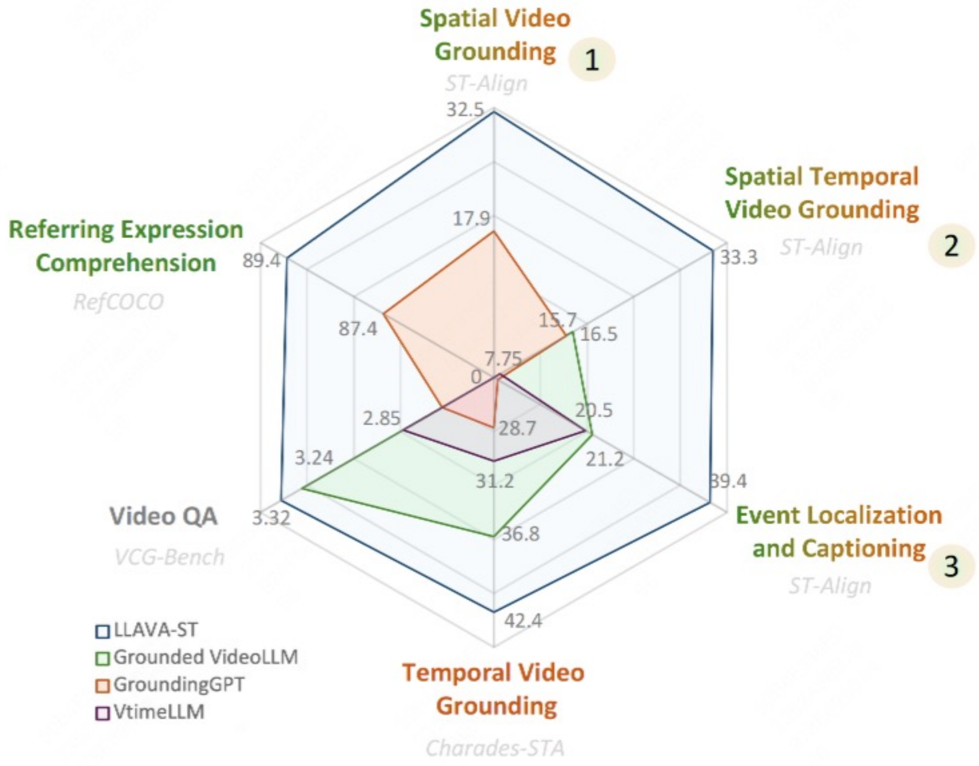
LLaVA-ST: A Multimodal Large Language Model for Fine-Grained Spatial-Temporal Understanding

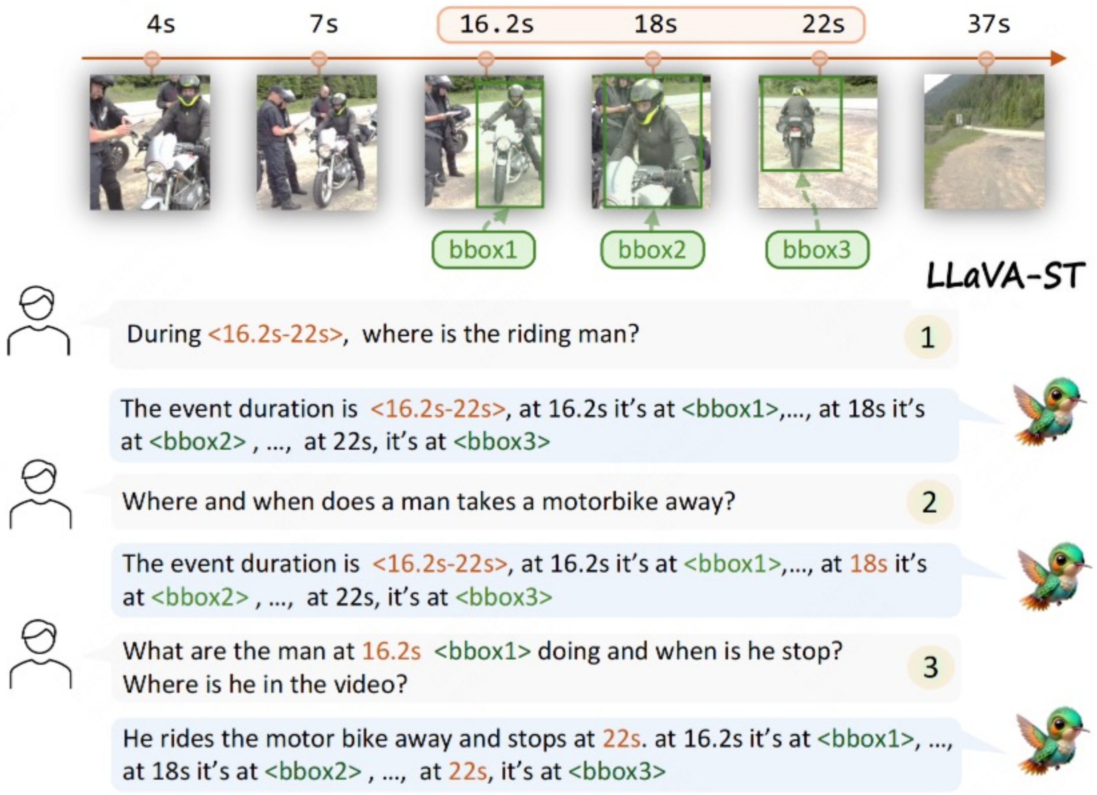
Hongyu Li[1*] Jinyu Chen[1*] Ziyu Wei[1*] Shaofei Huang[2,3] Tianrui Hui[2]
Jialin Gao[4†] Xiaoming Wei[4] Si Liu[1†]

([*]Equal contribution , [†]Corresponding author)

[1]BUAA [2]HFUT [3]CAS [4]Meituan

Overall Introduction





- ***LLaVA-ST*** is the **first MLLM capable of simultaneously processing spatial-temporal fine-grained understanding tasks** and demonstrates high performance across various tasks of fine-grained multimodal understanding.
- Examples of spatial-temporal interleaved fine-grained understanding tasks in the proposed **ST-Align Benchmark**, which include Spatial Temporal Video Grounding (STVG), Event Localization and Captioning (ELC), and Spatial Video Grounding (SVG).

