NEURAL INFORMATION PROCESSING SYSTEMS

ReXTime: A Benchmark Suite for Reasoning-Across-Time in Videos

Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu,
Yen-Chun Chen, Yu-Chiang Frank Wang
National Taiwan University, Microsoft

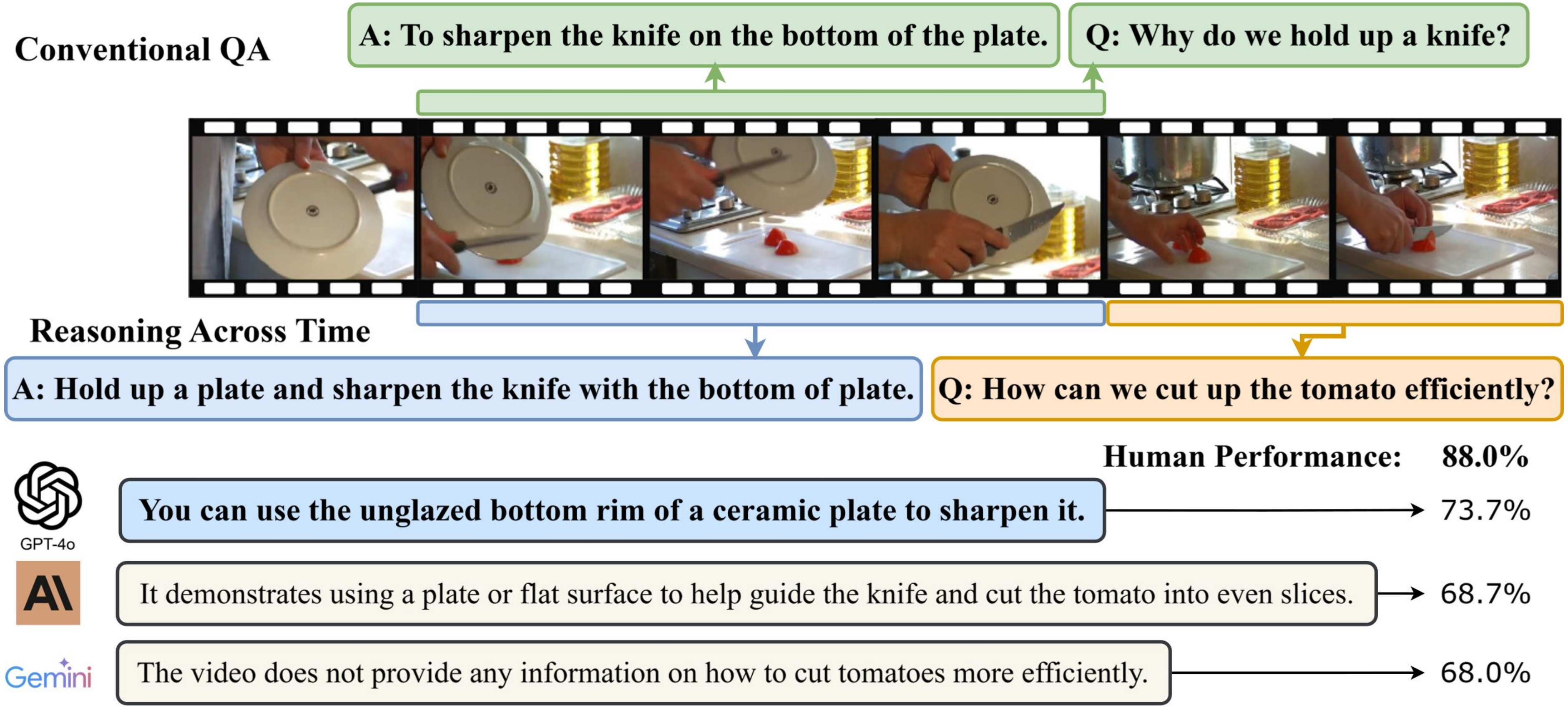
