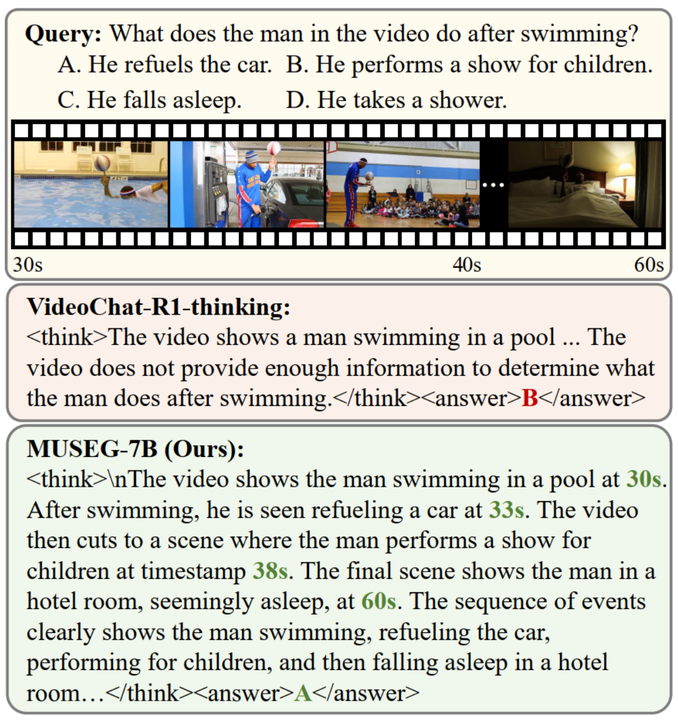
MUSEG: Reinforcing Video Temporal Understanding via Timestamp-Aware Multi-Segment Grounding

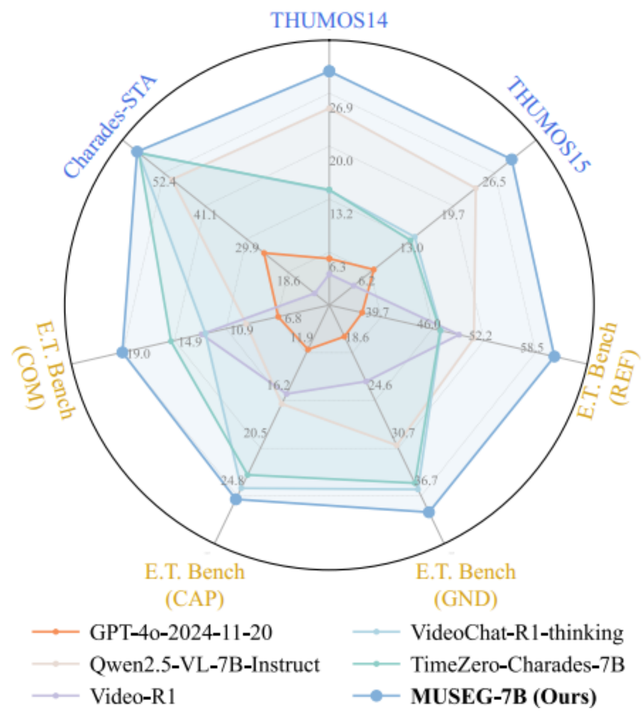
Fuwen Luo[1*], Shengfeng Lou[4*], Chi Chen[1*], Ziyue Wang[1*], Chenliang Li[3], Weizhou Shen[3], Jiyue Guo[1], Peng Li[2†], Ming Yan[3†], Ji Zhang[3], Fei Huang[3], Yang Liu[1,2†]

[1] Dept. of Comp. Sci. & Tech., Institute for AI, Tsinghua University
[2] Institute for AI Industry Research (AIR), Tsinghua University
[3] Institute of Intelligent Computing, Alibaba Group
[4] School of Computer Science and Technology (School of Artificial Intelligence), Zhejiang Sci-Tech University