Motivation

We refer to tasks that require processing visual coordinates based on linguistic input as fine-grained multimodal understanding, designing a unified MLLM for spatial-temporal fine-grained understanding tasks faces two major challenges:

- ➢ Difficulties in multimodal coordinate alignment
- ➢ Challenges in preserving visual details

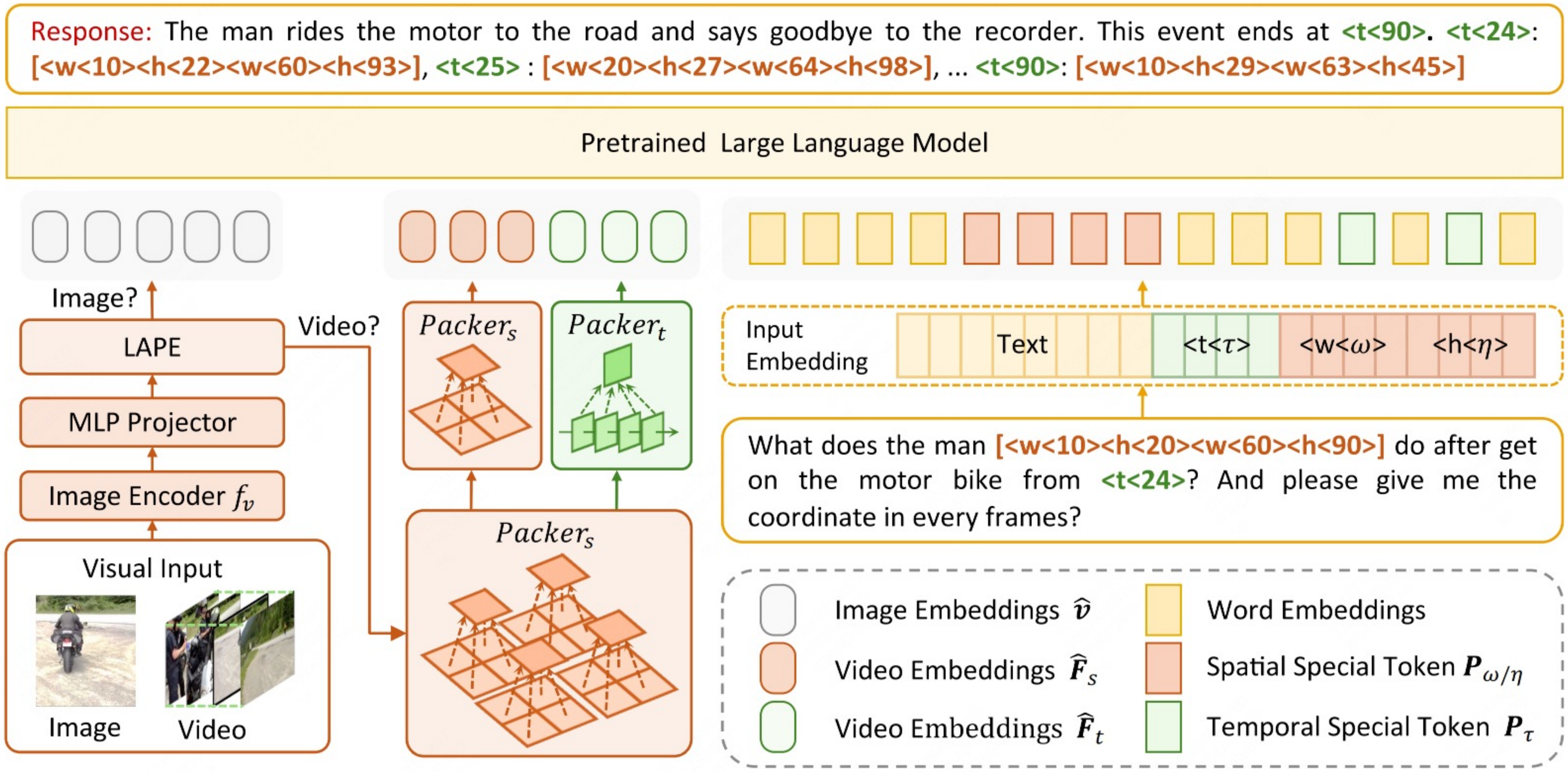
Our contributions are summarized as follows:

i) We propose **LLaVA-ST**, the first MLLM capable of end-to-end processing fine-grained spatial, temporal, and interleaved fine-grained multimodal understanding tasks. We introduce the LAPE to reduce the difficulty of aligning coordinate features between vision and language and STP to preserve fine-grained spatiotemporal context in the video feature compression process.

ii) We propose a progressive training strategy for LLaVA-ST, enabling the model to progressively learn content alignment, coordinate alignment and multi-task ability. To support the training process, we construct the **STAlign dataset** which includes 15 different tasks and 228K newly conducted data samples.

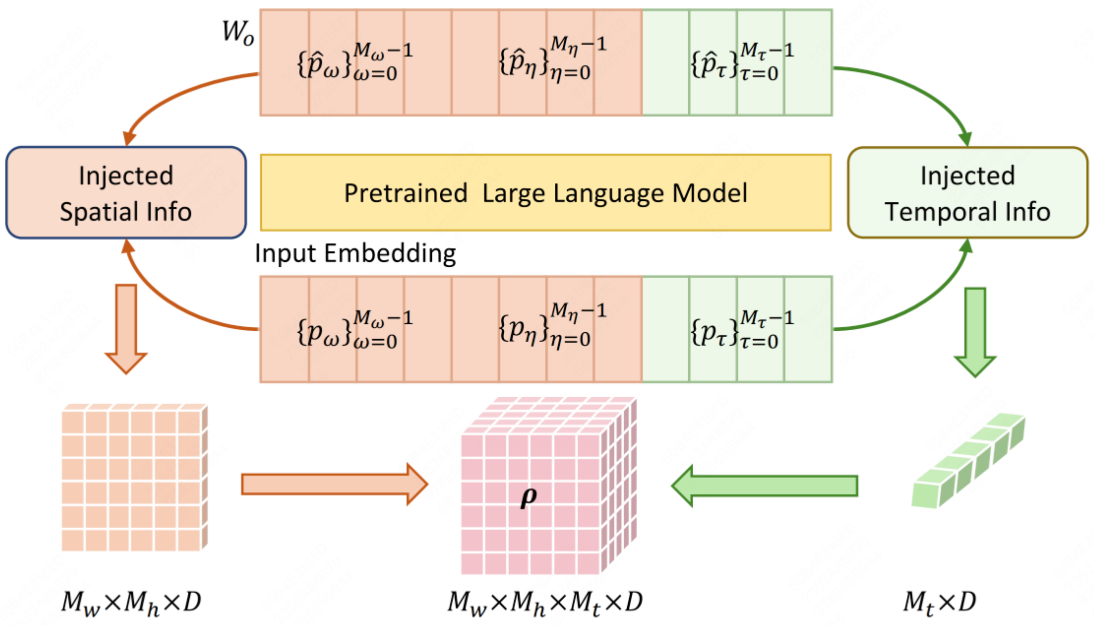
iii) We perform experiments for on 11 different benchmarks, including TVG, video QA, REC and spatial-temporal interleaved tasks, and the results demonstrate the outstanding capabilities of LLaVA-ST.

