to make it us. Noticeable as possible. Just try not to get into those situations again. This time not just the silhouette. Need to know how they felt their wings. Would help a lot if I did some studies on how these wings fold. I would have to do it a lot. So they don't have. Catholic. Three layers like this. And then there's some stuff sticking out here in there. I don't think I should hang too much stuff on it. I want to cover up the birds eyes some so that I don't have to draw it since I take up a lot of attention. But I'm not sure. It's doing that is a good idea too. Or those little tiny little helmets. Nighthawks where? OK, if I draw the. If I draw the ornaments of this little place on top of its head. Sound works I can. Have it stick out or lots and not block fishing and it can. Floxy the ice. Oh man, I need a real settle. So let's settle is. Part of clothing patterns. Material on the back. Of the animal. Strap here for the like breasts. Put the rest. And then on top is another layer and then the seats with a little back. So I really need to do is to draw the seat. The seat needs to be able to adjust for when the bird is flying so the person is sitting like this. Maybe it just means it's not. That crucial. Front I haven't had to do this in a while. And I just remember that every time I have to design something new. I spend like. I don't know 50 pages of the comic drawing them in panels. By the time I get used to drawing them, and I learn to draw them correctly, I wouldn't need to draw them anymore. Think they need to be able to sit back a little more. So that the bird doesn't get us much shoulder from flying. Just wait around. Or maybe what I need? Is. Little weights on the back of here to counterbalance the weight of the person so that. Even though the bird is carrying. Iload. That it's balanced across its back when it's flying so windy. Weird. It wouldn't feel unbalanced. It wouldn't have to strain specifically the the front part. And nothing else. Nice, I'm happy for you. OK, I haven't touched a person yet.

Transcript for a 25 min long segment

The tutorial is for designing a bird that a person can ride. The bird needs to be strong enough to carry the person, but also be able to fly and maneuver well. The person should be able to sit comfortably on the bird, and the bird should be able to see over the person's head. The bird should also be able to wear some armor.

Query generation for a 25 min long segment

Starts sketching the birds adding them shades and details by outlining and adding some color to the sketch of a bird, also sketch a man sitting on the bird. Gives more texture to the bird on which a man is sitting.

Human annotation for a 25 min long segment

Figure 4. **Visual comparison of transcript length (left side), query generation (upper right side) and human query (bottom right side).** As it can be seen, the length of the transcript for a 25 min segment is significantly longer.

### 2.2. Quantity of data

For this experiment, we study how the quantity of training data affects performance. Fig. 2 indicates that the more data we have, the better. This aspect further validates our approach of automatically generating segment annotations without incurring a large annotation cost.

## 3. Additional information

In this section, we provide more details on various design choices, starting with how we generate random segments. Then, additional details about the low-level adaptations of Moment-DETR [2] are provided. Further, we present how we use LLMs to summarize the transcripts and present some statistics about the query length.

### 3.1. Video Segmentation-Random

In Sec. **5.2** from the main paper we compared against a random segmentation baseline. Now, we will provide further explanation on how we randomly split into segments each video. We start by choosing a random duration for the first segment between 800 and 3000 seconds and then we continue doing this for the rest of the video. By using this approach, we ended up with an average of around 5 non-overlapping segments per video.

### 3.2. Adjusting Moment-DETR

As stated in the main paper, we started from Moment-DETR [2] and made several technical adjustments for the model to process longer videos. The original code did not work "out of the box" for our videos, since it assumes the videos are shorter than 3 minutes. Firstly, we removed the original constraint to trim the video to three minutes. Fur-

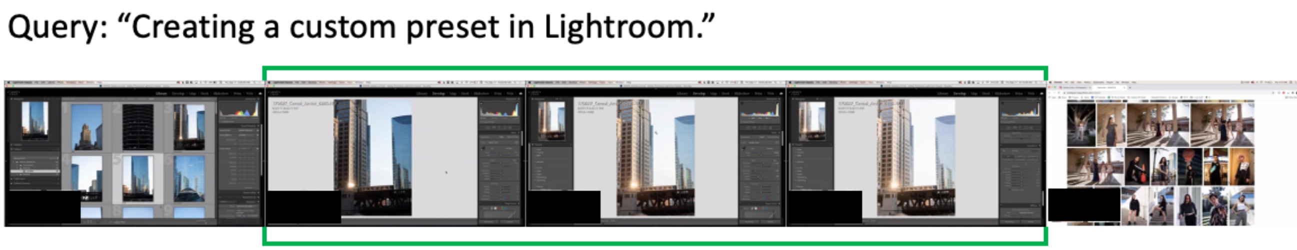


Figure 5. **YouTube Chapters example.** We collected the chapter annotations from YouTube for some long tutorial videos.

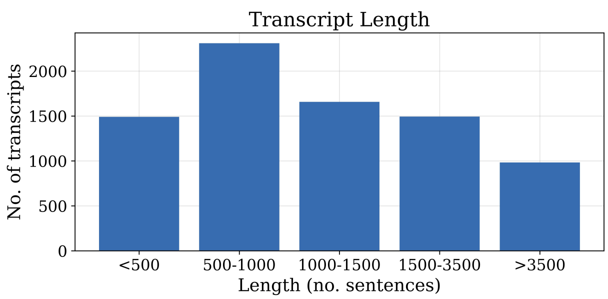


Figure 6. **Histogram of transcript length.** We present the transcript length in number of sentences per video on the BMD dataset.

