NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Two causally related needles in a video haystack

NeuIPS 2025

Miaoyu Li, Qin Chao, Boyang Li

Nanyang Technological University

Two crucial abilities insufficiently evaluated by existing long-context video benchmarks:

1. The ability to extract information from two separate locations in a long video and understand them jointly. (“Needle in a haystack” benchmarks: mainly focusing on understanding one needle or understanding multiple needles independently.)

2. The ability to model the world in terms of cause and effect in human behaviors.

The first long video benchmark for jointly understanding causally related events in the videos.

Two causally related needles in a video haystack (NeuIPS 2025)

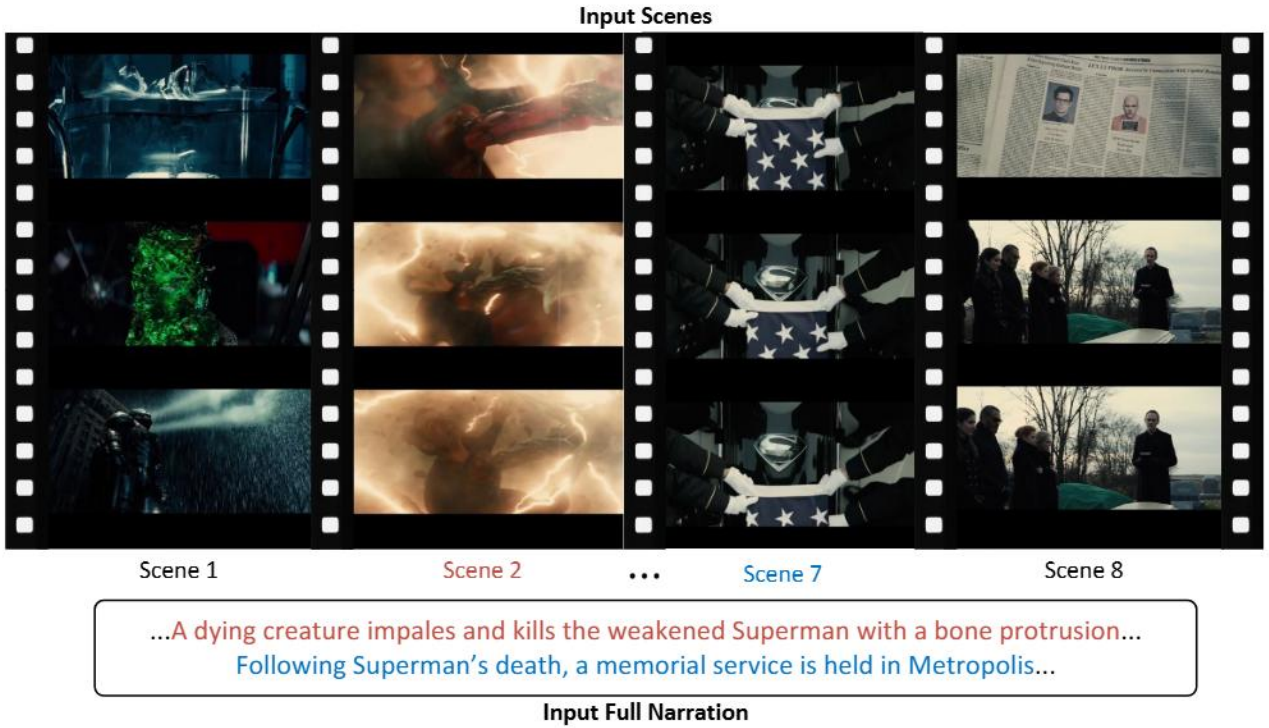
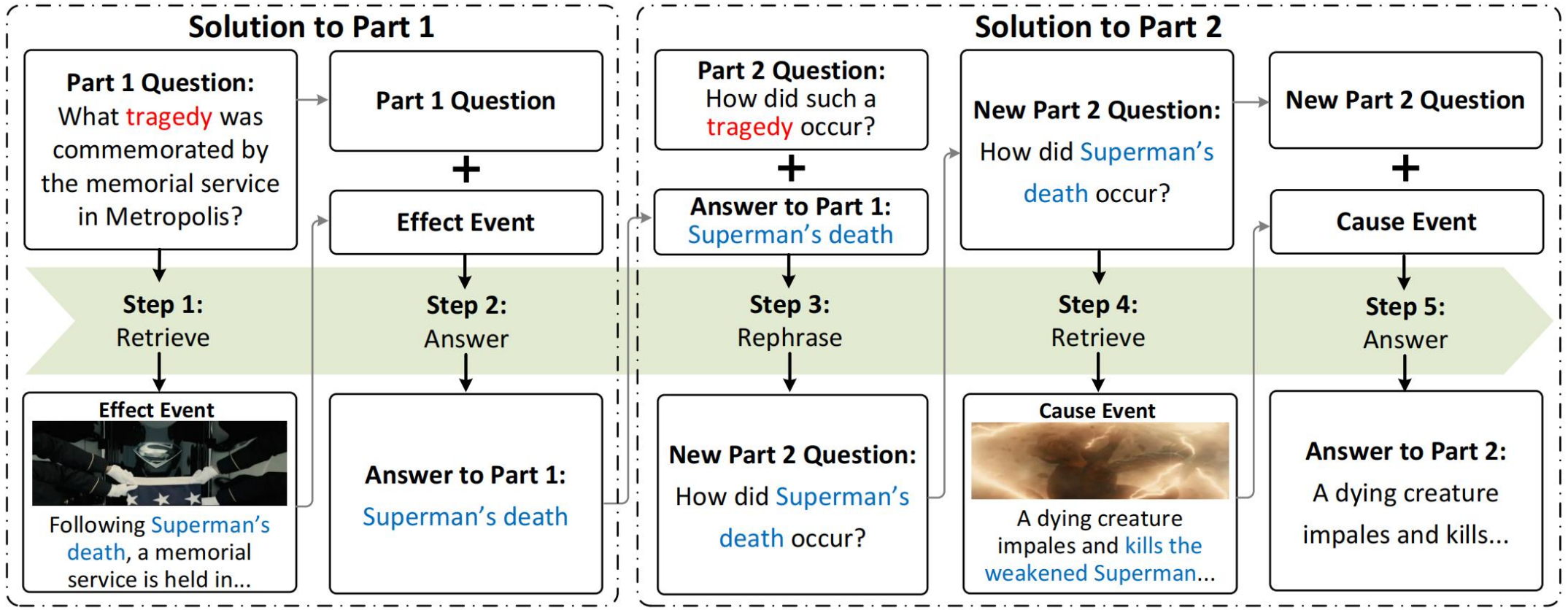
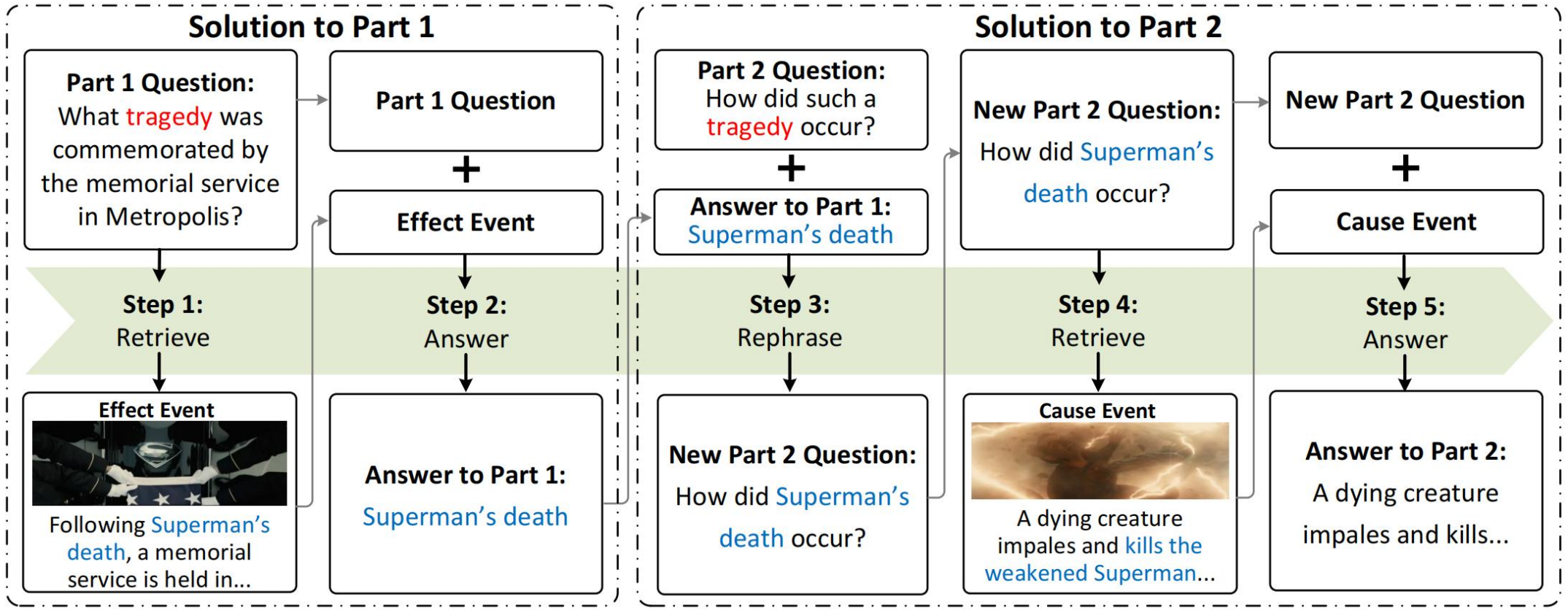