National Taiwan University

Microsoft

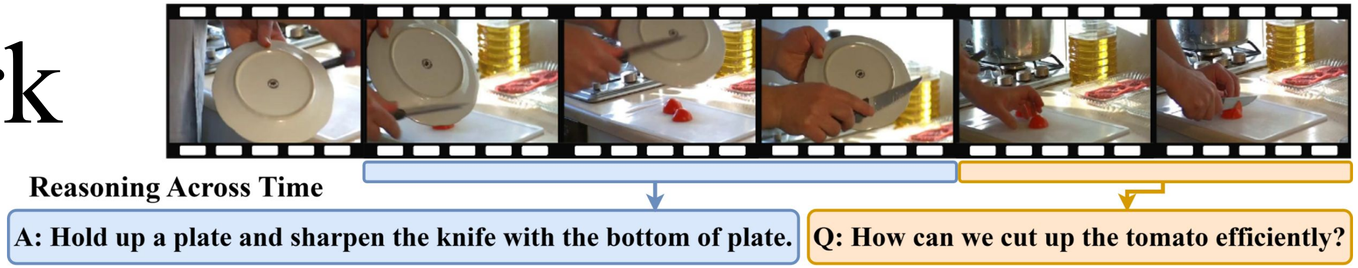
Project Page

Introduction

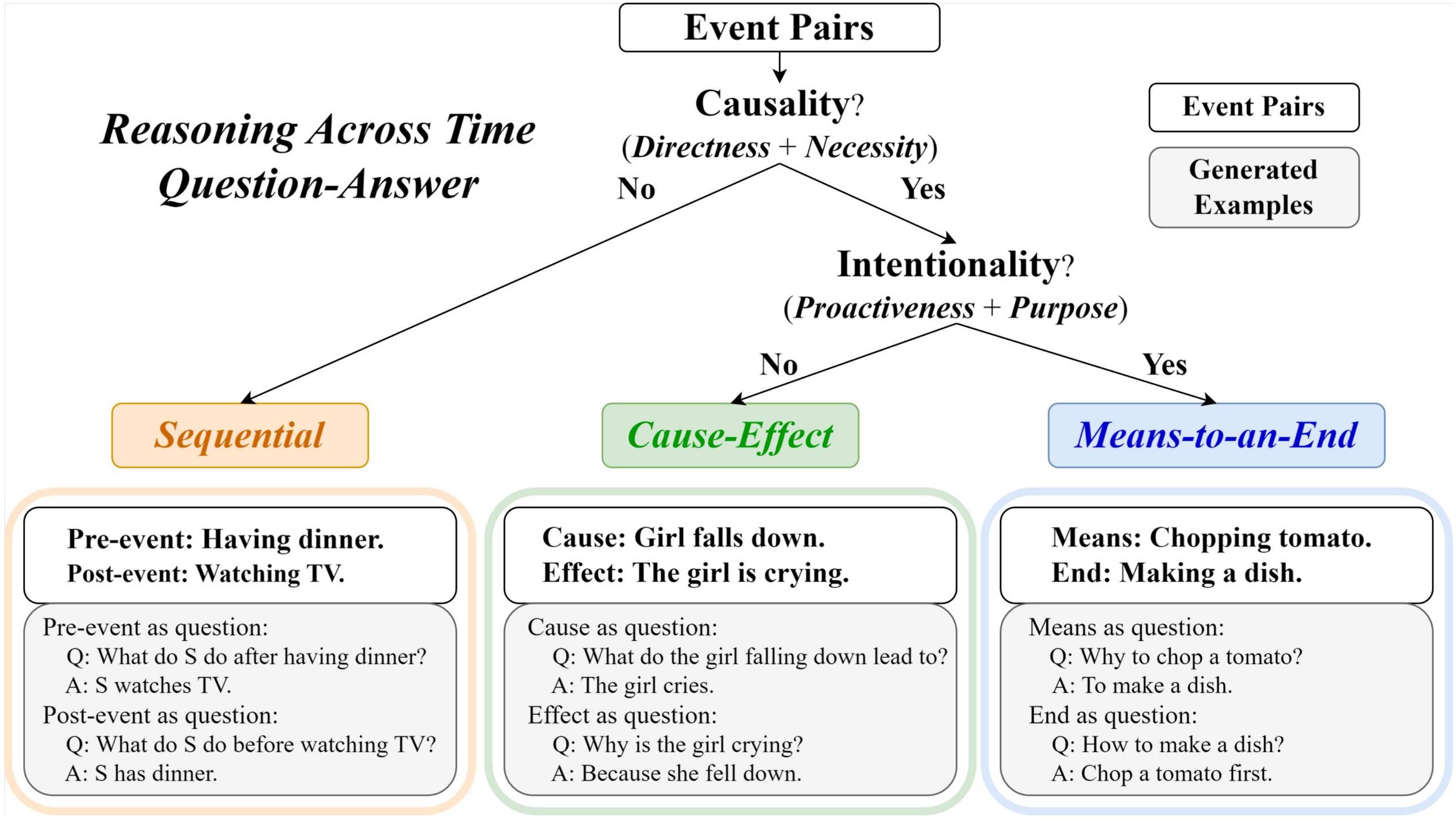
- QA with Reasoning-Across-Time
  - Question and answer each belongs to different time spans.



