DisTime: Distribution-based Time Representation for Video Large Language Models

Supplementary Material

Appendix

In the appendix, we describe (1) additional details about the training strategy (Section A), (2) further ablation experiments (Section B), (3) additional results for moment retrieval (Section C), (4) the construction of training data instructions (Section D), and (5) qualitative results (Section E). For sections, figures, tables, and equations, we use numbers (*e.g.* Sec. 1) to refer to the main paper and capital letters (*e.g.* Sec. A) to refer to this appendix.

A. Training Details

For DisTime-InternVL, we employed a total batch size of 16 throughout the training process. The AdamW optimizer [29] was applied with a cosine learning rate decay and an initial warm-up period. During training, we used a single epoch with a learning rate set to 4×10^{-5}. The LoRA [14] parameters were configured with $r = 16$ and $\alpha = 32$. We complete the model training process on 8 A100 GPUs. The completion time for InternVL2.5-1B is approximately 40 hours, while for InternVL2.5-8B, it amounts to around 61 hours. For DisTime-LLaVAOV, we employed a total batch size of 512 throughout the training process. During training, we used a single epoch with a learning rate set to 2×10^{-5}. The LoRA [14] parameters were configured with $r = 64$ and $\alpha = 16$. Other settings are consistent with [18]. We complete the model training process on 8 A100 GPUs approximately 50 hours.

B. More Ablation Studies

Effectiveness of the scoring and ensemble in InternVid-TG. To validate the effectiveness of the scoring, we explore using the query and video data from the Charades-STA dataset, including 3720 queries. We perform inference on these three models in a zero-shot setting and calculate the score and temporal mean-IoU (mIoU) with the ground truth. Fig. A shows the mIoU results of prediction under various scores, indicating a positive correlation between the score and mIoU. Next, we explore how to leverage this score to perform an ensemble on the model results. First, we use queries from the Charades-STA dataset to predict and score the events, setting a series of offset values to examine the bias in the scoring strategy for each model. We find that the biases for each model are equal, which allows us to directly consider the temporal result with the highest score as the ensemble output. To directly verify the benefits of this ensemble method, we manually annotate the temporal positions for 1k queries extracted from step 1 (referred to as InternVid-TG-1k). Tab. A presents the metrics for independent inference of each model and the ensemble results on the InternVid-TG-1k dataset. Our method achieves a mIoU of 43.97%, which closely approaches the SOTA performance on public datasets (*e.g.* 46.83% on ANet-Caption), indicating that the ensemble results enhance each component, resulting in more accurate annotation localization.

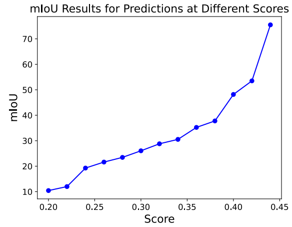


Figure A. mIoU *vs*. Confidence Score Curve.

Method	mIoU	R1@50
UniMD	26.63	22.22
Mr.Blip	39.81	34.34
TFVTG	38.95	40.40
Ensemble	**43.97**	**45.45**

Table A. Ensemble *vs*. Individual Model Comparison

Number of bins in time decoder. The results for different numbers of regression bins in the time decoder are shown in Tab. B. As depicted in the table, utilizing 32 bins achieves the best performance in both Charades-STA and YouCook2, except for a slightly lower F1 score compared to using 16 bins (20.5% *vs*. 20.9%). Notably, using a larger number of 64 bins does not result in more accurate time predictions. We believe the large number of bins complicates the model's ability to capture the flow of event timings smoothly.

$\#reg_{max}$	Charades-STA (MR)				YouCook2 (DVC)		
	R@1 (IoU=0.3)	R@1 (IoU=0.5)	R@1 (IoU=0.7)	mIoU	SODA_c	CIDEr	F1 Score
16	77.6	54.4	27.6	50.2	4.0	12.1	**20.9**
32	**78.1**	**56.3**	**29.7**	**51.6**	**4.2**	**15.6**	20.5
64	77.0	53.7	27.4	50.3	2.9	11.2	16.4

Table B. Ablation study on the number of bins (reg_{max}) in time decoder.