Figure 2: The evaluation framework of CAUSAL2NEEDLES. To help models understand the storyline, we also feed the full textual narration into the model. Four types of questions are designed for each pair of causally related events.

The logical process of answering our 2-needle questions



The design is based on “bridge entity”



Bridge Entity: a piece of shared information connecting the cause and effect events

Two question formats to mitigate textual bias from narration text

Visual grounding: requires the model to select the video clip corresponding to the event it needs to retrieve.

2-Needle Question (Visual Grounding)

Question: (Part 1) What tragedy was commemorated by the memorial service in Metropolis, **(Part 2)** and how did such a tragedy occur?
Select the scenes that contain the answers to the question.
Answer: Scene 7 for Part 1; Scene 2 for Part 2

Maybe OOD?
Underestimation?

← Complement →

Image description: requires the model to answer multiple-choice questions about the appearance of the retrieved video clip.

2-Needle Question (Image Description)

Question: (Part1) What tragedy was commemorated by the memorial service in Metropolis? **(Part2)** When the tragedy that causes the memorial service occurs, what color were the energy blasts emanating from the creature?
Choose the correct option.
A. Yellow B. Blue
C. Red D. White
Answer: A

Knowledge leakage?

Dataset statistics and question quality evaluation

Table 1: A comparison of CAUSAL2NEEDLES with other multi-needle long video benchmarks. CAUSAL2NEEDLES is the only benchmark dedicated to needle-in-haystack problems (Diagnostic Precision) that requires joint understanding of the two needles and the identification of cause events from effect events.