ReXTime Benchmark

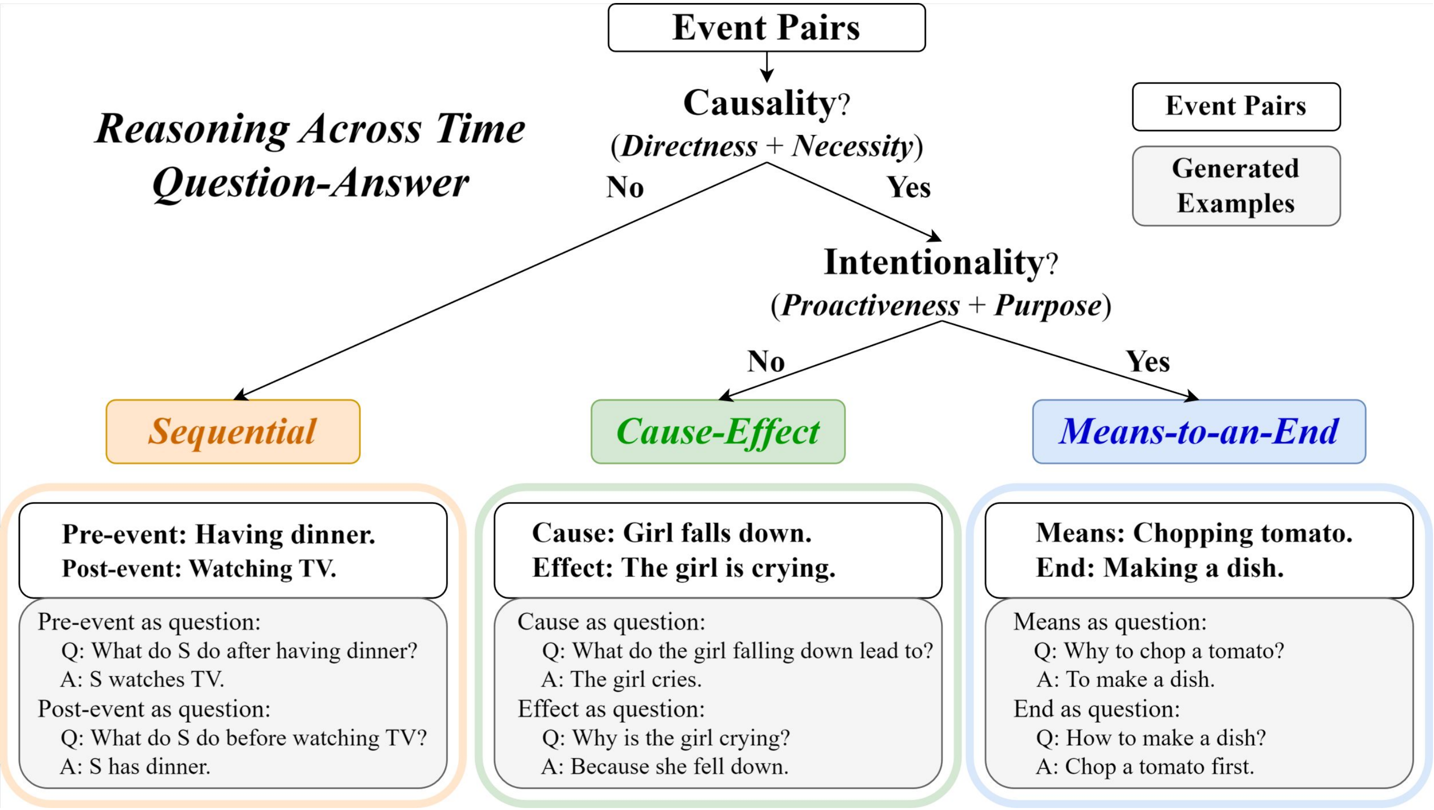


- Grounding-VQA data pairs:
  - ❑ Sequential
  - ❑ Cause-Effect
  - ❑ Means-to-an-End
- ReXTime tasks:
  - ❑ Multi-choice VQA
  - ❑ Moment localization



