Number it: Temporal Grounding Videos like Flipping Manga

Supplementary Material

6. Experimental Details

6.1. Moment Retrieval

In the training-free (NumPro) setting, we extract frames from videos at 1 FPS, with each frame resized to a resolution of 336 × 336. In the fine-tuned (NumPro-FT) setting, frames are extracted at 0.5 FPS during both the training and inference phases due to GPU memory constraints.

6.2. Highlight Detection

In both the training-free (NumPro) and fine-tuned (NumPro-FT) settings, we extract frames from videos at 0.5 FPS because the saliency score is labeled every 2 seconds. Each frame is resized to a resolution of 336 × 336.

7. Hallucination in Vid-LLMs for VTG

7.1. General Vid-LLMs

Qwen2-VL-7B. Figure 8a shows the results of Qwen2-VL-7B [66] suffer from severe hallucinations. For instance, the model generates responses like "from frame 000 to frame 200" even when the input video contains only 19 frames.
Qwen2-VL-72B. A larger-scale Vid-LLM, Qwen2-VL-72B [66], as shown in Figure 8b, also exhibits significant hallucination issues. The model produces illogical outputs, such as incomplete sentences like "The given query happens in344-," further emphasizing its struggle with coherent and accurate temporal reasoning.
LLaVA-Video-7B. Figure 9a displays the distribution of the top 10 most common time intervals predicted by LLaVA-Video-7B [89]. The model frequently outputs very short segments, such as $[1, 3]$, $[2, 4]$, and $[2, 3]$, which together account for over 50% of predictions. This behavior suggests a significant bias in the model toward producing overly simplistic temporal spans.
LLaVA-OneVision-7B. Figure 9b presents the 10 most frequently predicted intervals of LLaVA-OneVision-7B [35]. The outputs are dominated by segments like $[10, 12]$ and $[1, 3]$, which together account for over 50% of all results. The repetitive predictions indicate a substantial hallucination problem limiting temporal reasoning capability.

7.2. VTG-Tuned Vid-LLMs

VTimeLLM. Figure 9c shows the predictions on the Charades-STA dataset of VTimeLLM [25]. The model frequently predicts certain frame intervals, such as $[17, 34]$, which accounts for 49.34% of predictions, and $[0, 17]$, which constitutes 16.34%. This pattern suggests significant hallucination and overfitting to specific frame numbers.

Table 5. The ablation results on the QVHighlights dataset.

Model	QVHighlights	
	mAP	HIT@1
LLaVA-OneVision-7B	17.2	19.9
+*NumPro*	20.9 (+3.7)	27.6 (+7.7)
LLaVA-Video-7B	20.7	34.8
+*NumPro*	22.3 (+1.6)	38.4 (+4.4)
Qwen2-VL-72B	21.6	37.5
+*NumPro*	24.2 (+2.6)	44.3 (+6.8)
LongVA-7B-DPO	14.2	20.4
+*FT*	21.9	30.8
+*NumPro-FT*	25.0 (+10.8)	37.2 (+16.8)

Table 6. Ablation study with different font size of NumPro-FT.

Size	Color	Position	Charades-STA			
			R@0.3	R@0.5	R@0.7	mIoU
40	Red	Bottom Right	**63.8**	**42.0**	**20.6**	**41.4**
60	Red	Bottom Right	56.0	37.6	**20.6**	40.9

Table 7. Performance comparison between the original Qwen2-VL-7B and the "Attention Map" method, which selects the two frames with the highest attention scores as the temporal boundaries.

Method	Charades-STA			
	R@0.3	R@0.5	R@0.7	mIoU
Qwen2-VL-7B	8.7	5.4	2.4	7.9
Attention Map	**18.1**	**11.4**	**3.1**	**19.8**

TimeChat. As shown in Figure 9d, we analyze the output of the TimeChat [58] model. It tends to produce results in multiples of 5, such as 5, 10, 15, and 20. Notably, intervals like $[0, 5]$ and $[0, 10]$ appear in over 42% of predictions. This indicates both hallucinations and overfitting.

8. Additional Attention Analysis

We present additional attention analysis results in Figure 10. In the examples on the left, the model produces incorrect or incomplete outputs, such as "from 2 to .". On the right, the examples display severe hallucinations, with outputs extending beyond the video's actual duration. Despite these issues, the attention maps in both cases consistently highlight the relevant video segments. These findings show that while Vid-LLMs identify correct video segments, they fail to output accurate temporal boundaries due to the inability to translate these segments into precise textual locations.