❖ Introduction

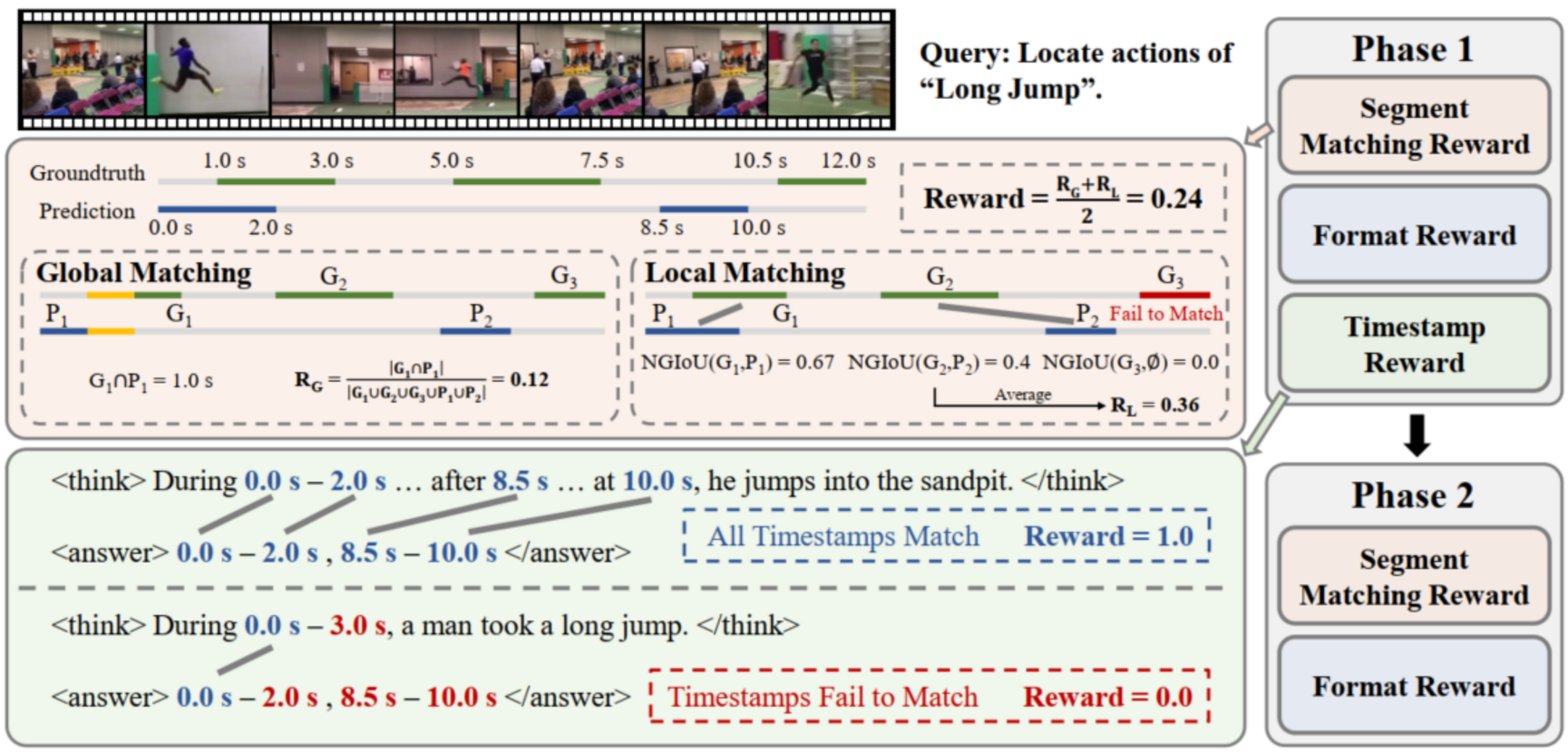
- Video temporal understanding requires reasoning over when events happen and how they are ordered.
- Current MLLMs often rely on global video descriptions and remain weak on fine-grained temporal cues.
- MUSEG improves temporal reasoning by aligning questions with multiple timestamped video segments.
- Key challenges: single-segment grounding can be answered by object shortcuts instead of temporal reasoning; answer-only rewards do not verify whether the model used temporal evidence in its reasoning.





❖ MUSEG

- The output contains both a reasoning process and a final answer in segment format.
- Segment Matching Reward evaluates predicted intervals against ground-truth intervals.
- Timestamp Reward checks whether answer timestamps are explicitly used in the reasoning process.
- Phased rewards use timestamp supervision early and remove it later.



❖ Results

Model	In-Domain				Out-of-Domain									
	Charades-STA (Single-Seg)	THUMOS14 (Multi-Seg)	THUMOS15 (Multi-Seg)	Perception Test (Multi-Seg)	E.T. Bench REF	E.T. Bench GND	E.T. Bench CAP	E.T. Bench COM	E.T. Bench AVG	E.T. Bench (Subset) REF	E.T. Bench (Subset) GND	E.T. Bench (Subset) CAP	E.T. Bench (Subset) COM	E.T. Bench (Subset) AVG
API-based Models														
GPT-4o	25.1	5.5	6.7	-	-	-	-	-	-	37.4	16.5	11.6	6.8	18.1
Open-source ∼ 7B Models														
Qwen2.5-VL-7B	50.2	24.9	23.4	25.3	53.1	30.7	16.2	11.3	27.8	51.0	30.3	16.5	9.3	26.8
+vanilla SFT	28.1	15.5	15.6	20.3	24.3	11.3	15.3	6.6	14.4	27.8	12.6	15.0	8.7	16.0
+vanilla GRPO	53.9	25.8	25.6	30.0	54.4	37.6	23.5	20.6	34.0	50.9	36.6	23.7	17.8	32.3
E.T. Chat	45.6	23.7	24.9	9.2	38.4*	**38.0***	16.7*	13.5*	26.7	31.8*	33.8*	17.1*	11.1*	23.5
TRACE-7B	29.9*	7.6	7.8	14.0	33.6*	33.8*	20.3*	**25.8***	28.4	25.2	17.2	14.7	5.3	15.6
Video-R1	11.3	3.5	3.4	5.7	50.3	25.3	15.6	12.4	25.9	49.2	22.2	15.6	12.8	25.0
VideoChat-R1	59.4	14.3	13.4	27.1	55.8	35.6	22.1	19.5	33.3	47.0	35.9	24.1	12.5	29.9
TimeZero	59.2	14.4	12.7	26.8	55.9	35.8	21.4	17.1	32.6	46.9	35.1	22.9	15.2	30.0
MUSEG-7B (Ours)	**59.7**	**29.7**	**29.3**	**31.7**	**61.9**	37.5	**23.7**	24.0	**36.8**	**60.8**	**38.8**	**25.1**	**19.0**	**35.9**
Open-source ∼ 3B Models														
Qwen2.5-VL-3B	41.4	12.6	12.8	19.4	51.7	20.4	13.6	8.0	23.4	52.9	20.4	12.7	7.6	23.4
TEMPURA	44.5	8.7	12.1	20.7	46.3	26.1	14.4	**10.2**	24.3	**56.4**	22.8	13.3	3.5	24.0
MUSEG-3B (Ours)	**53.7**	**21.0**	**20.3**	**29.1**	**53.9**	**30.0**	**18.7**	8.8	**27.9**	54.3	**28.7**	**18.3**	**11.8**	**28.3**

Email: lfw23@mails.tsinghua.edu.cn

Page: https://github.com/THUNLP-MT/MUSEG/