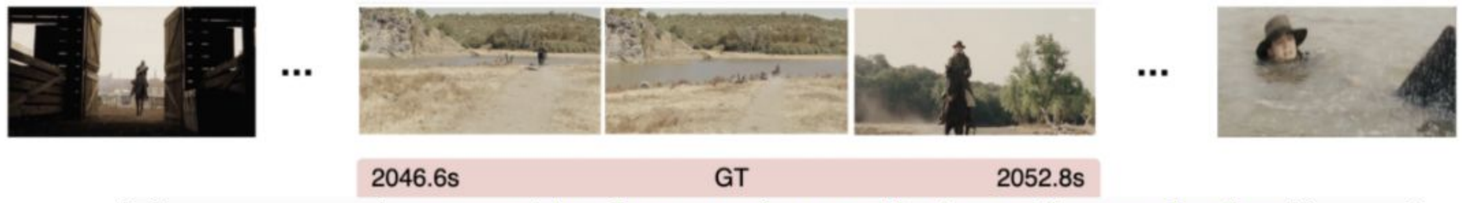
ReVisionLLM: Recursive Vision-Language Model for Temporal Grounding in Hour-Long Videos

Tanveer Hannan, Md Mohaiminul Islam, Jindong Gu, Thomas Seidl, Gedas Bertasius

https://sites.google.com/view/revisionllm

Problem Overview

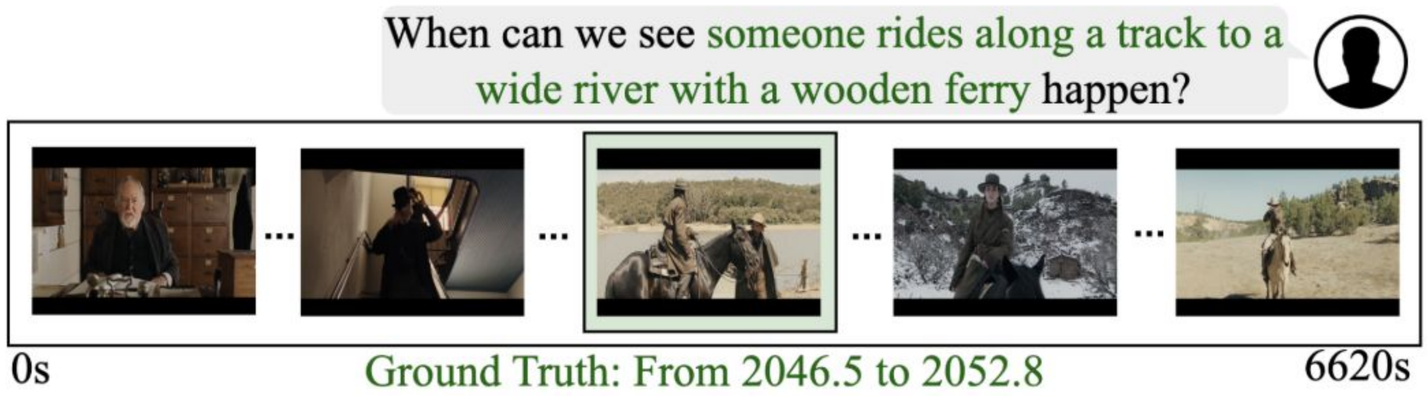
Given a long video (20–120 minutes), we want to predict the moment boundary specified by textual queries.



When can we see 'someone rides along a track to a wide river with a wooden ferry' happen?

Prior Works

Existing vision-language models are not equipped to process hour-long videos effectively and struggle to pinpoint precise temporal boundaries for events within extended video durations.