To further examine this challenge of Vid-LLMs quantitatively, we conduct an experiment with the Qwen2-VL-7B model on the Charades-STA dataset. In this experiment, the two frames with the highest attention scores are selected as the predicted start and end frames of the segment. Specifically, we selected the two frames with the highest attention scores as the start and end frames, treating these as the predicted segment boundaries. The results, presented in Table 7, showing that this naïve attention-based solution achieves an improvement of 11.9% in mIoU compared to direct predictions of the original model. This significant gain supports our observation that Vid-LLMs have the inherent capacity to locate relevant video segments but struggle to express temporal boundaries accurately in text.

Overall, these analyses highlight the primary bottleneck of Vid-LLMs in addressing temporal grounding tasks and emphasize the need to overcome this limitation, which is what our proposed NumPro method is designed to address.

9. Additional Video Benchmark Results

We conducted experiments on additional video question answering (QA) benchmarks, MVBench [37] and VideoMME [16], as summarized in Table 8. We use 1FPS as the sampling rate, and adopt a design of red color, font size 40, and bottom right positioning for the number prompt. Our results demonstrate that Vid-LLMs enhanced with NumPro achieve robust performance across diverse downstream tasks. Notably, NumPro significantly improves the Vid-LLMs' generalization capabilities in temporal understanding tasks, such as Scene Transition and Temporal Perception. These findings align with our previous results presented in Table 4 on VideoInstruct [50] in the main paper.

Table 8. Evaluations on two video QA benchmarks: MVBench and VideoMME. The results demonstrate that our NumPro approach enhances Vid-LLMs' generalization capabilities on downstream tasks involving temporal understanding.

MVBench			**Video-MME**		
Scene Transition	**State Change**	**Overall**	**Temp. Per.**	**Temp. Rea.**	**Overall**
80.0 **(+2.5)**	42.0 **(+1.0)**	51.8 **(+0.2)**	72.7 **(+12.7)**	49.7 **(+8.8)**	63.7 **(+0.3)**

10. Ablation Results on Highlight Detection

We present additional ablation results on the QVHighlights dataset, as shown in Table 5. The results show that our method generalizes well across various General Vid-LLMs, achieving notable improvements in both mAP and HIT@1 metrics. Specifically, fine-tuning with NumPro-FT obtains a 10.8% increase in mAP and a 16.8% increase in HIT@1, surpassing state-of-the-art results.

11. Ablation Results on NumPro-FT Designs

In this section, we present the ablation results for NumPro-FT. While a font size of 60 achieves better Number Accuracy than a size of 40 (Figure 5 of the main paper), it reduces Caption Accuracy and introduces more outliers. Table 6 further supports this finding, showing that a font size of 60 results in generally lower performance, including a 7.8% drop in R@0.3 compared to size 40. We attribute this to interference with the model's understanding of video content.

Table 9. Comparison between overlaying timestamps with overlaying frame numbers in NumPro design.

Dataset	R@0.3	R@0.5	R@0.7	mIoU
Charades-STA	58.4 **(-2.3)**	37.8 **(+1.0)**	16.6 **(+0.7)**	37.6 **(-0.9)**
ActivityNet	31.6 **(-12.6)**	18.1 **(-6.3)**	10.2 **(-4.2)**	23.0 **(-8.3)**

12. NumPro with Accurate Timestamps

In the main paper, we choose frame numbers because they serve as the smallest discrete units of a video and can be directly mapped to precise timestamps using the frame sampling rate. In this section, we compare the performance by directly overlaying actual timestamps in videos. However, timestamps (*e.g.*, "10.5s") may introduce decimals, which can increase parsing complexity for Vid-LLMs. In Tabel 9, we compare minute-level temporal annotations (*e.g.*, "01:10") with frame numbers annotations sampled with 1FPS. The performance on Charades-STA is comparable, while frame numbers outperformed timestamps on ActivityNet, which includes longer videos and more annotations. These suggest that overlaying numbers with temporal information is an effective strategy for Vid-LLMs in temporal grounding, while frame numbers offer a simpler and more scalable solution.

13. Additional Visualization Cases

13.1. Dialogue

Figure 11 illustrates a real-world application of our NumPro method within the Qwen2-VL-7B model, highlighting its ability to handle complex video-based dialogue tasks. Compared to the VTG-specific models [10, 41], NumPro equipped with Vid-LLMs facilitates multi-turn dialogue that adapt to user queries in real-time. For instance, the model can track score changes across video segments, identify celebrities through advanced facial recognition, and even extract textual information via OCR. These capabilities demonstrate the enhanced contextual understanding and practical value of Vid-LLMs for video comprehension tasks. By integrating NumPro, our approach further refines the temporal grounding process, enabling more precise and interactive video analysis for real-world applications.