Model Architecture



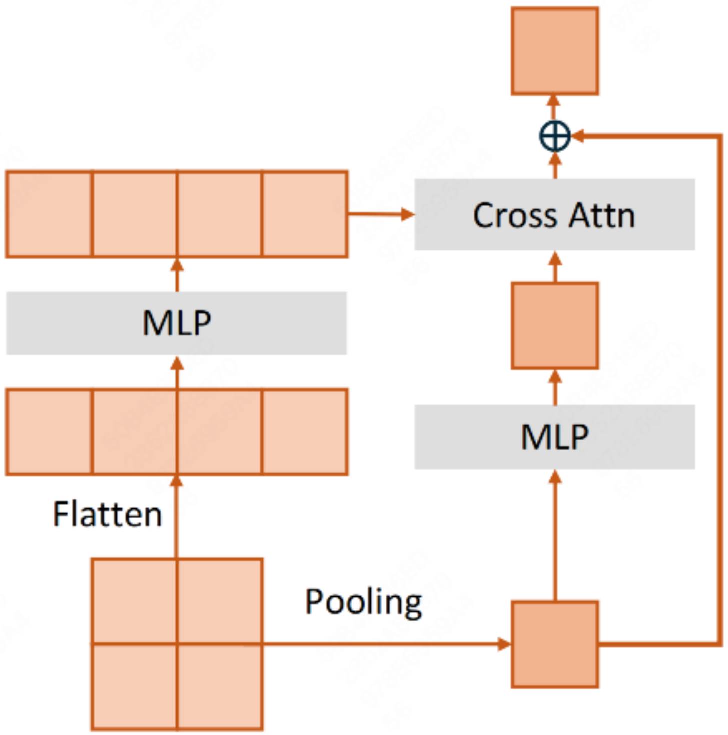
We introduce discrete special tokens to represent spatiotemporal coordinates within the language modality. LAPE embed these coordinate representations into the visual feature space. Furthermore, the STP module utilizes a two-stream packing mechanism to efficiently compress the features.

Design of LAPE



To avoid the inefficiency of tokenizing numerical text, we extend the vocabulary of the LLM by introducing special tokens for coordinates. These special tokens are incorporated into the input text embeddings and output layers of the LLM. We then embed the spatial tokens into the visual features as positional embeddings. LAPE leverages coordinate-related input text embeddings and features within the output layer matrix as visual positional embeddings.

Design of STP



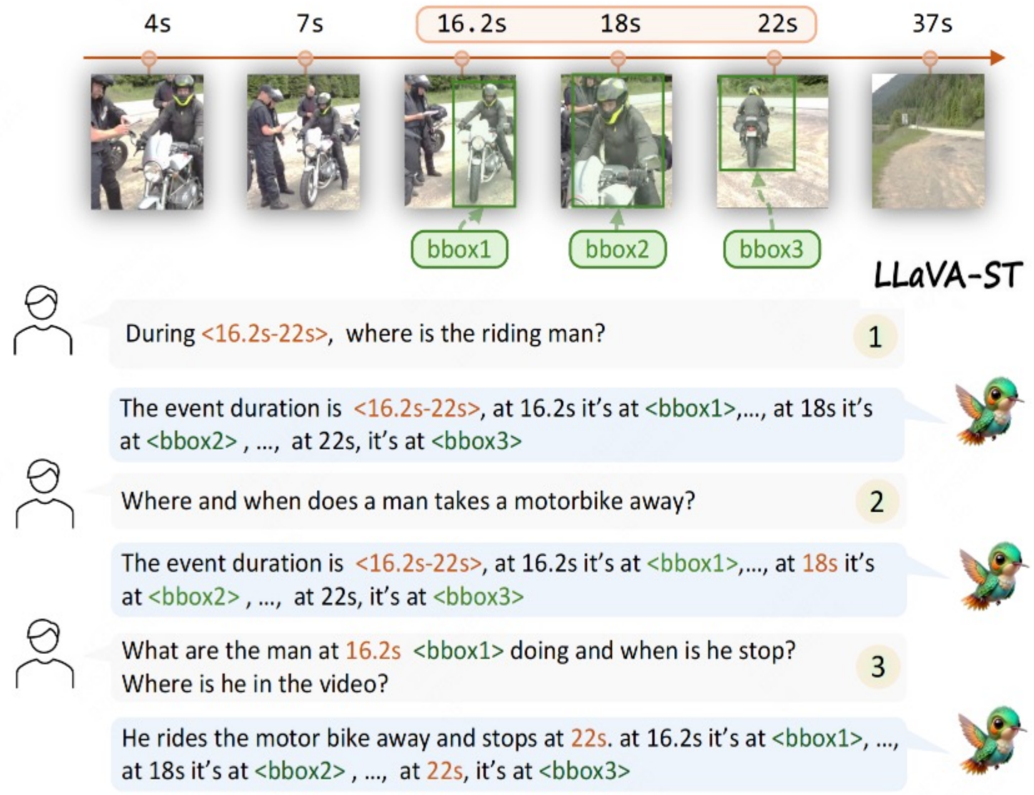
Above illustrates the calculation process of a patch feature in $packer_s$. After obtaining the region feature and the pooled feature, each is processed through an MLP, followed by cross-attention computation. The pooled feature is then added to the output feature through a residual connection. $packer_t$ share the similar architecture but with temporal pooling dimension.