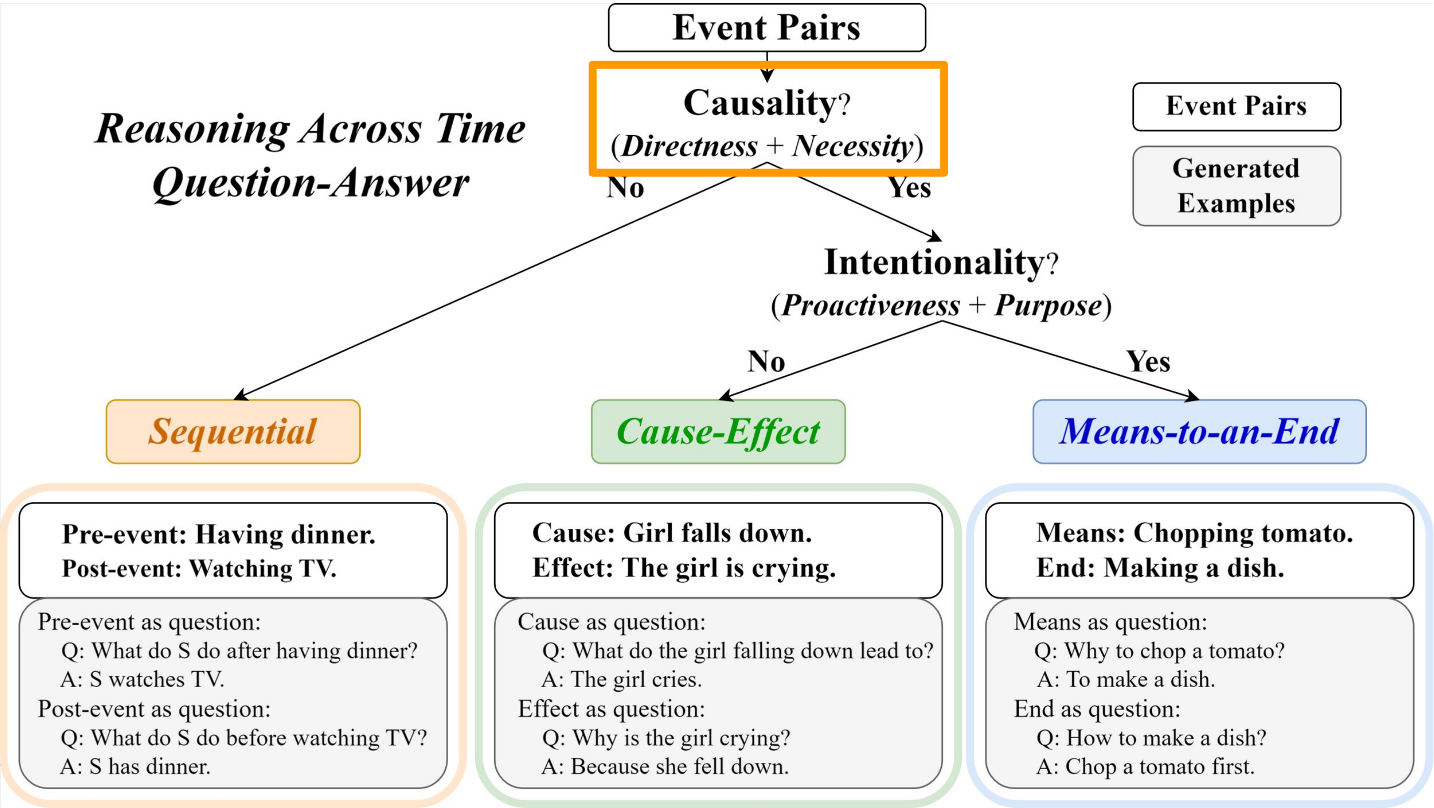
Grounding-VQA Classification Criteria

- ***Directness***: This criterion assesses the directness of the causal link between events.
- ***Necessity***: This criterion measures whether the second event is inevitable due to the first.
- ***Proactiveness***: This evaluates whether an event is carried out with deliberate intention.
- ***Purpose***: This evaluates whether the intention has been fulfilled.