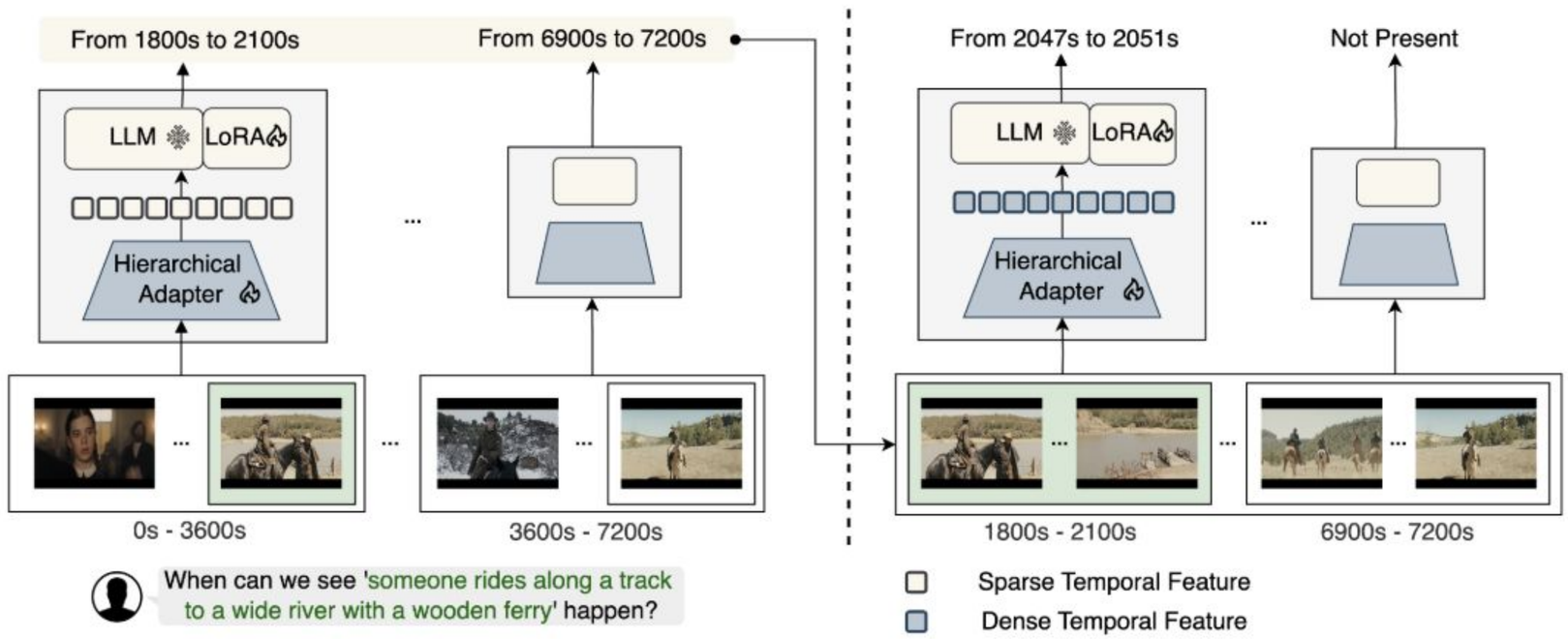


Tanveer Hannan, Md Mohaiminul Islam, Jindong Gu, Thomas Seidl, Gedas Bertasius

https://sites.google.com/view/revisionllm

Recursive Vision-Language Model

ReVisionLLM is a recursive vision-language model designed to localize events in hour-long videos. Inspired by human search strategies, it first scans the entire video to identify relevant intermediate segments and zooms in to precisely locate event boundaries.