Benchmark	Video Length	# QA	Diagnostic Precision	Needle Type	Joint Understanding	Causal Reasoning
EgoScheme [Mangalam et al., 2023]	180 s	5,000	✗	Natural	✗	✓
MVBench [Li et al., 2024c]	16 s-40 s	4,000	✓	Natural	✗	✓
TVBench [Cores et al., 2024]	16 s-40 s	2,525	✓	Natural	✗	✗
MLVU [Zhou et al., 2024]	180 s-3600 s	3,102	✓	Artificial	✗	✗
VideoMME [Fu et al., 2024]	1018s	2,700	✓	Natural	✗	✓
LVB Wang et al. [2024a]	473 s	6,679	✓	Natural	✗	✗
CAUSAL2NEEDLES (Ours)	438 s	4,200	✓	Natural	✓	✓

Table 2: Evaluation results of generated questions. 'VG' and 'ID' refer to visual grounding and image description, respectively, while '1-N' and '2-N' denote 1-needle and 2-needle questions. Numbers in parentheses indicate the performance of random baselines.

Models	Shared Existence of Bridge Entities	Correctness of Vague References	Factual Correctness of Questions				Readability of Questions			
			Noncausal 1-N	Causal 1-N	VG 2-N	ID 2-N	Noncausal 1-N	Causal 1-N	VG 2-N	ID 2-N
ChatGPT-4.1	95.6% (0.3%)	91.0% (3.6%)	4.71 (1.10)	4.62 (1.12)	4.99 (1.10)	4.74 (1.13)	4.91	4.85	4.83	4.67
Gemini-2.0-flash	95.0% (3.8%)	98.7% (2.5%)	4.75 (1.05)	4.66 (1.02)	4.96 (1.01)	4.83 (1.00)	4.75	4.18	4.69	4.25
Human	82.4%	98.5%	-	-	4.50	-	-	-	4.80	-

Leaderboard

Table 3: Quantitative results (accuracy, %) of VLMs on our benchmark. 'Forward' refers to inputting video scenes in chronological order, while 'Reverse' uses reverse order. 'Avg' denotes results averaged over both orders. Best scores are in **bold**.

Models	Noncausal 1-N Questions	Causal 1-N Questions	VG 2-N Questions									ID 2-N Questions
			Forward			Reverse			Avg			
			Part 1	Part 2	Both	Part 1	Part 2	Both	Part 1	Part 2	Both	
Human	–	78.2	83.7	85.9	79.3		-			-		88.2
Proprietary Models												
ChatGPT-4o	**56.8**	**39.2**	16.7	39.2	9.4	**45.4**	21.2	**13.4**	**31.1**	30.2	**11.4**	59.2
Gemini-1.5-pro	55.4	35.6	**21.0**	**40.0**	**10.2**	35.7	**21.4**	8.4	28.4	**30.7**	9.3	**60.9**
ChatGPT-4o-mini	39.9	33.4	17.4	22.9	5.0	32.4	11.9	5.2	24.9	17.4	5.1	52.3
Claude-3.5-sonnet	37.6	26.5	16.6	22.4	4.8	19.3	13.9	2.9	17.9	18.1	3.9	60.5
Open-source Models												
Qwen2.5VL-32B	**30.7**	11.7	26.3	17.7	5.4	10.3	20.4	1.9	18.3	19.0	3.6	**53.5**
Qwen2.5VL-7B	17.5	13.6	**27.6**	**17.7**	**5.0**	11.2	**18.9**	**1.9**	**19.4**	**18.3**	**3.4**	43.2
LLaVA-Next-Video-34B	12.4	12.3	0.8	17.4	0.0	11.8	0.9	0.0	6.3	9.2	0.0	48.6
LLaVA-OneVision-7B	12.3	**18.0**	4.6	14.7	0.0	17.0	5.6	0.1	10.8	10.2	0.1	28.3
InternVL2-8B	11.6	7.4	14.5	8.3	1.2	9.5	9.1	0.5	12.0	8.7	0.9	40.2
LLaVA-Next-Video-7B	11.7	17.2	0.0	4.4	0.0	15.9	0.0	0.0	8.0	2.2	0.0	27.3
LongVA-7B	9.2	14.7	2.8	5.0	0.0	10.3	0.7	0.0	6.6	2.8	0.0	49.7
Aria-28B	7.0	12.1	19.0	14.8	0.6	**18.7**	18.1	0.1	18.9	16.5	0.4	43.0
LongVU-7B	3.3	12.2	3.2	1.3	0.3	4.4	2.1	0.5	3.8	1.7	0.4	34.2
Random Chance	9.8	9.8	9.8	9.8	1.0	9.8	9.8	1.0	9.8	9.8	1.0	25.0