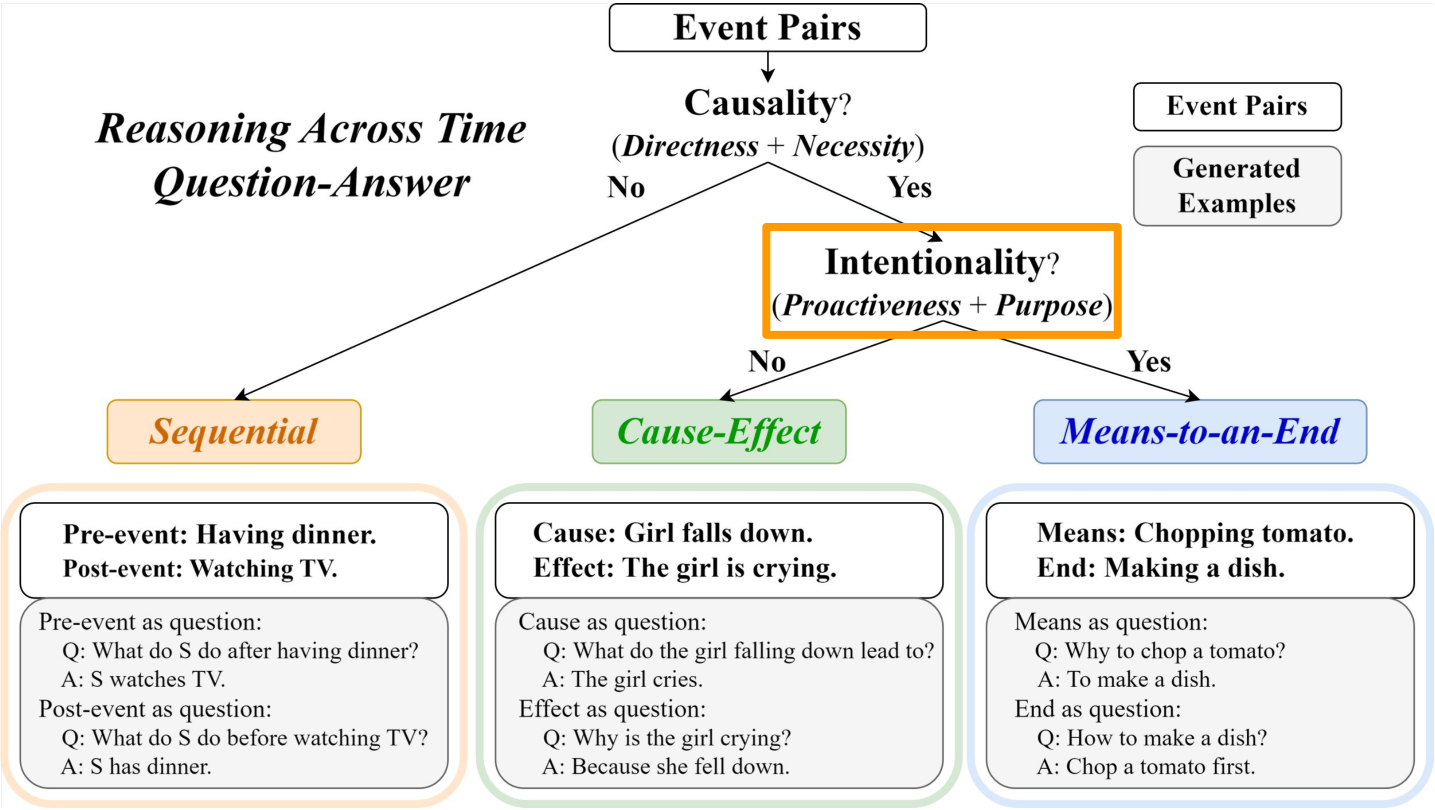
Grounding-VQA Classification Criteria

- ***Directness***: This criterion assesses the directness of the causal link between events.
- ***Necessity***: This criterion measures whether the second event is inevitable due to the first.
- ***Proactiveness***: This evaluates whether an event is carried out with deliberate intention.
- ***Purpose***: This evaluates whether the intention has been fulfilled.