13.2. Moment Retrieval

Figure 12 showcases additional visualization examples highlighting the effectiveness of our method in moment retrieval tasks. Our approach demonstrates robust temporal grounding capabilities by accurately identifying event boundaries across videos of varying lengths and content. Compared to previous state-of-the-art methods, it achieves substantial performance improvements.

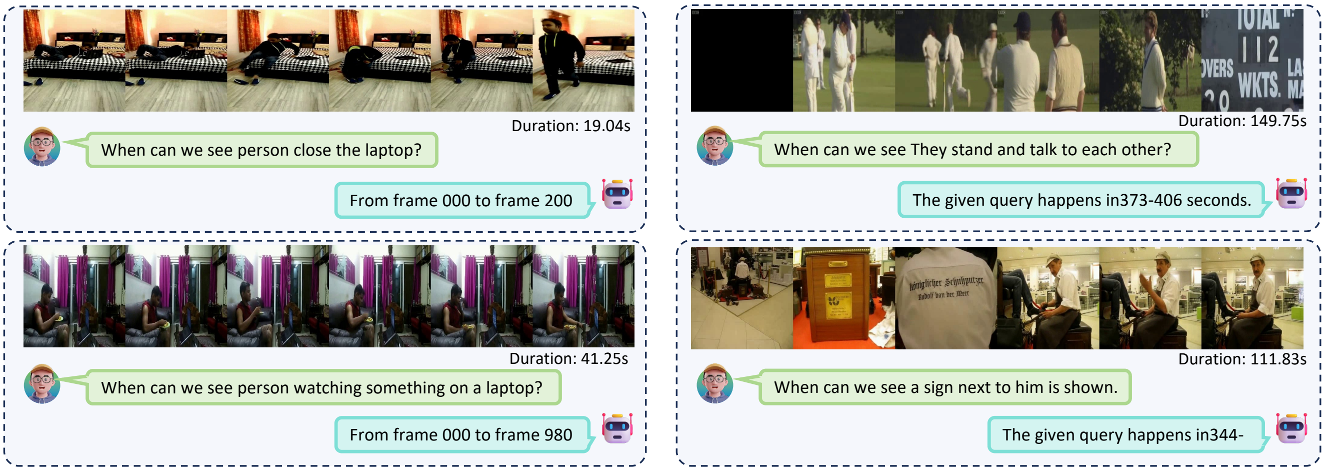
13.3. Highlight Detection

Highlight detection [33] focuses on identifying video segments that match a given query while also assessing their relative importance. The model generates timestamps for the relevant segments and assigns a saliency score on a scale from 1 to 5. As shown in Figure 13, our method excels in accurately predicting segment start and end times, achieving consistently high mAP values. Additionally, it demonstrates precision in saliency score assessment, highlighting its suitability for tasks requiring detailed temporal localization and importance evaluation.

14. Limitations

While NumPro and NumPro-FT have proven effective across multiple models and datasets, significantly surpassing previous state-of-the-art models, there are still some limitations:

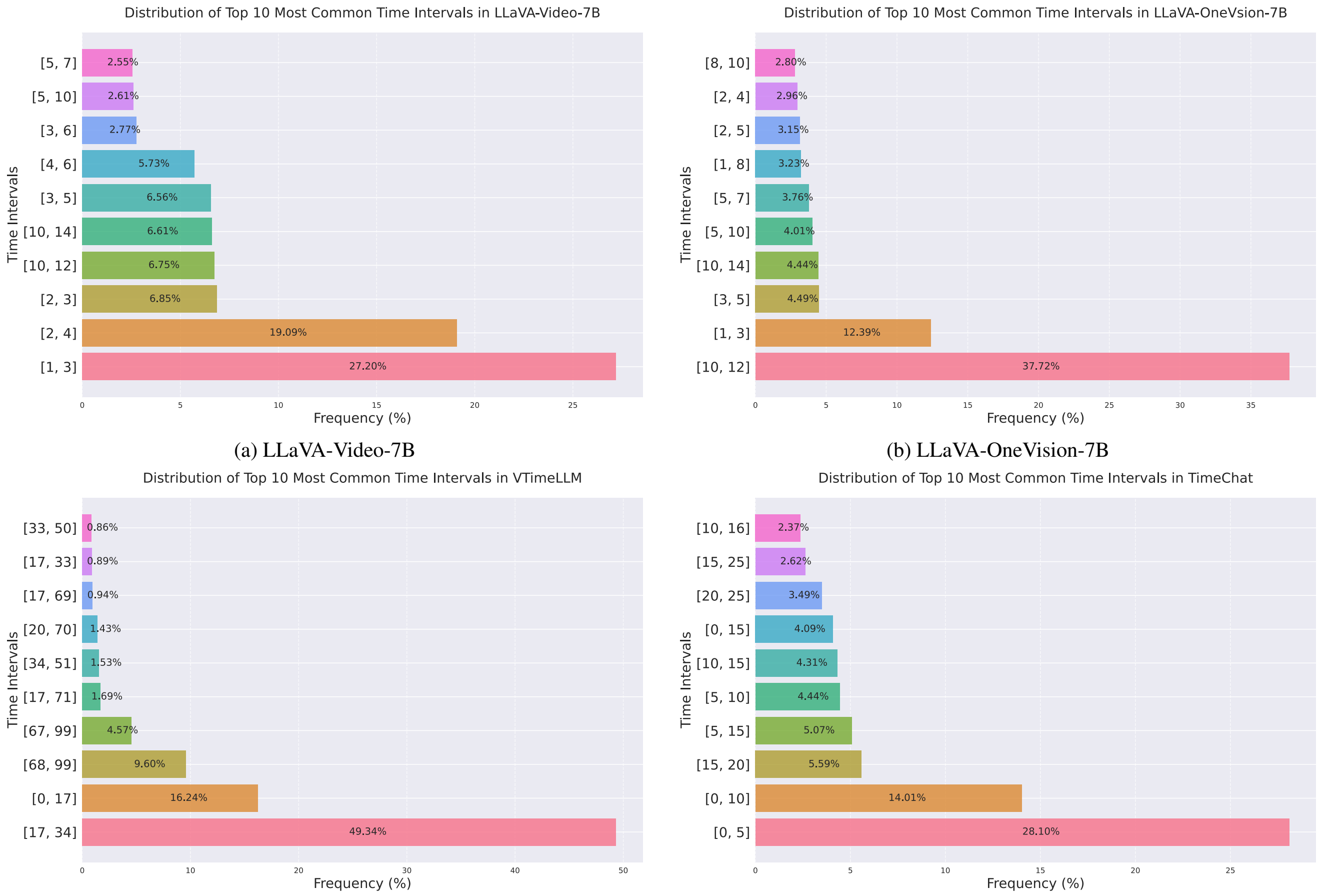
- Limited Dataset Scope: Current datasets for video temporal grounding (VTG) tasks are predominantly focused on short videos, typically ranging from 30 seconds to 3 minutes in duration. Expanding evaluation to include longer videos, such as hour-long recordings, is essential for testing the scalability and generalizability of our approach.
- Potential Visual Obstruction: Although NumPro is designed to minimize its impact on video content, there are scenarios where it might obscure critical visual elements, such as details, watermarks, or logos. Future enhancements could involve dynamic adjustments to the opacity of numbers or the implementation of adaptive number positioning to avoid blocking essential visual information.
- Frame Rate Optimization: The effect of different sampling frame rates on performance remains underexplored. This study used a fixed frame rate of 1 FPS for NumPro, which may not be universally optimal. Investigating adaptive frame rates that align with the perceptual and computational characteristics of various models could lead to further improvements in accuracy and efficiency.


Duration: 19.04s
When can we see person close the laptop?
From frame 000 to frame 200
Duration: 41.25s
When can we see person watching something on a laptop?
From frame 000 to frame 980
Duration: 149.75s
When can we see They stand and talk to each other?
The given query happens in373-406 seconds.
Duration: 111.83s
When can we see a sign next to him is shown.
The given query happens in344-

(a) Qwen2-VL-7B

(b) Qwen2-VL-72B

Figure 8. Video temporal grounding results where the models exhibit serious hallucination and output incorrect results. In all cases, frames are sampled at 1 FPS.