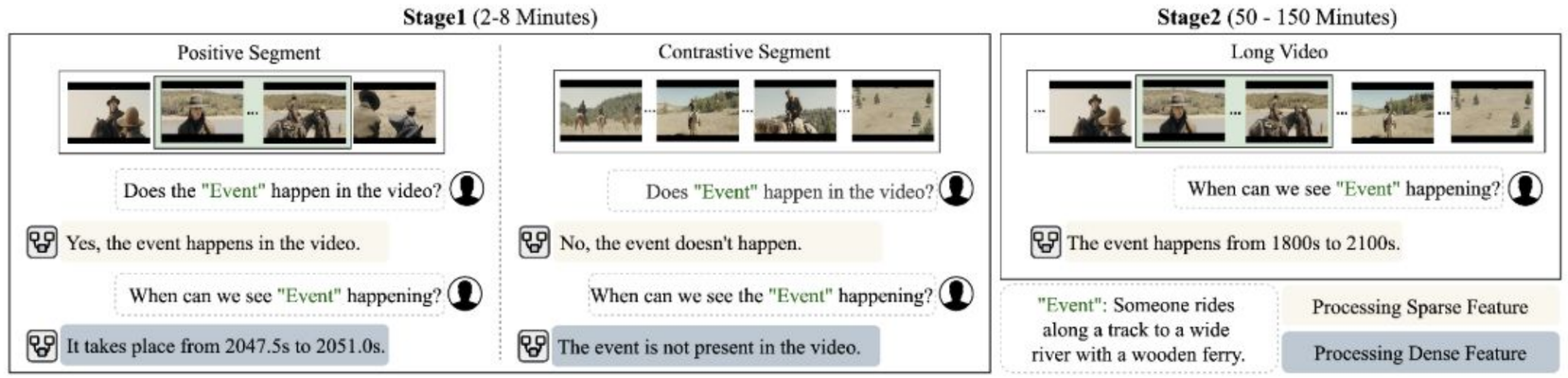
- First, we detect segments (e.g., a few minutes) from an hour-long video using sparse temporal features produced by the Hierarchical Adapter.
- Then ReVisionLLM produces a precise temporal boundary using dense temporal features within the predicted segments.

ReVisionLLM: Recursive Vision-Language Model for Temporal Grounding in Hour-Long Videos

Tanveer Hannan, Md Mohaiminul Islam, Jindong Gu, Thomas Seidl, Gedas Bertasius

https://sites.google.com/view/revisionllm

Progressive Training Method



Our model is trained progressively: first on short video segments and then on hour-long videos.

1. In the first stage, the model learns to detect whether an event is present in the input video and, if so, predicts its precise start and endpoints. Sparse features help determine an event's presence, while dense features additionally facilitate exact localization.
2. In the second stage, we utilize the sparse features learned in Stage 1 to identify event segments within hour-long videos.

Tanveer Hannan, Md Mohaiminul Islam, Jindong Gu, Thomas Seidl, Gedas Bertasius

https://sites.google.com/view/revisionllm

Experiments

1. RGNet outperforms the previous best MAD method (RGNet) by **+2.6%** and **+1.5%** on R1@.1 and R1@.3 scores. It further achieves SOTA on VidChapters-7M by outperforming the previous best (M-DETR) **+4.2%** in R1@.7 and **+8.8%** in R1@.9.

Model	MAD [48]							VidChapters-7m [62]				
	R1@.1	R5@.1	R1@.3	R5@.3	R1@.5	R5@.5	Avg.↑	R1@.3	R1@.5	R1@.7	R1@.9	Avg.↑
M-Guide [5]	9.3	18.9	4.6	13.1	2.2	7.4	9.3	-	-	-	-	-
CONE [16]	8.9	20.5	6.9	16.1	4.1	9.6	11.0	-	-	-	-	-
SOONet [42]	11.3	23.2	9.0	19.6	5.3	13.1	13.6	-	-	-	-	-
SnAG [39]	10.3	24.4	8.5	20.6	5.5	**13.7**	13.8	-	-	-	-	-
RGNet [15]	12.4	25.1	9.5	18.7	5.6	10.9	13.7	-	-	-	-	-
BERT [10]	-	-	-	-	-	-	-	0.6	0.3	0.1	0.0	0.3
VTimeLLM* [69]	1.4	3.1	1.3	2.5	0.6	1.1	1.7	10.6	4.1	1.6	0.2	4.1
CLIP [45]	6.6	15.1	3.1	9.9	1.5	5.4	6.9	10.7	5.2	2.3	0.5	4.7
M-DETR [26]	3.6	13.0	2.8	9.9	1.7	5.6	6.1	**37.4**	27.3	17.6	6.4	22.1
Ours†	**17.3**	**31.4**	**12.7**	**23.5**	**6.7**	13.1	**17.5**	-	-	-	-	-
Ours	15.0	25.1	11.0	18.8	5.8	10.5	14.4	33.8	**27.4**	**21.8**	**15.2**	**24.6**

2. We validate the effectiveness of each proposed modules with a cumulative ablation study. Each of our proposed modules contributes to a significant improvement in grounding capability, with the **Recursive Process** achieving the highest gains.