Number of layers in time decoder and encoder. As shown in Tab. C, the number of time decoder and encoder layers significantly impacts the final results. Utilizing three lay-

$\#layers$	Charades-STA (MR)				YouCook2 (DVC)		
	R@1 (IoU=0.3)	R@1 (IoU=0.5)	R@1 (IoU=0.7)	mIoU	SODA_c	CIDEr	F1 Score
2	77.1	52.4	26.6	50.1	3.4	10.1	17.0
3	**78.1**	**56.3**	**29.7**	**51.6**	**4.2**	**15.6**	**20.5**
4	77.3	54.4	27.9	50.0	3.6	15.5	18.0

Table C. Ablation study on the number of layers (L) in time decoder and time encoder.

Model	Size	QVHighlights			
		R@1 (IoU=0.3)	R@1 (IoU=0.5)	mAP@.5	mAP@.75
Dedicated					
Mr.BLIP [32]	3B	74.8	60.5	68.1	53.4
LLaVA-MR [30]	3B	**76.6**	**61.5**	**69.4**	**54.4**
Video-LLMs					
Momentor [35]	7B	17.0	-	7.6	-
DisTime-InternVL	1B	54.1	27.8	47.9	19.2
DisTime-LLaVAOV	7B	44.1	14.9	37.9	8.4
DisTime-InternVL	8B	61.1	37.5	53.8	28.1

Table D. Comparison with different models on moment retrieval task in QVHighlights. “Dedicated” refers to the model fine-tuned with benchmark-specific training data. Notably, our DisTime is in a zero-shot setting.

ers yields optimal performance, while further increasing the number of layers does not lead to better results. We believe that the embeddings produced by the LLM inherently contain substantial temporal information, and excessive layers may disrupt this intrinsic data.

C. More Results

Moment retrieval task on QVHighlights. As shown in Tab. D, we surpass previous video-LLMs by a large margin (7.6% → 53.8% in mAP@0.5). However, we still lag behind dedicated models. This is due to the presence of multiple segments for a single event in the QVHighlight annotations, which makes it challenging for video-LLM models to recall the targets effectively.

D. Instructions for Time-Sensitive Task

In addition to the rich instruction data included in the ET-Instruct dataset [28], we expanded the instructions for the moment retrieval (used in InternVid-TG) and the dense video captioning. Specifically, we referred to TimeChat [37] and ET-Instruct, used GPT-4o [1] to expand more high-quality instructions, and finally manually selected the generated instructions as the final templates. Tab. E shows our instruction examples, answer format, and output examples for the moment retrieval and dense video captioning.

E. Qualitative Analyses

Impact of temporal distribution. We demonstrate the importance of temporal distribution in temporal grounding tasks, as illustrated in Fig. B. When the model receives the event query, it initially perceives an open cupboard door in the frame at 0 seconds, causing a slight response in the start time distribution curve. However, as time progresses, this response diminishes until the person begins to close the door, at which point the start time responds again. As the door-closing action comes to an end, the response for the end time gradually weakens. In the frame at 30 seconds, a hand blocks the cupboard door, resulting in another slight response from the model. Compared to the Dirac distribution, representing time as a broader distribution is more suitable for capturing events with blurred boundaries.