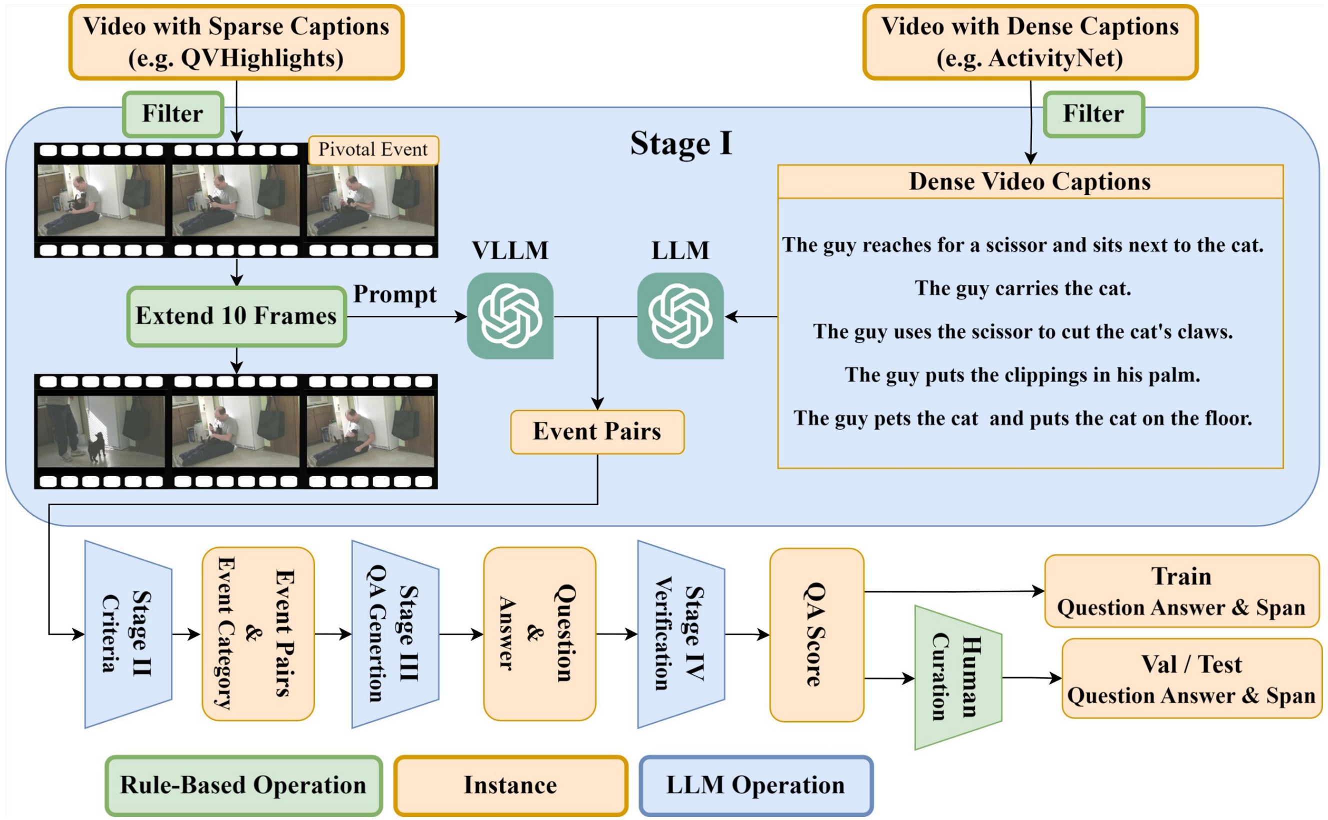
Grounding-VQA Classification Criteria

- *Directness*: This criterion assesses the directness of the causal link between events.
- *Necessity*: This criterion measures whether the second event is inevitable due to the first.
- ***Proactiveness***: This evaluates whether an event is carried out with deliberate intention.
- ***Purpose***: This evaluates whether the intention has been fulfilled.

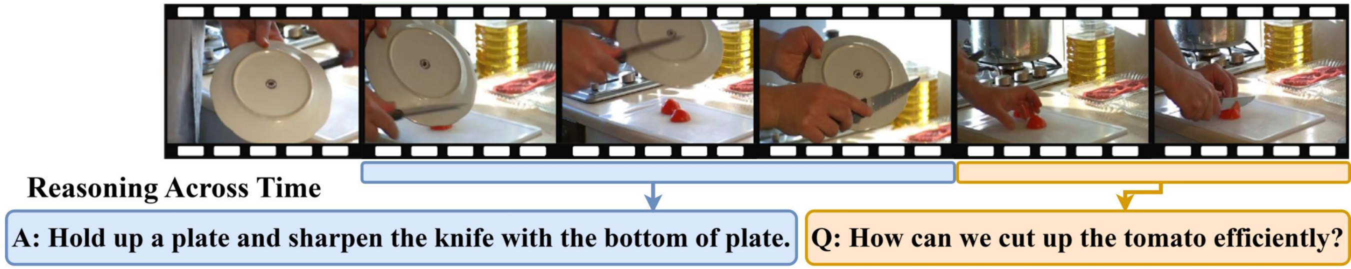


Performances on ReXTime

- Dataset sources:
  - ❑ ActivityNet [1], QVHighlights [2]
- Machine generated / verified
- Human verified validation / test set
- Reduce about 55% of overall cost



ReXTime Evaluation



- QA-IoU
  - ❑ Question-Answer Intersection over Union
- Lower QA-IoU indicates:
  - ❑ Less overlapping between the question span and the answer span.
  - ❑ More challenging for temporal reasoning.

Datasets	# of Reasoning Across Time Samples			C.L. (s) ↑	QA-mIoU (%) ↓
	Train	Val	Test		
Ego4D-NLQ	2,212†	775†	705†	5.2	85.5
NExTGQA	–	1,403†	2,301†	11.7	66.1
ReXTime (Ours)	9,695	921	2,143	**66.0**	**15.5**

Table: Frontier Models' Performances

Results of Frontier Models on ReXTime

- Moment localization
  - ❑ mIoU, R@1 (IoU=0.3), R@1 (IoU=0.5)
- VQA / Grounding VQA
  - ❑ Accuracy
  - ❑ Acc@IoU>0.5
- Human evaluation
  - ❑ 3 testers per question