Modules	R1@.1↑	R5@.1↑	R1@.3↑	R5@.3↑
Baseline: VTimeLLM [18]	0.0	0.0	0.0	0.0
(+) CONE [16]	1.4	2.4	1.3	2.5
(+) Contrastive Segment	4.8	6.7	4.2	7.2
(+) Calibration (-) CONE	8.4	12.7	6.6	8.9
(+) Recursive Process*	**15.0**	**25.1**	**11.0**	**18.8**

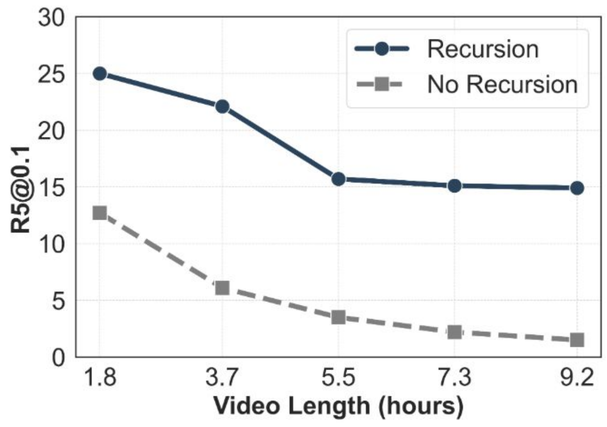
Tanveer Hannan, Md Mohaiminul Islam, Jindong Gu, Thomas Seidl, Gedas Bertasius
https://sites.google.com/view/revisionllm

Experiments

3. The model's performance improves with the number of hierarchies in the recursive structure, becoming more effective with each additional level.

Hierarchies	R1@.1↑	R5@.1↑	R1@.3↑	R5@.3↑
0	0.0	0.0	0.0	0.0
1	8.4	12.7	6.6	8.9
2	11.9	17.5	8.7	13.2
3	**15.0**	**25.1**	**11.0**	**18.8**

4. Our recursive approach maintains strong performance even with videos up to 10 hours long, while the baseline method fails entirely in these cases.

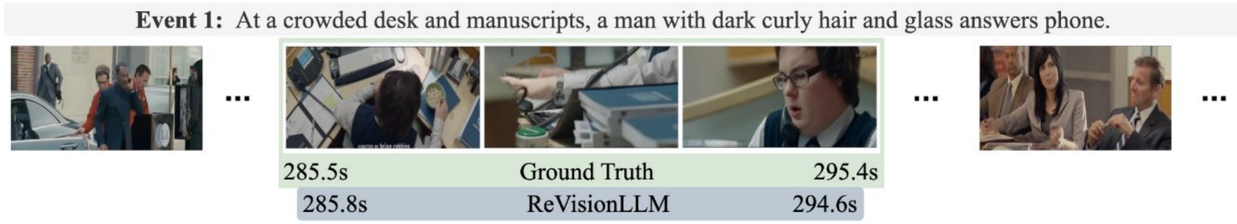


Tanveer Hannan, Md Mohaiminul Islam, Jindong Gu, Thomas Seidl, Gedas Bertasius

https://sites.google.com/view/revisionllm

Experiments

5. ReVisionLLM accurately locates precise event boundaries involving intricate actions and complex visual details within hour-long Videos.



6. ReVisionLLM successfully performs Text-to-video retrieval task. Here the model has to retrieve the positive video from a large database of videos.

Event 1: People are walking down a street holding signs.

Event 2: A man playing guitar and singing on road side.

Event 3: Dogs are walking across the road in a video.

Event 4: A car is racing on the road.

Tanveer Hannan, Md Mohaiminul Islam, Jindong Gu, Thomas Seidl, Gedas Bertasius
https://sites.google.com/view/revisionllm

Conclusion

- ✔ We introduce ReVisionLLM, the first VLM specifically designed with a recursive structure for temporal event grounding in **hour-long videos**.
- ✔ Its recursive architecture effectively can locate events within extensive videos and establishes a new state-of-the-art, outperforming specialized models.
- ✔ Future work could focus on integrating audio for better event comprehension and expanding the capabilities to handle even longer videos spanning multiple days.