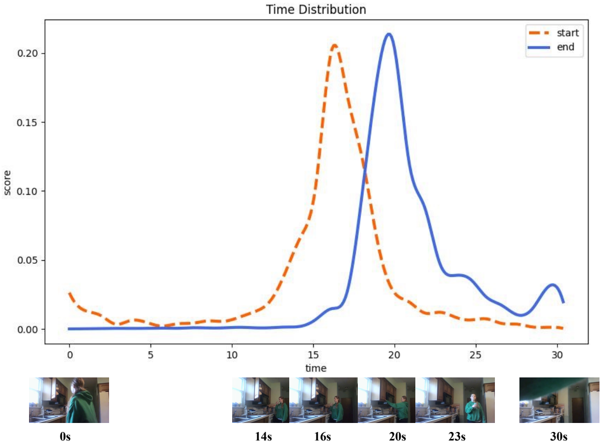
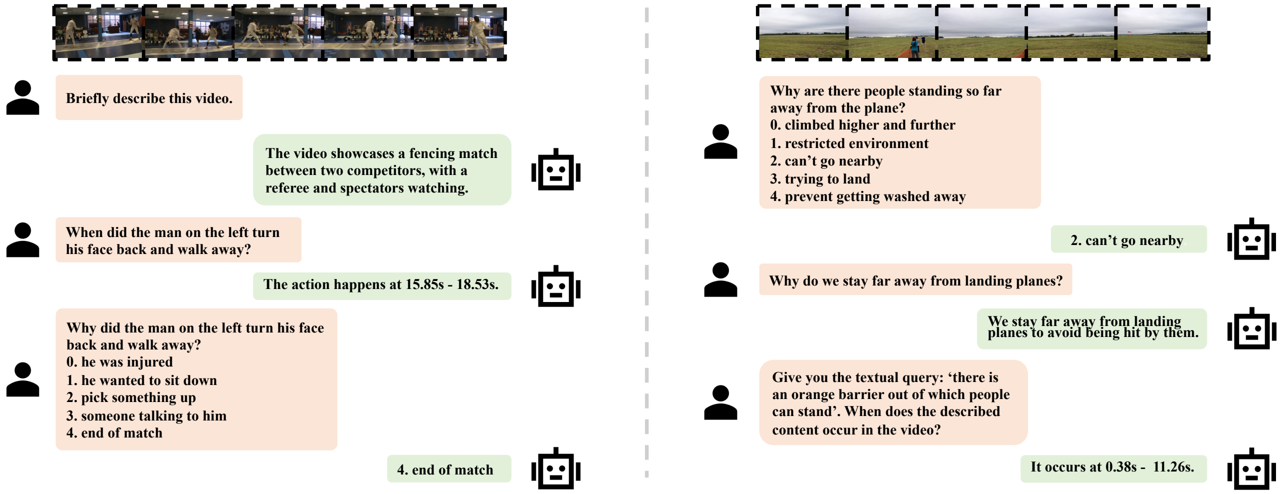


Figure B. Visualization curves of the start and end time in moment retrieval. Query: person closes a cupboard door. Orange curve represents the start time distribution; Blue curve represents the end time distribution.

Qualitative results. This section presents the qualitative results from videos involving multiple tasks, including open-ended question answering and moment retrieval, as illustrated in Fig. C. Additionally, we provide a visual comparison between DisTime’s predictions and the ground truth on temporal grounding tasks. Specifically, for moment retrieval, we show the visualization results of DisTime on the Charades-STA and ANet-Caption, as depicted in Fig. D. For dense video captioning, we display the visualization results of DisTime on the step description dataset YouCook2 and the event description dataset ANet-Caption, as illustrated in Fig. E and Fig. F, respectively. This comparison demonstrates the advanced capabilities of our method in accurately modeling and predicting temporal events.

Task	Type	Example
Moment Retrieval	Instruction Example	Give you a textual query: <query_placeholder> . When does the described content occur in the video? Please return the timestamp. Here is a text query: <query_placeholder> . At what point in the video does the described event happen? Please provide the timestamp. Analyze the event description: <query_placeholder> . At what moment in the video does the described event take place? Return the timestamp. Consider the query: <query_placeholder> . When is the described event occurring in the video? Kindly provide the timestamp. Examine the following text query: <query_placeholder> . When is the described event taking place in the video? Please return the timestamp.
	Answer Format	The event occurs at <TIME_STAMP> . The described event takes place at <TIME_STAMP> . This situation happens at <TIME_STAMP> . This event is at <TIME_STAMP> . It takes place at <TIME_STAMP> .
	Output Example	The event occurs at 1.2s - 5.8s. The described event takes place at 3.4s - 7.2s. This situation happens at 22.1s - 36.0s. This event is at 12.5s - 30.1s. It takes place at 33.1 - 41.0s.
Dense Video Caption	Instruction Example	Identify and localize a series of steps or actions occurring in the video, providing start and end timestamps and related descriptions. Localize a series of action steps in the given video, output a start and end timestamp for each step, and briefly describe the step. Capture and describe the activity events in the given video, specifying their respective time intervals, and output the time. Pinpoint the time intervals of activity events in the video, and provide detailed descriptions for each event. Detect and report the start and end timestamps of activity events in the video, along with descriptions.
	Answer Format	<TIME_STAMP> , Step1. <TIME_STAMP> , Step2. <TIME_STAMP> , Step3... <TIME_STAMP> , Step1. <TIME_STAMP> , Step2. <TIME_STAMP> , Step3... <TIME_STAMP> , Event1. <TIME_STAMP> , Event2. <TIME_STAMP> , Event3 ... <TIME_STAMP> , Event1. <TIME_STAMP> , Event2. <TIME_STAMP> , Event3 ... <TIME_STAMP> , Event1. <TIME_STAMP> , Event2. <TIME_STAMP> , Event3 ...
	Output Example	23.7s - 69.8s, spread the meat on the foil. 74.5s - 111.6s, cut the meat into four pieces. 114.7s - 146.0s, ... 26.7s - 50.8s, add oil to the wok. 49.3s - 82.2s, add garlic and green onions to the wok.81.7 - 107.7s, add ... 2.1s - 96.4s, a large orange leaf blower blows leaves in a yard. 96.5s - 196.8s, a man drives the leaf blower. 2.5s - 18.1s, a man is seen speaking to the camera and leads into him pouring oil into a pot. 18.2s - 58.8s, he ... 0.3s - 17.2s, a girl is seen climbing across a set of monkey bars while looking back to the camera. 17.5s - 32.8s, ...