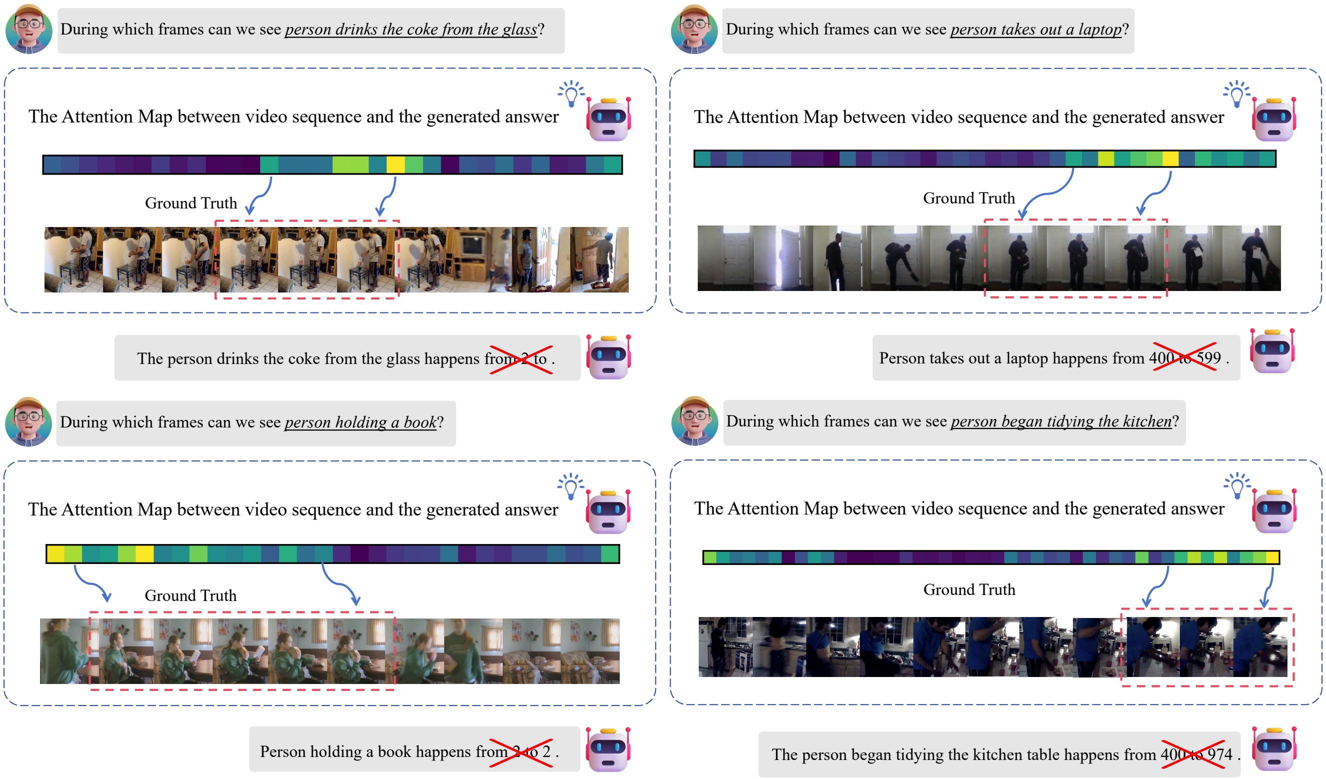
(a) LLaVA-Video-7B

(b) LLaVA-OneVision-7B

(c) VTimeLLM

(d) TimeChat

Figure 9. Distribution of the Top 10 Most Common Time Intervals within the Charades-STA Dataset for Different Models.



During which frames can we see *person drinks the coke from the glass*?

The Attention Map between video sequence and the generated answer
Ground Truth

The person drinks the coke from the glass happens from ~~2 to~~ .

During which frames can we see *person takes out a laptop*?

The Attention Map between video sequence and the generated answer
Ground Truth

Person takes out a laptop happens from ~~400 to 599~~ .

During which frames can we see *person holding a book*?

The Attention Map between video sequence and the generated answer
Ground Truth

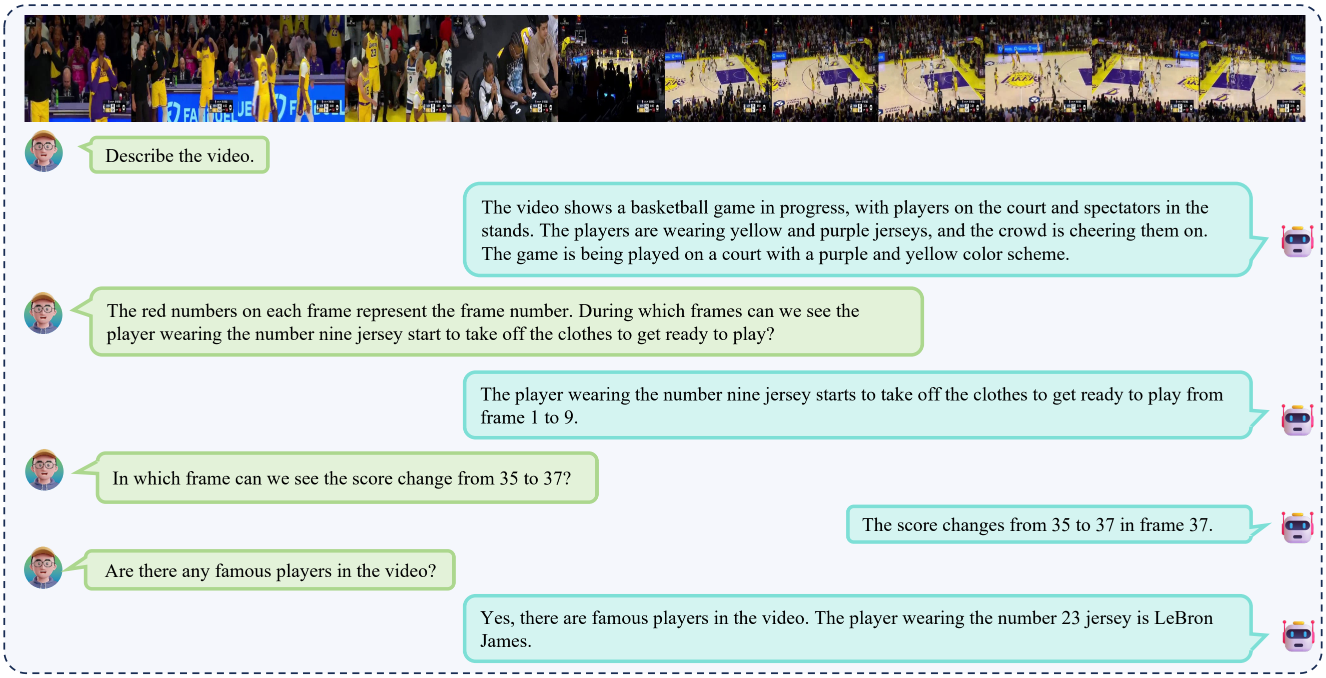
Person holding a book happens from ~~2 to 2~~ .