STAlign Dataset

Training Stage	Task	# of Samples	Data Source
Content Alignment (Stage-1)	Video Captioning	1.28M	WebVid-10M [2], Panda-70M [6], InternVid-10M [63]
Coordinate Alignment (Stage-2)	Temporal Video Grounding	149K	VTimeLLM-Stage2 [15]
	Dense Video Captioning	92K	VTimeLLM-Stage2 [15], Moment-10M [50], InternVid-G [63]
	Temporal Referring	95K	VTimeLLM-Stage2 [15], InternVid-G [62]
	Referring Expression Comprehension	101K	GranD [51]*
	Dense Grounded Captioning	150K	GranD [51]*
	Region Caption	100K	GranD [51]*
Multi-Task Instruction Tuning (Stage-3)	Temporal Grounded Conversation	459K	ANet-RTL [16], Moment-10M [50], Grounded-VideoLLM [58]
	Temporal Video Grounding	84K	DiDeMo [1], HiREST [77], QuerYD [48], VTG-IT [13]
	Dense Video Caption	41K	COIN [56], ViTT [17], YouCook2 [81], VTG-IT [13]
	Image Grounded Conversation	190K	MUSE [53]*, Flickr30k Entities [75]
	Referring Expression Comprehension	288K	RefCOCO [25], RefCOCO+ [25], RefCOCOg [47]
	Dense Grounded Captioning	22K	Flickr30k Entities [75]
	Region Caption	288K	RefCOCO [25], RefCOCO+ [25], RefCOCOg [47]
	Spatial-Temporal Video Grounding	81K	Self collected
	Event Localization and Captioning	36K	Self collected
	Spatial Video Grounding	81K	Self collected
	Converstation	222K	VCG-Plus-112K [45], Videochatgpt-100K [43], Videochat2-Conv [33]
	VideoQA	338K	EgoQA [11], NExT-QA [68], Intent-QA [31], AGQA [12], STAR [66], CLEVRER [73], WebVid-QA [72]
	Classification	66K	SthSthV2 [46], Kinetics [32]
	Video Captioning	168K	TextVR [67], YouCook2 [81], WebVid [2], ShareGPT4Video [5]

To enhance the stability of the training process and improve the model's final performance, we partitioned the training data into three stages based on data quality and the granularity of visual-text alignment. These stages are: content alignment, coordinate alignment, and multi-task instruction tuning.

STAlign Benchmark



To address the current deficiency of data involving fine-grained multimodal understanding with spatiotemporal interleaving, we employ GPT-4-turbo to revise and enhance the textual annotations of VidSTG, adapting them for the following three tasks:

i) **Spatial Video Grounding (SVG)**: based on the event duration and textual description, locate the tracklet of the event subject.

ii) **Spatial Temporal Video Grounding (STVG)**: localizing the spatiotemporal tube of the event subject in the video based on the event description.

iii) **Event Localization and Captioning (ELC)**: given a starting position and the bounding box of the event subject, locate the ending position of the event and provide a description of the event.

Results on Fine-Grained Spatial-Temporal Grounding

Model	Spatial-Temporal Video Grounding				Event Localization and Captioning					Spatial Video Grounding		
	tIoU@0.5	m_tIoU	sIoU@0.5	m_sIoU	tIoU@0.5	m_tIoU	sIoU@0.5	m_sIoU	METEOR	sIoU@0.3	sIoU@0.5	m_sIoU
GroundingGPT	7.1	12.2	2.9	9.2	4.8	6.6	2.1	6.4	8.2	19.7	5.4	17.9
VTimeLLM	7.1	15.5	-	-	33.1	40.3	-	-	6.0	-	-	-
Grounded-VideoLLM	30.0	33.0	-	-	53.1	56.4	-	-	7.2	-	-	-
LLaVA-ST	**44.6**	**43.8**	**21.1**	**22.8**	**60.4**	**60.0**	**32.4**	**33.5**	**24.7**	**47.2**	**30.9**	**32.5**

Model	Charades-STA			
	R@0.3	R@0.5	R@0.7	mIoU
Video-LLaMA-7B	25.2	10.6	3.4	16.8
SeViLA-3B	27.0	15.0	5.8	18.3
Video-ChatGPT-7B	27.2	6.2	1.9	19.7
Valley-7B	28.4	1.8	0.3	21.4
VideoChat2-7B	38.0	14.3	3.8	24.6
VideoChat-7B	32.8	8.6	0.0	25.9
Momenter-7B	42.6	26.6	11.6	28.5
VTimeLLM-7B	51.0	27.5	11.4	31.2
GroundingGPT-7B	-	29.6	11.9	28.7
TimeChat-7B	-	32.2	13.4	-
VTG-LLM-7B	-	33.8	15.7	-
HawkEye-7B	50.6	31.4	14.5	33.7
Grounded-VideoLLM-4B	54.2	36.4	19.7	36.8
LLaVA-ST-7B	**63.1**	**44.8**	**23.4**	**42.4**

Model	RefCOCO			RefCOCO+			RefCOCOg	
	val	test-A	test-B	val	test-A	test-B	val-u	test-u
Generalist models w. full resolution input								
Ferret v2-7B	92.8	94.7	88.7	87.4	92.8	79.3	89.4	89.3
Generalist models w/o full resolution input								
OFA-L	80.0	83.7	76.4	68.3	76.0	61.8	67.6	67.6
Shikra-7B	87.0	90.6	80.2	81.6	87.4	72.1	82.3	82.2
GroundingGPT-7B	88.0	91.6	82.5	81.6	87.2	73.2	81.7	82.0
MiniGPT-v2-7B	88.7	91.7	85.3	80.0	85.1	74.5	84.4	84.7
Ferret-7B	87.5	91.4	82.5	80.8	87.4	73.1	83.9	84.8
VistaLLM-7B	88.1	91.5	83.0	82.9	89.8	74.8	83.6	84.4
Elysium-7B	89.1	92.1	85.0	82.9	88.9	75.6	82.9	83.6
Groma-7B	89.5	92.1	**86.3**	83.9	88.9	78.1	86.3	87.0
LLaVA-ST-7B	**90.1**	**93.2**	85.0	**86.0**	**91.3**	**78.8**	**86.7**	**87.4**