Table E. Example of instructions for moment retrieval and dense video captioning. <query_placeholder> is the query placeholder for moment retrieval, and <TIME_STAMP>is the original response of LLM, which will be replaced by the time result decoded by the Time Decoder.



Briefly describe this video.

The video showcases a fencing match between two competitors, with a referee and spectators watching.

When did the man on the left turn his face back and walk away?

The action happens at 15.85s - 18.53s.

Why did the man on the left turn his face back and walk away?
0. he was injured
1. he wanted to sit down
2. pick something up
3. someone talking to him
4. end of match

4. end of match

Why are there people standing so far away from the plane?
0. climbed higher and further
1. restricted environment
2. can't go nearby
3. trying to land
4. prevent getting washed away

2. can't go nearby

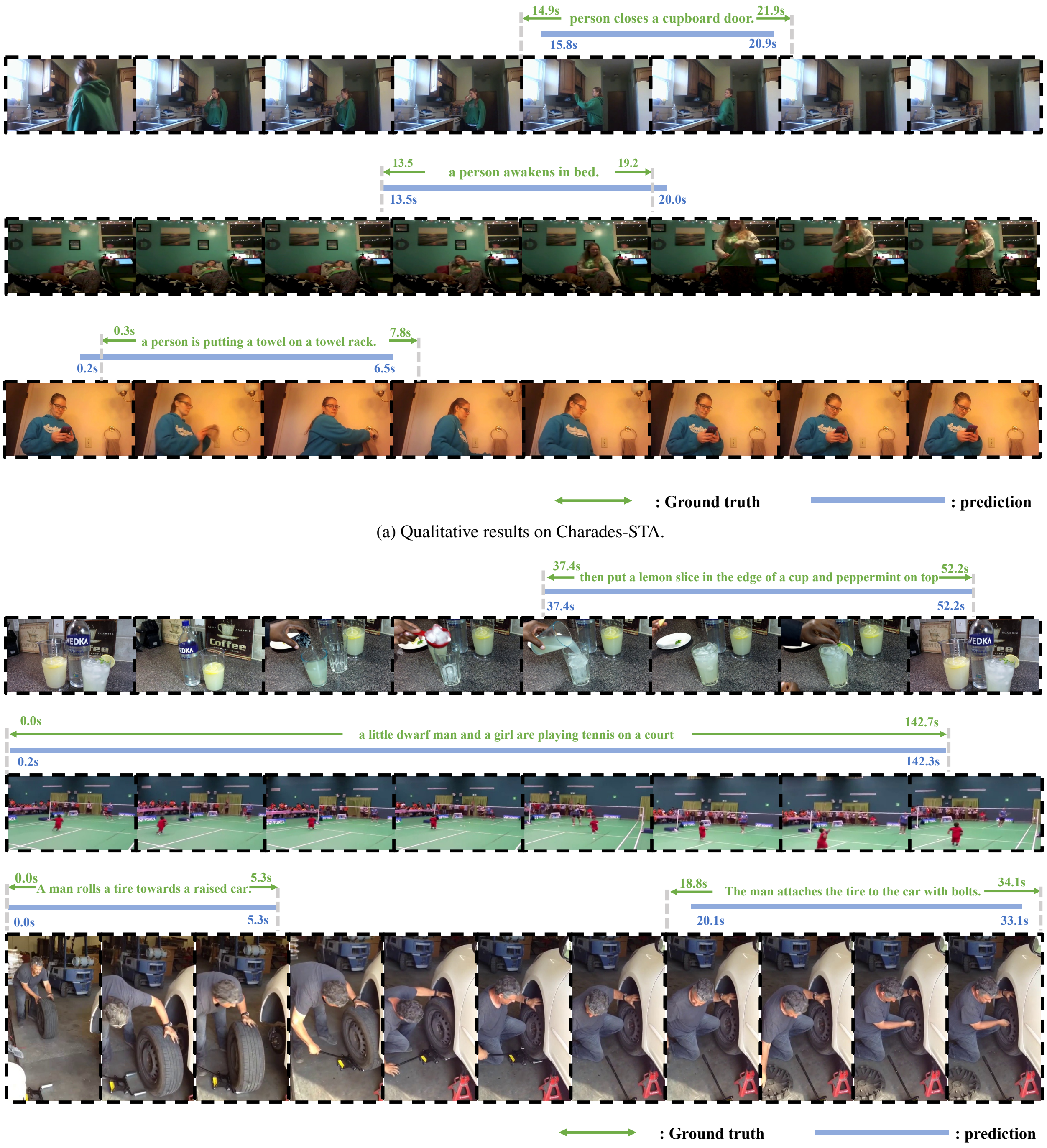
Why do we stay far away from landing planes?