Main Findings

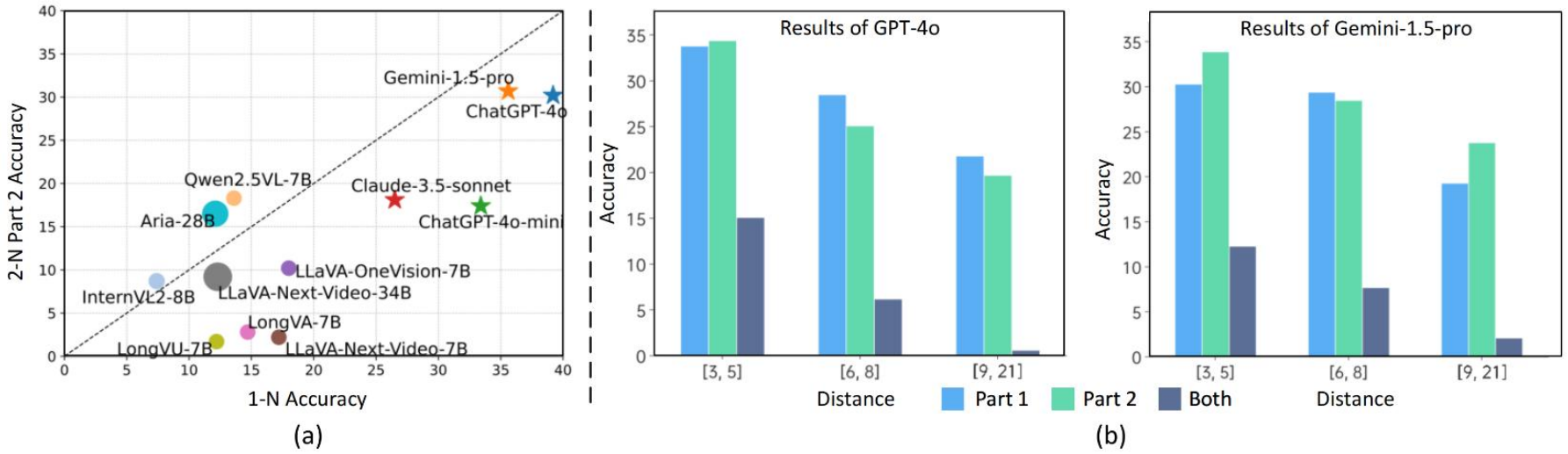


Figure 3: (a) The performance of models on VG 2-needle questions (Avg, Part 2) and causal 1-needle questions. The size of the dots indicates the model size. The stars indicate proprietary models. (b) The 2-needle visual grounding performance (Avg) on questions with different needle distances.

1. Causal questions are more challenging than noncausal.

2. 2-needle questions are more challenging than 1-needle.