Results on General Video QA

Model	MSVD-QA		MSRVTT-QA		VCG-Bench					
	Acc.	Score	Acc.	Score	CI	DO	CU	TU	CO	Avg.
Video-LLaMA [79]	51.6	2.5	29.6	1.8	1.96	2.18	2.16	1.82	1.79	1.98
Video-ChatGPT [44]	64.9	3.3	49.3	2.8	2.50	2.57	2.69	2.16	2.20	2.42
GroundingGPT [36]	67.8	3.7	51.6	3.1	-	-	-	-	-	-
Momentor [50]	68.9	3.6	55.6	3.0	-	-	-	-	-	-
MovieChat [55]	75.2	3.8	52.7	2.6	2.76	2.93	3.01	2.24	2.42	2.67
VTimeLLM [15]	-	-	-	-	2.78	3.10	3.40	2.49	2.47	2.85
LongVLM [65]	70.0	3.8	59.8	3.3	2.76	2.86	3.34	2.39	3.11	2.89
VideoChat2 [35]	70.0	3.9	54.1	3.3	3.02	2.88	3.51	2.66	2.81	2.98
Chat-UniVi [24]	65.0	3.6	54.6	3.1	2.89	2.91	3.46	2.89	2.81	2.99
LITA [16]	-	-	-	-	2.94	2.98	3.43	2.68	3.19	3.04
P-LLaVA-7B [70]	76.6	**4.1**	62.0	3.5	3.21	2.86	3.62	2.33	2.93	3.12
ST-LLM [38]	74.6	3.9	63.2	3.4	3.23	3.05	3.74	2.93	2.81	3.15
VideoGPT+ [45]	-	-	-	-	3.27	**3.18**	3.74	2.83	3.39	3.28
Elysium [59]	75.8	3.7	**67.5**	3.2	-	-	-	-	-	-
Grounded-VideoLLM [58]	**76.3**	**4.1**	60.3	**3.6**	**3.34**	2.94	3.66	**3.12**	3.14	3.24
LLaVA-ST	75.9	**4.1**	59.0	3.5	3.29	**3.18**	**3.80**	2.9	**3.41**	**3.32**

Model	Avg.	AS	AP	AA	FA	UA	OE	OI	OS	MD	AL	ST	AC	MC	MA	SC	FP	CO	EN	ER	CI
VideoChatGPT [44]	32.7	23.5	26.0	62.0	22.5	26.5	54.0	28.0	40.0	23.0	20.0	31.0	30.5	25.5	39.5	48.5	29.0	33.0	29.5	26.0	35.5
VideoLLaMA [79]	34.1	27.5	25.5	51.0	29.0	39.0	48.0	40.5	38.0	22.5	22.5	43.0	34.0	22.5	32.5	45.5	32.5	40.0	30.0	21.0	37.0
VideoChat [34]	35.5	33.5	26.5	56.0	33.5	40.5	53.0	40.5	30.0	25.5	27.0	48.5	35.0	20.5	42.5	46.0	26.5	41.0	23.5	23.5	36.0
TimeChat [52]	38.5	40.5	36.0	61.0	32.5	53.0	53.5	41.5	29.0	19.5	26.5	66.5	34.0	20.0	43.5	42.0	36.5	36.0	29.0	35.0	35.0
Video-LLaVA [79]	43.0	46.0	42.5	56.5	39.0	53.5	53.0	48.0	41.0	29.0	31.5	82.5	45.0	26.0	53.0	41.5	33.5	41.5	27.5	38.5	31.5
P-LLaVA-7B [70]	46.6	58.0	49.0	55.5	41.0	61.0	56.0	61.0	36.0	23.5	26.0	82.0	39.5	42.0	52.0	45.0	42.0	53.5	30.5	48.0	31.0
VideoChat2 [35]	51.1	66.0	47.5	83.5	49.5	60.0	58.0	71.5	**42.5**	23.0	23.0	88.5	39.0	42.0	58.5	44.0	49.0	36.5	35.0	40.5	**65.5**
ShareGPT4Video [5]	51.2	49.5	39.5	79.5	40.0	54.5	82.5	54.5	32.5	**50.5**	41.5	84.5	35.5	62.5	75.0	51.0	25.5	46.5	28.5	39.0	51.5
ST-LLM [38]	54.9	66.0	53.5	84.0	44.0	58.5	80.5	73.5	38.5	42.5	31.0	86.5	36.5	56.5	78.5	43.0	44.5	46.5	34.5	41.5	58.5
VideoGPT+ [45]	58.7	69.0	60.0	83.0	48.5	66.5	85.5	75.5	36.0	44.0	34.0	89.5	39.5	71.0	90.5	45.0	53.0	50.0	29.5	44.0	60.0
Grounded-VideoLLM [58]	59.4	76.0	**75.5**	77.0	48.0	67.5	85.5	77.0	34.5	39.5	**59.5**	86.5	**44.5**	60.5	79.0	**51.5**	49.0	46.0	35.0	42.5	54.0
LLaVA-ST	**64.2**	**77.0**	69.0	**91.5**	**50.0**	**68.5**	**93.5**	**84.5**	40.0	44.5	49.5	**93.0**	44.0	**77.5**	**97.5**	41.0	**57.0**	**56.5**	**37.0**	**49.0**	63.0

Ablation Study

Method	RefCOCO (val) Acc@0.5	ST-Align (STVG) m_tIoU	m_sIoU	Charades-STA mIoU
LLaVA-ST	**89.6**	**43.5**	**21.0**	**38.5**
w/o LAPE	88.7	42.1	17.9	36.1
w/o STP	89.2	42.0	18.2	35.5

Table 7. Ablation studies on *LLaVA-ST* modules. Results for REC, STVG, and TVG evaluated on RefCOCO, ST-Align, and Charades-STA. The LAPE and STP modules significantly improve *LLaVA-ST* performance.