We stay far away from landing planes to avoid being hit by them.

Give you the textual query: 'there is an orange barrier out of which people can stand'. When does the described content occur in the video?

It occurs at 0.38s - 11.26s.

Figure C. Qualitative results on multiple different tasks, such as temporal grounding task, video question answer, and open-ended question answer.


14.9s
person closes a cupboard door.
21.9s
15.8s
20.9s
13.5
a person awakens in bed.
19.2
13.5s
20.0s
0.3s
a person is putting a towel on a towel rack.
7.8s
0.2s
6.5s
: Ground truth
: prediction
37.4s
then put a lemon slice in the edge of a cup and peppermint on top
52.2s
37.4s
52.2s
0.0s
a little dwarf man and a girl are playing tennis on a court
142.7s
0.2s
142.3s
0.0s
A man rolls a tire towards a raised car.
5.3s
0.0s
5.3s
18.8s
The man attaches the tire to the car with bolts.
34.1s
20.1s
33.1s
: Ground truth
: prediction

(a) Qualitative results on Charades-STA.

(b) Qualitative results on ANet-Caption.

Figure D. Visual comparison of ground truth and prediction of moment retrieval. The arrow indicates the ground truth, and the rectangular progress bar indicates the prediction of DisTime.

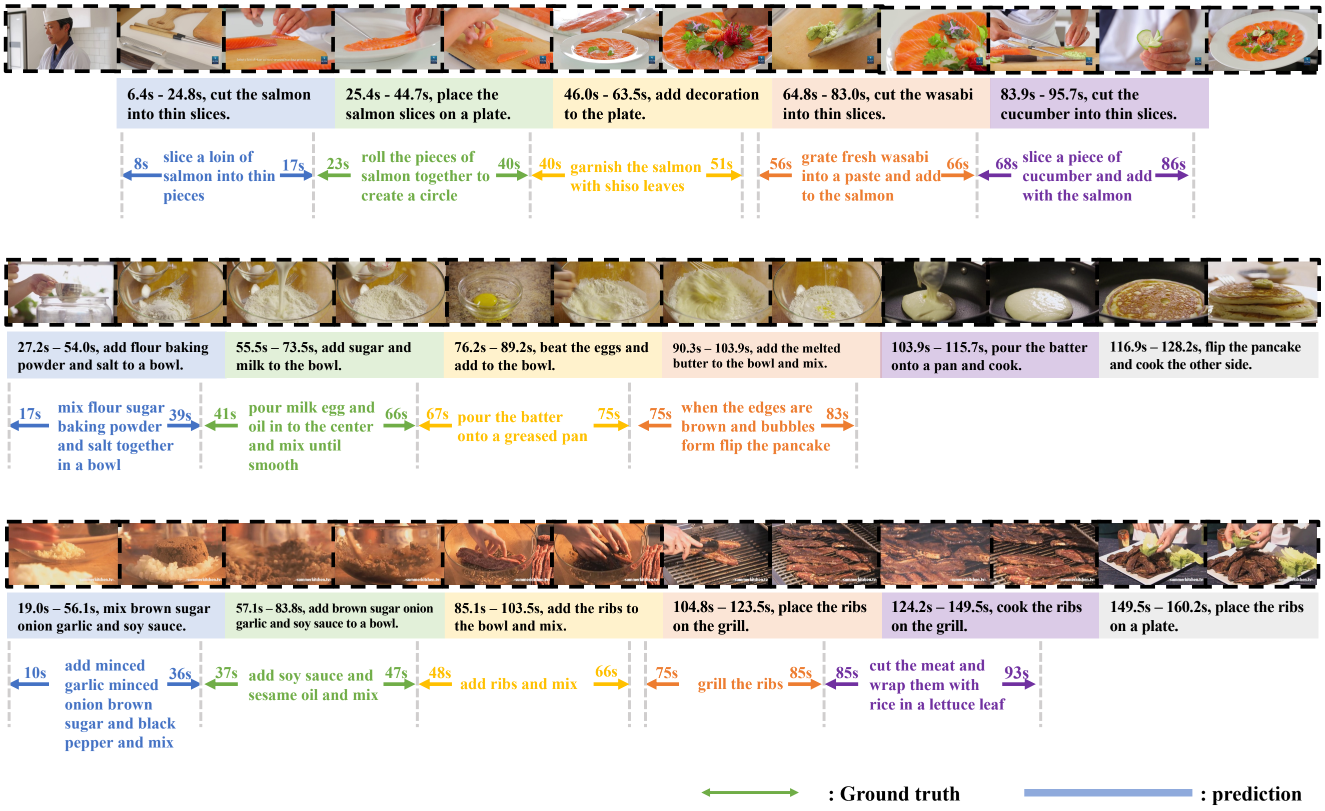
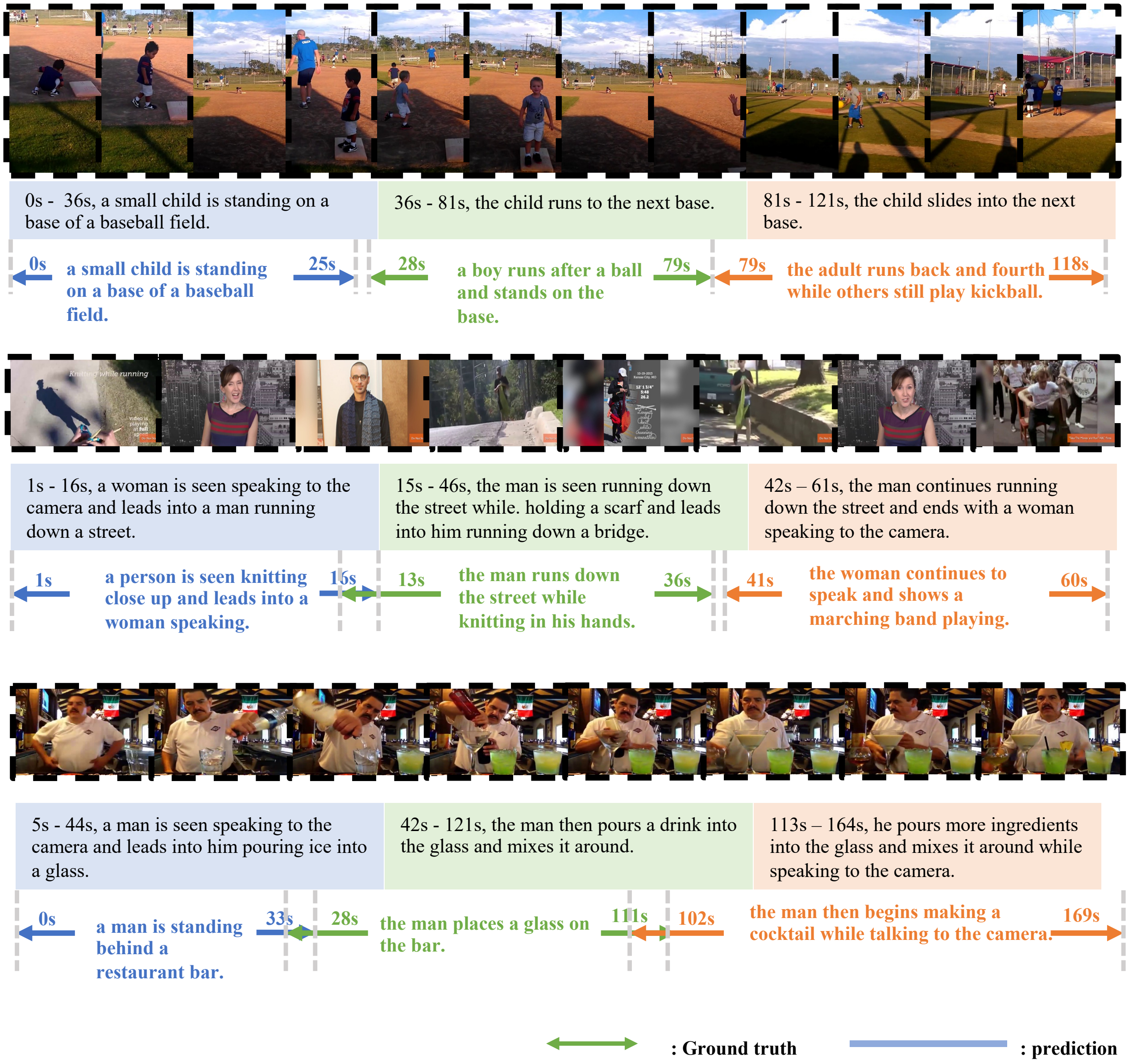