3. The performance on 2-needle questions decreases as the needle distance increases.

Pathological Behaviors of VLMs

1. Positional Bias.

VG 2-N Questions								
Forward			Reverse			Avg		
Part 1	Part 2	Both	Part 1	Part 2	Both	Part 1	Part 2	Both
83.7	85.9	79.3		-			-	
16.7	39.2	9.4	**45.4**	21.2	**13.4**	**31.1**	30.2	**11.4**
21.0	**40.0**	**10.2**	35.7	**21.4**	8.4	28.4	**30.7**	9.3
17.4	22.9	5.0	32.4	11.9	5.2	24.9	17.4	5.1
16.6	22.4	4.8	19.3	13.9	2.9	17.9	18.1	3.9
26.3	17.7	5.4	10.3	20.4	1.9	18.3	19.0	3.6
27.6	**17.7**	**5.0**	11.2	**18.9**	**1.9**	**19.4**	**18.3**	**3.4**
0.8	17.4	0.0	11.8	0.9	0.0	6.3	9.2	0.0
4.6	14.7	0.0	17.0	5.6	0.1	10.8	10.2	0.1
14.5	8.3	1.2	9.5	9.1	0.5	12.0	8.7	0.9
0.0	4.4	0.0	15.9	0.0	0.0	8.0	2.2	0.0
2.8	5.0	0.0	10.3	0.7	0.0	6.6	2.8	0.0
19.0	14.8	0.6	**18.7**	18.1	0.1	18.9	16.5	0.4
3.2	1.3	0.3	4.4	2.1	0.5	3.8	1.7	0.4
9.8	9.8	1.0	9.8	9.8	1.0	9.8	9.8	1.0

Pathological Behaviors of VLMs

2. Static Output Bias.

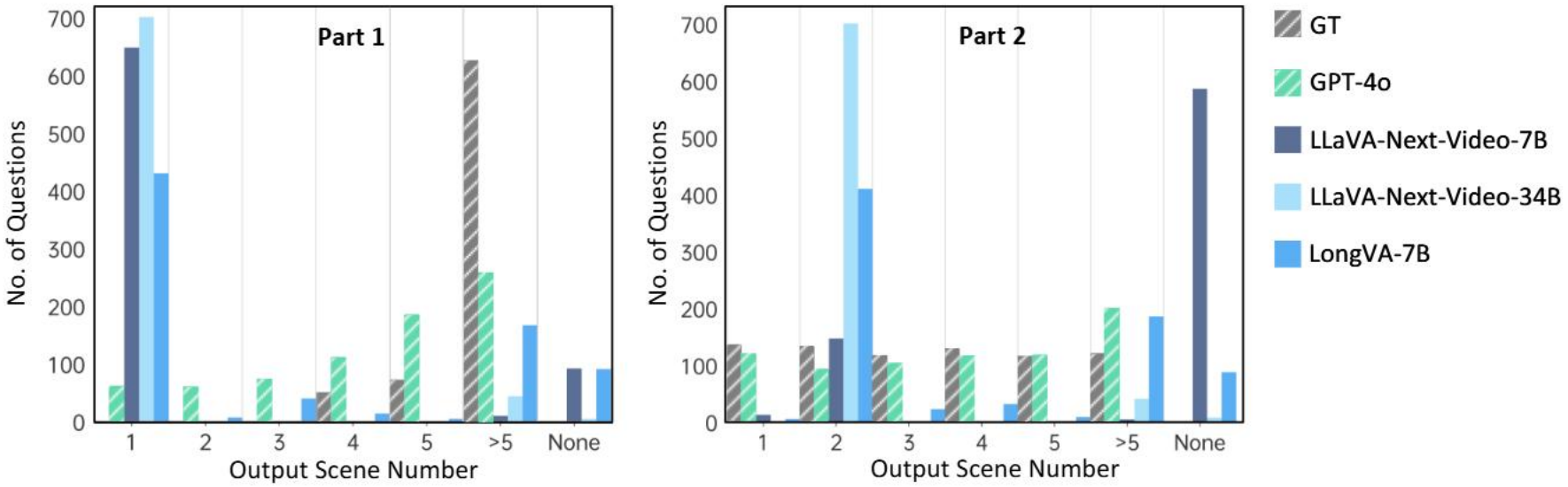


Figure 4: The answer distribution of various models in the forward evaluation of visual grounding 2-needle questions. "GT" denotes ground truth. "None" means no scene number is output.

Thank you for your attention!