Training Data			RefCOCO			Charades-STA		
T	S	T+S	val	test-A	test-B	R@0.3	R@0.5	mIoU
✓			-	-	-	48.7	32.0	34.6
✓	✓		**89.8**	**92.5**	**85.1**	52.9	34.1	32.3
✓	✓	✓	89.6	92.1	**85.1**	**58.0**	**39.8**	**38.5**

Table 8. Ablation studies on training data. Results evaluated on RefCOCO and Charades-STA for REC and TVG tasks. (T: Temporal, S: Spatial, T+S: Spatial temporal interleaved data)

Training Data	RefCOCO (val) ACC@0.5	ST-Align (STVG) m_tIoU	m_sIoU	Charades-STA mIoU
S.1 + S.3	88.4	37.5	16.6	33.3
S.1 + (S.2 + S.3)	**88.7**	41.2	18.3	**35.0**
S.1 + S.2 + S.3	88.6	**41.5**	**21.2**	33.3

Table 9. Ablation study about training stage. (S.2 + S.3) indicates mix training data of S.2 and S.3.

Model			ST-Align (STVG)		Charades-STA
packer_t	packer_s	token#	m_tIoU	m_sIoU	mIoU
25×9×9	0	2025	36.2	8.3	8.8
0	100×5×5	2500	42.5	18.9	36.7
20×9×9	100×3×3	2520	**43.5**	**21.0**	**38.5**

Table 10. The ablation study about packer_s and packer_t. We display the feature resolutions of the outputs of packer_s and packer_t in spatial and temporal dimensions.

Quantitative Results

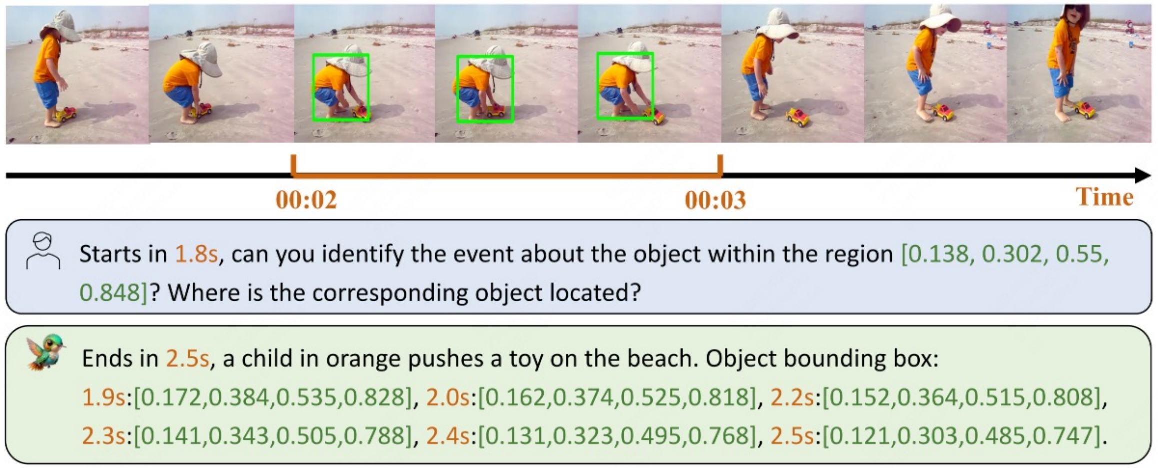


Figure 8. Visualized result of event localization and captioning.

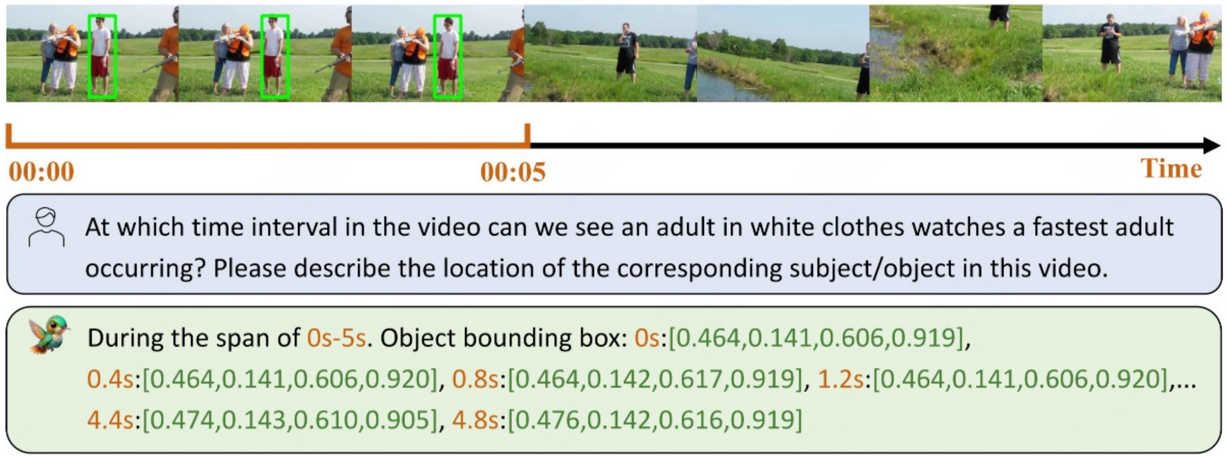


Figure 7. Visualized result of spatial-temporal video grounding.

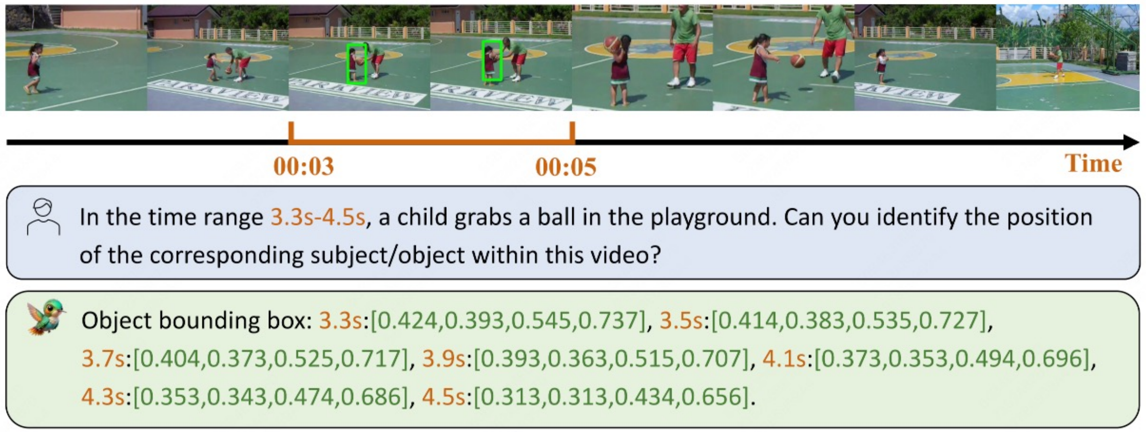


Figure 9. Visualized result of spatial video grounding.