Models	Moment Localization			VQA	
	mIoU	R@1 (IoU= 0.3)	R@1 (IoU= 0.5)	Accuracy(%)	Accuracy(%) @IoU ≥ 0.5
Human	**61.11**	**74.30**	**62.85**	**87.98**	**58.51**
GPT-4o	**36.28**	**45.33**	**34.00**	**73.67**	**28.67**
Claude3-Opus	23.61	30.67	17.67	68.67	13.67
Gemini-1.5-Pro	28.43	35.67	25.00	68.00	18.33
GPT-4V	26.74	33.33	22.00	63.33	16.67
Reka-Core	27.95	36.33	24.00	59.67	17.00

Table: Frontier MLLMs' Performances on ReXTime

Results of the Fine-tuned Performance

- Fine-tuned on ReXTime generated training data
- Performance boost after fine-tuned with our generated training data

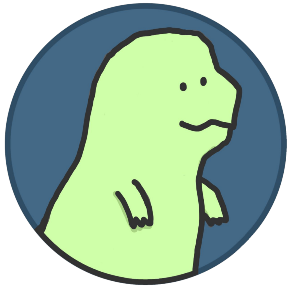
Models	Moment Localization			VQA	
	mIoU	R@1 (IoU=0.3)	R@1 (IoU=0.5)	Accuracy(%)	Accuracy(%) @ IoU $\geq$ 0.5
UniVTG (Zero-shot)	28.17	41.34	26.88	–	–
UniVTG (Finetuned)	34.63 (+6.46)	53.48 (+12.14)	34.53 (+7.65)	–	–
CG-DETR (Zero-shot)	23.87	31.31	16.67	–	–
CG-DETR (Finetuned)	26.53 (+2.66)	39.71 (+8.40)	22.73 (+6.06)	–	–
VTimeLLM (Zero-shot)	20.14	28.84	17.41	36.16	–
VTimeLLM (Finetuned)	29.92 (+9.78)	43.69 (+14.85)	26.13 (+8.72)	57.58 (+21.42)	17.13
TimeChat (Zero-shot)	11.65	14.42	7.61	40.04	–
TimeChat (Finetuned)	26.29 (+14.64)	40.13 (+25.71)	21.42 (+13.81)	49.46 (+9.42)	10.92

Conclusion

- Reasoning across time remains a challenge for current MLLMs.
- ReXTime is the first benchmark for reasoning-across-time with 2143 test samples
- ReXTime generated data is effective in enhancing reasoning across time.

- Thank you for your listening!

Project Page



Personal Page

References

- [1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In CVPR, 2015. 1, 4
- [2] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In NeurIPS, 2021. 1, 3, 4, 6
- [3] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In ICCV, 2023. 3, 7, 8
- [4] WonJun Moon, Sangeek Hyun, SuBeen Lee, and Jae-Pil Heo. Correlation-guided query-dependency calibration in video representation learning for temporal grounding. arXiv preprint arXiv:2311.08835, 2023. 3, 7, 8
- [5] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In CVPR, 2024. 3, 7, 8
- [6] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In CVPR, 2024. 3, 7, 8
- [7] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In CVPR, 2022. 3, 8, 9

References

- [8] Junbin Xiao, Angela Yao, Yicong Li, and Tat Seng Chua. Can i trust your answer? visually grounded video question answering. In CVPR, 2024. 3, 6, 8, 9
- [9] The claude 3 model family: Opus, sonnet, haiku. Technical report, Anthropic, 2024. 1, 6, 7
- [10] Gpt-4 system card. Technical report, OpenAI, 2024. 1, 6, 7
- [11] Reka core, flash, and edge: A series of powerful multimodal language models. Technical report, Reka, 2024. 6, 7
- [12] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023. 1, 3, 6, 7
- [13] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multi-modal models. arXiv preprint arXiv:2312.11805, 2023. 1, 3, 6, 7