During which frames can we see *person began tidying the kitchen*?

The Attention Map between video sequence and the generated answer
Ground Truth

The person began tidying the kitchen table happens from ~~400 to 974~~ .

Figure 10. Additional attention analysis results on general Vid-LLMs



Describe the video.

The video shows a basketball game in progress, with players on the court and spectators in the stands. The players are wearing yellow and purple jerseys, and the crowd is cheering them on. The game is being played on a court with a purple and yellow color scheme.

The red numbers on each frame represent the frame number. During which frames can we see the player wearing the number nine jersey start to take off the clothes to get ready to play?

The player wearing the number nine jersey starts to take off the clothes to get ready to play from frame 1 to 9.

In which frame can we see the score change from 35 to 37?

The score changes from 35 to 37 in frame 37.

Are there any famous players in the video?

Yes, there are famous players in the video. The player wearing the number 23 jersey is LeBron James.

Figure 11. An example of multi-turn dialogue using our NumPro method in the Qwen2-VL-7B model.

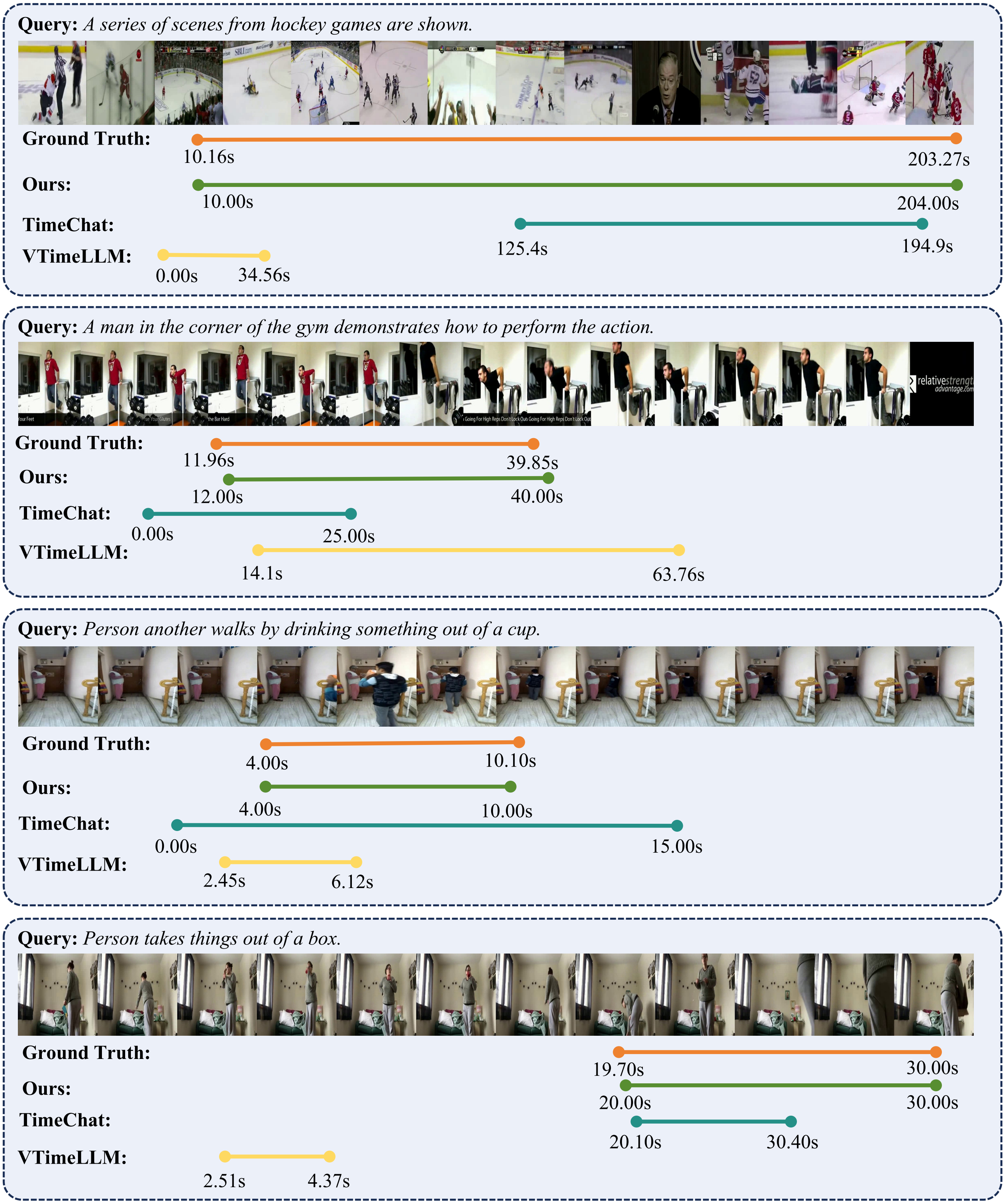
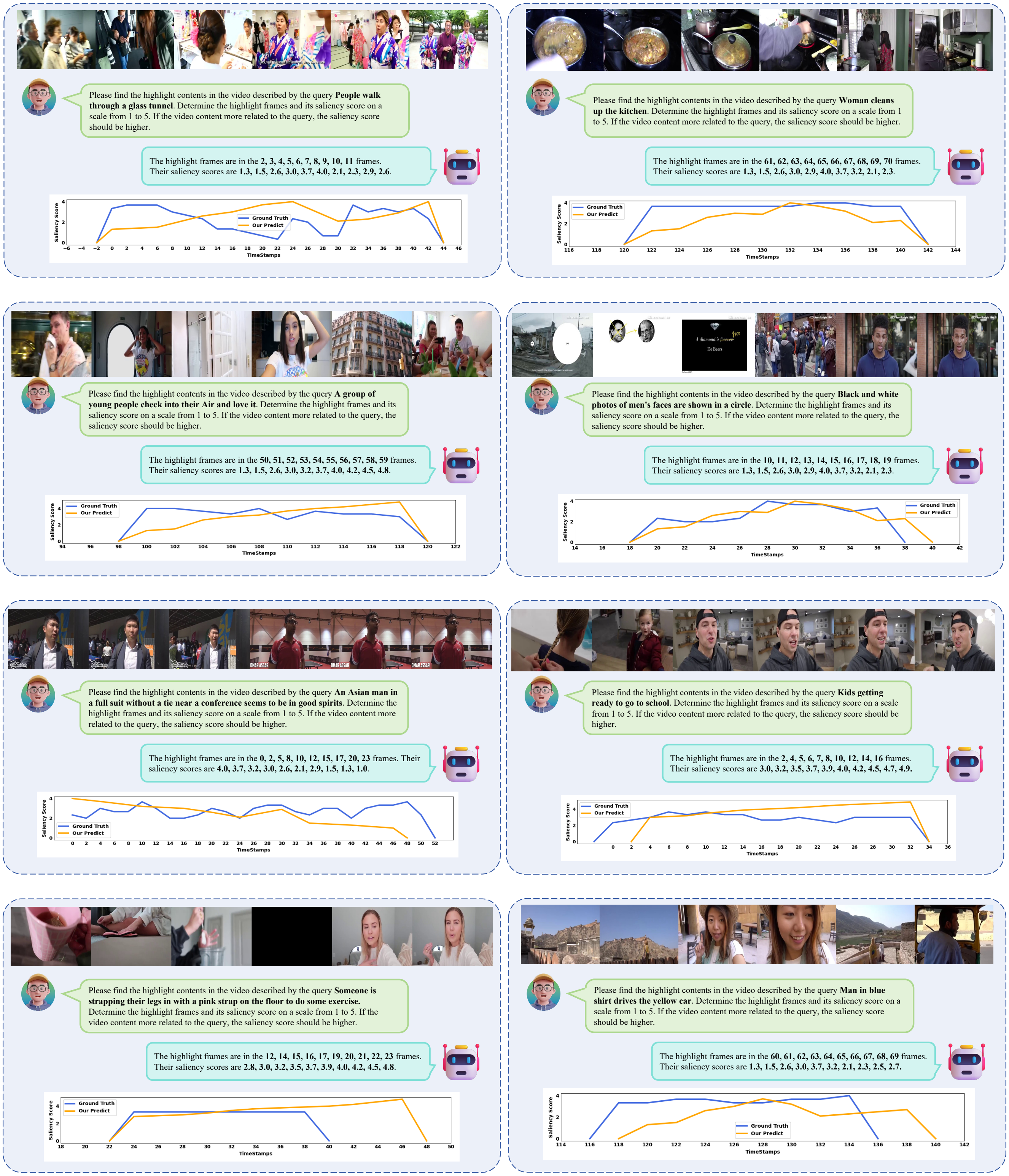