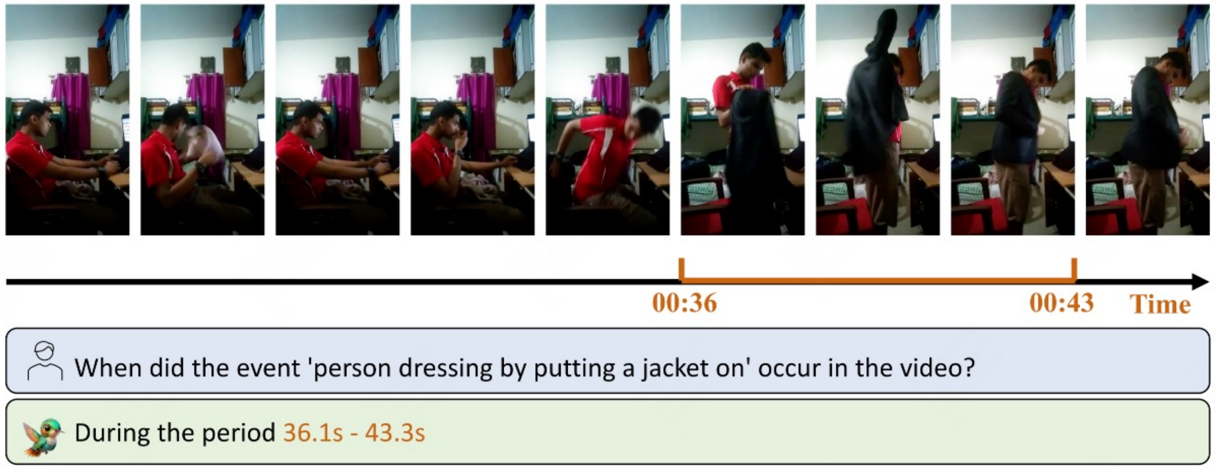


Figure 10. Visualized result of temporal video grounding.

Quantitative Results

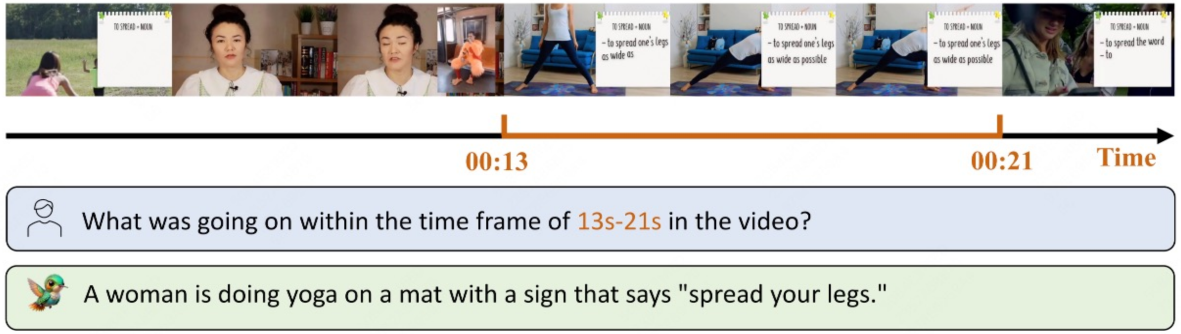


Figure 11. Visualized result of temporal referring.

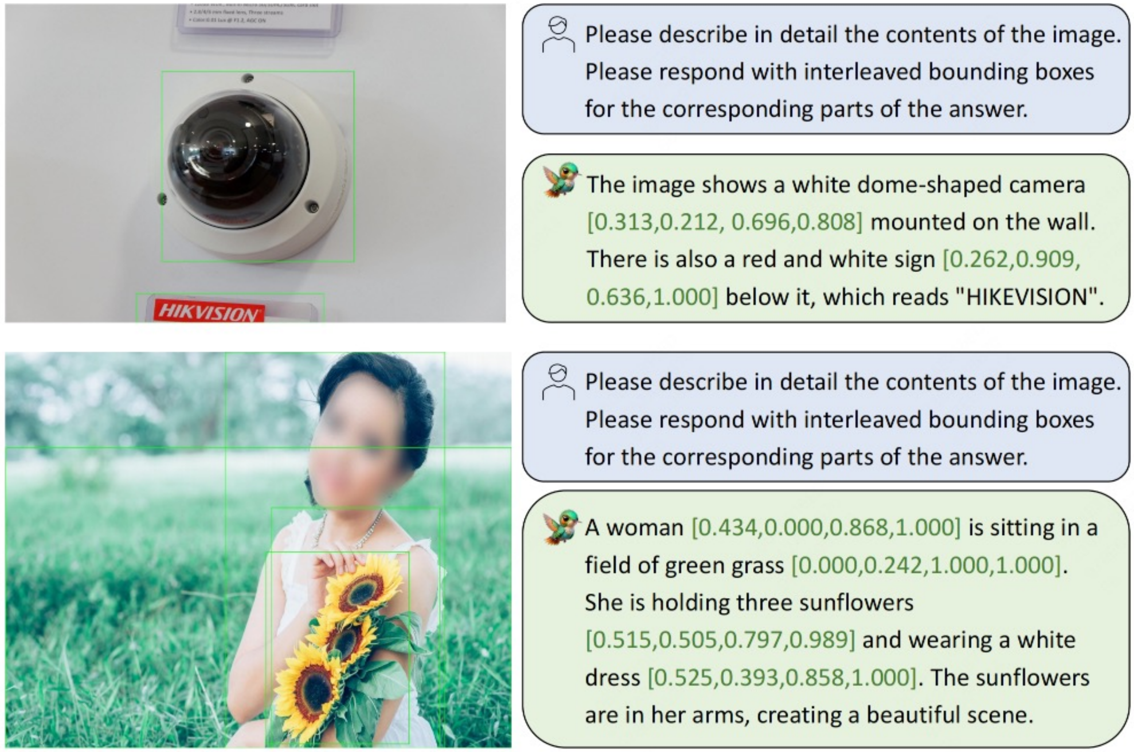


Figure 12. Visualized result of dense grounded captioning.