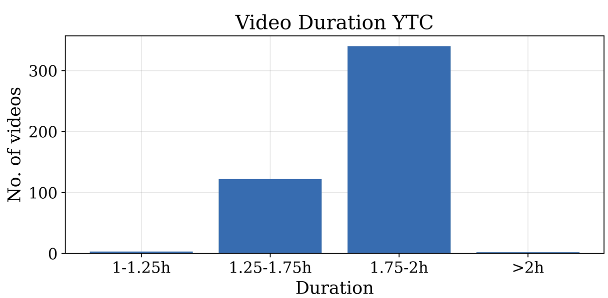


Figure 7. **Histogram of video duration YTC.** The majority of the videos from YTC have around 2 hours.

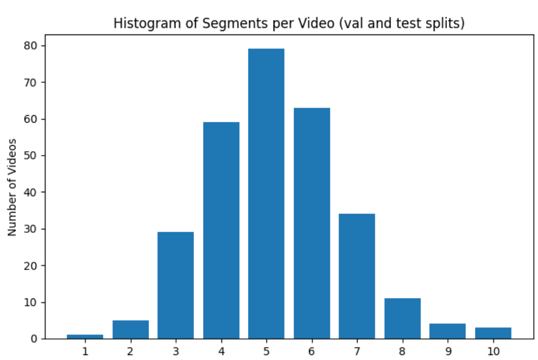


Figure 8. **Histogram of segments per video.**

ther we changed the evaluation to consider longer segment durations. Another difference is accounting for a variable sampling rate (which was originally hardcoded to 2). The sampling rate at which the videos features are extracted influences the loss and the training step of the model. Also, we opted to use the GPT2-xl [4] features for the text side, not CLIP [3] (we presented ablation studies in the main paper to justify this design choice). We will make the code available online along with the data.

| Dataset | Avg sents | Avg words |
|---|---|---|
| BMD-Train | 6.5 | 131.9 |
| BMD-Eval | 1.7 | 28.5 |
| YTC | 1.0 | 4.8 |
| Transcript | 330 | 4217 |

Table 3. **Query statistics.** It can be observed that the queries in YTC are very short containing only essential keywords.

### 3.3. Transcript usage with LLMs

As already stated, the transcript is very long, even for a segment proposal. Since the LLMs usually have an input length limit, in order to get a summary (which will act as the final query generation text), we have an iterative approach, where we split the transcript in several parts (that can be processed at once) and feed them independently through the LLM. In the end, the final query is obtained by concatenating all the subparts.

### 3.4. Query length

In Tab. 3 we present the average query length per segment in our BMD and YTC datasets. The chapter annotation in YTC are very short and contain around 5 words on average, while in BMD-Eval there are around 30 words on average per query.

### 3.5. Number of segments

The average number of segments per video in BMD-Train is around 4.5, while in YTC there are around 9.4 segments per video. The segments in BMD-Train where obtained by using OSG [5] with $scenes\_count = 5$. The segments were then filtered to have an associated transcript and to be shorter than 1.5h. For YTC, the chapter annotations were extracted from YouTube and were manually added by the creator of the video.

| Model | Pre-training | Training | $R1@0.5 \uparrow$ | $R1@0.7 \uparrow$ |
|---|---|---|---|---|
| CHAPTER-DETR | - | YTC | $12.6_{\pm 0.3}$ | $5.8_{\pm 0.6}$ |
| CHAPTER-DETR | BMD-val | YTC | $14.4_{\pm 0.3}$ | $5.9_{\pm 0.7}$ |
| CHAPTER-DETR | BMD-train | YTC | $\mathbf{16.1}_{\pm 0.5}$ | $\mathbf{6.6}_{\pm 0.3}$ |

Table 4. **Results on YouTube-Chapters.** The best results are obtained by pre-training on our automatically curated BMD-train split.

# 4. BMD

In Fig. 3 we show some additional qualitative examples for LONGMOMENT-DETR. Moreover, in Fig. 6 we present the histogram of transcript lengths in our BMD dataset. As expected, the transcripts are considerable long and contain a lot of wide-ranging dialogue. A visual representation to better understand the difference between a transcript and a human query is presented in Fig. 4. The validation and testing split are manually annotated and have a variable number of scenes per videos. A histogram of number of segments per video for the validation and testing splits is presented in Fig. 8.

# 5. YTC

In Fig. 5 we present a visual example of a video from YTC. In Fig. 7 we present the histogram of the YTC videos duration. We observe that the majority of the videos from YTC are about 2 hours long.

Additionally, in Tab. 4 we present the performance of using the features obtained by training with supervised data from BMD validation split on the downstream task of YouTube chapter detection. As can be observed, using the BMD-val as pre-training for YTC slightly improves performance. However using our proposed automatically curated BMD-train, the increase in performance is greater.

# References


[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 1

[2] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021. 1, 2

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 3

[4] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3

[5] Daniel Rotman, Dror Porat, and Gal Ashour. Robust and efficient video scene detection using optimal sequential grouping. In *2016 IEEE international symposium on multimedia (ISM)*, pages 275–280. IEEE, 2016. 1, 3