6.4s - 24.8s, cut the salmon into thin slices.
25.4s - 44.7s, place the salmon slices on a plate.
46.0s - 63.5s, add decoration to the plate.
64.8s - 83.0s, cut the wasabi into thin slices.
83.9s - 95.7s, cut the cucumber into thin slices.
8s slice a loin of salmon into thin pieces 17s
23s roll the pieces of salmon together to create a circle 40s
40s garnish the salmon with shiso leaves 51s
56s grate fresh wasabi into a paste and add to the salmon 66s
68s slice a piece of cucumber and add with the salmon 86s
27.2s – 54.0s, add flour baking powder and salt to a bowl.
55.5s – 73.5s, add sugar and milk to the bowl.
76.2s – 89.2s, beat the eggs and add to the bowl.
90.3s – 103.9s, add the melted butter to the bowl and mix.
103.9s – 115.7s, pour the batter onto a pan and cook.
116.9s – 128.2s, flip the pancake and cook the other side.
17s mix flour sugar baking powder and salt together in a bowl 39s
41s pour milk egg and oil in to the center and mix until smooth 66s
67s pour the batter onto a greased pan 75s
75s when the edges are brown and bubbles form flip the pancake 83s
19.0s – 56.1s, mix brown sugar onion garlic and soy sauce.
57.1s – 83.8s, add brown sugar onion garlic and soy sauce to a bowl.
85.1s – 103.5s, add the ribs to the bowl and mix.
104.8s – 123.5s, place the ribs on the grill.
124.2s – 149.5s, cook the ribs on the grill.
149.5s – 160.2s, place the ribs on a plate.
10s add minced garlic minced onion brown sugar and black pepper and mix 36s
37s add soy sauce and sesame oil and mix 47s
48s add ribs and mix 66s
75s grill the ribs 85s
85s cut the meat and wrap them with rice in a lettuce leaf 93s
: Ground truth
: prediction

Figure E. Visualization of ground truth and prediction from YouCook2. The arrow indicates the ground truth, and the rectangular progress bar indicates the prediction of DisTime.



0s - 36s, a small child is standing on a base of a baseball field.

36s - 81s, the child runs to the next base.

81s - 121s, the child slides into the next base.

a small child is standing on a base of a baseball field.

a boy runs after a ball and stands on the base.

the adult runs back and fourth while others still play kickball.

1s - 16s, a woman is seen speaking to the camera and leads into a man running down a street.

15s - 46s, the man is seen running down the street while. holding a scarf and leads into him running down a bridge.

42s – 61s, the man continues running down the street and ends with a woman speaking to the camera.

a person is seen knitting close up and leads into a woman speaking.

the man runs down the street while knitting in his hands.

the woman continues to speak and shows a marching band playing.

5s - 44s, a man is seen speaking to the camera and leads into him pouring ice into a glass.

42s - 121s, the man then pours a drink into the glass and mixes it around.

113s – 164s, he pours more ingredients into the glass and mixes it around while speaking to the camera.

a man is standing behind a restaurant bar.

the man places a glass on the bar.

the man then begins making a cocktail while talking to the camera.

Figure F. Visualization of ground truth and prediction from ANet-Caption. The arrow indicates the ground truth, and the rectangular progress bar indicates the prediction of DisTime.