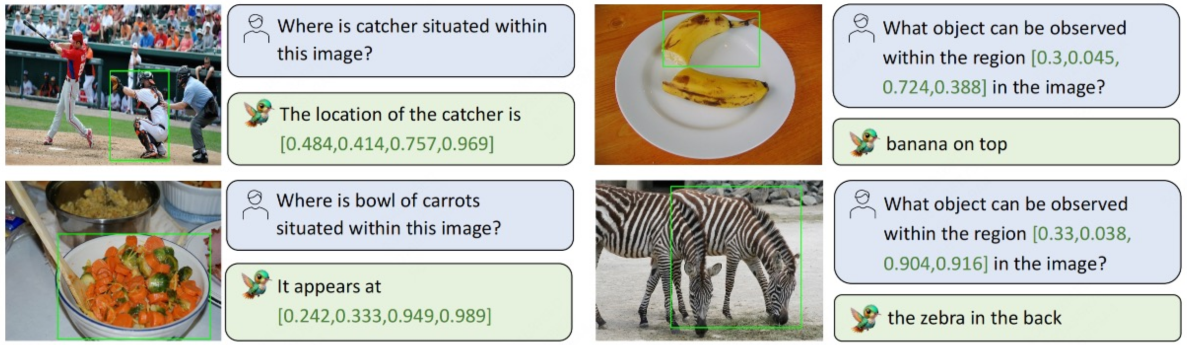


Figure 13. Referring expression comprehension and region caption visualize result.

Quantitative Results

Can you describe the video in detail?

Sure, the video shows a man sitting in a chair and watching TV. He is wearing a white shirt and black pants. The TV is placed on a stand next to him, and there are two chairs facing it. The room has a pink wall with a window that lets in natural light. There is also a staircase behind the man. The man seems to be engrossed in the TV show, and he is not moving much. The lighting in the room is bright, and the colors of the furniture and walls are muted. Overall, the video depicts a peaceful scene of a man enjoying his favorite TV show while relaxing in a comfortable setting.

Where is the scene in the video?

The scene in the video is set in a living room.

Can you write a short poem inspired from the video?

Sure, here's a short poem inspired from the video:
A man in a chair,
Watching TV with ease,
Relaxing and enjoying,
His favorite show!

Figure 14. Example for multi-turn open-ended video QA of LLaVA-ST.

Quantitative Results

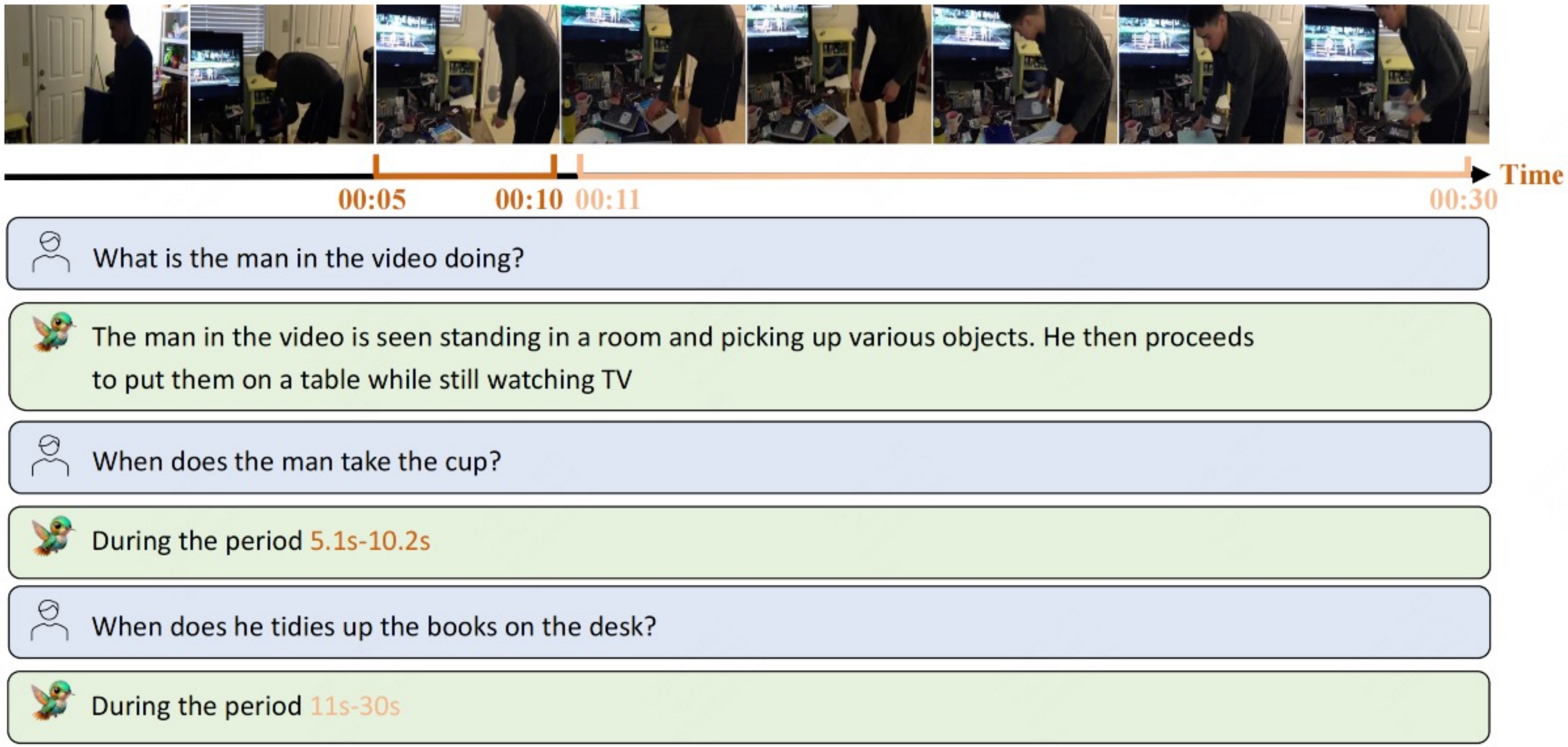


Figure 15. Example for multi-turn open-ended video QA of LLaVA-ST.

Thanks for Listening!