Figure 12. Additional visualization cases of Video Temporal Grounding task.



Please find the highlight contents in the video described by the query **People walk through a glass tunnel**. Determine the highlight frames and its saliency score on a scale from 1 to 5. If the video content more related to the query, the saliency score should be higher.

The highlight frames are in the **2, 3, 4, 5, 6, 7, 8, 9, 10, 11** frames. Their saliency scores are **1.3, 1.5, 2.6, 3.0, 3.7, 4.0, 2.1, 2.3, 2.9, 2.6**.

Please find the highlight contents in the video described by the query **Woman cleans up the kitchen**. Determine the highlight frames and its saliency score on a scale from 1 to 5. If the video content more related to the query, the saliency score should be higher.

The highlight frames are in the **61, 62, 63, 64, 65, 66, 67, 68, 69, 70** frames. Their saliency scores are **1.3, 1.5, 2.6, 3.0, 2.9, 4.0, 3.7, 3.2, 2.1, 2.3**.

Please find the highlight contents in the video described by the query **A group of young people check into their Air and love it**. Determine the highlight frames and its saliency score on a scale from 1 to 5. If the video content more related to the query, the saliency score should be higher.

The highlight frames are in the **50, 51, 52, 53, 54, 55, 56, 57, 58, 59** frames. Their saliency scores are **1.3, 1.5, 2.6, 3.0, 3.2, 3.7, 4.0, 4.2, 4.5, 4.8**.

Please find the highlight contents in the video described by the query **Black and white photos of men's faces are shown in a circle**. Determine the highlight frames and its saliency score on a scale from 1 to 5. If the video content more related to the query, the saliency score should be higher.

The highlight frames are in the **10, 11, 12, 13, 14, 15, 16, 17, 18, 19** frames. Their saliency scores are **1.3, 1.5, 2.6, 3.0, 2.9, 4.0, 3.7, 3.2, 2.1, 2.3**.

Please find the highlight contents in the video described by the query **An Asian man in a full suit without a tie near a conference seems to be in good spirits**. Determine the highlight frames and its saliency score on a scale from 1 to 5. If the video content more related to the query, the saliency score should be higher.

The highlight frames are in the **0, 2, 5, 8, 10, 12, 15, 17, 20, 23** frames. Their saliency scores are **4.0, 3.7, 3.2, 3.0, 2.6, 2.1, 2.9, 1.5, 1.3, 1.0**.

Please find the highlight contents in the video described by the query **Kids getting ready to go to school**. Determine the highlight frames and its saliency score on a scale from 1 to 5. If the video content more related to the query, the saliency score should be higher.

The highlight frames are in the **2, 4, 5, 6, 7, 8, 10, 12, 14, 16** frames. Their saliency scores are **3.0, 3.2, 3.5, 3.7, 3.9, 4.0, 4.2, 4.5, 4.7, 4.9**.

Please find the highlight contents in the video described by the query **Someone is strapping their legs in with a pink strap on the floor to do some exercise.** Determine the highlight frames and its saliency score on a scale from 1 to 5. If the video content more related to the query, the saliency score should be higher.

The highlight frames are in the **12, 14, 15, 16, 17, 19, 20, 21, 22, 23** frames. Their saliency scores are **2.8, 3.0, 3.2, 3.5, 3.7, 3.9, 4.0, 4.2, 4.5, 4.8**.

Please find the highlight contents in the video described by the query **Man in blue shirt drives the yellow car**. Determine the highlight frames and its saliency score on a scale from 1 to 5. If the video content more related to the query, the saliency score should be higher.

The highlight frames are in the **60, 61, 62, 63, 64, 65, 66, 67, 68, 69** frames. Their saliency scores are **1.3, 1.5, 2.6, 3.0, 3.7, 3.2, 2.1, 2.3, 2.5, 2.7**.

Figure 13. Additional visualization cases of Highlight Detection task on QVHighlights dataset.