IMPERIAL

ICML International Conference On Machine Learning

MatchLab

Hypo3D

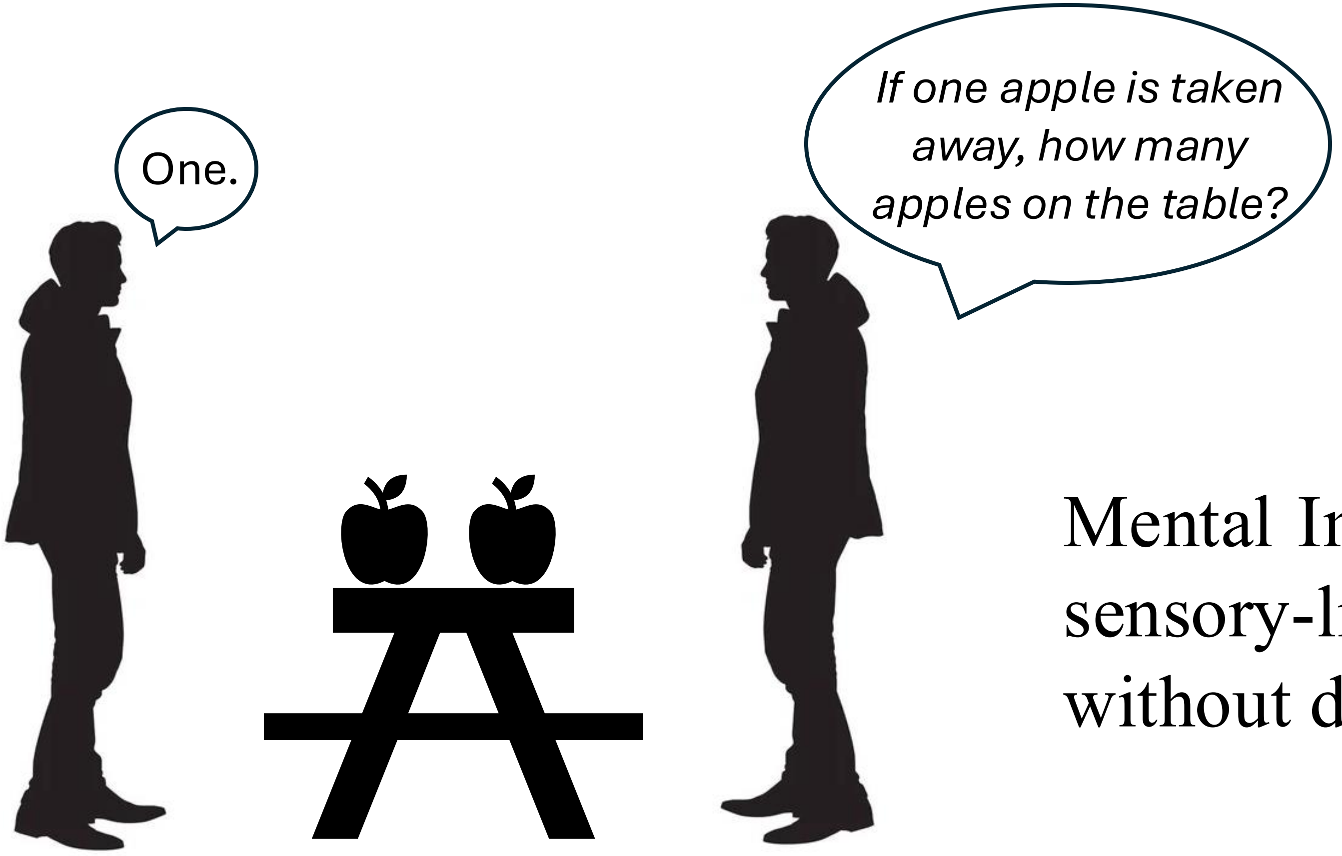
Exploring Hypothetical Reasoning in 3D

Ye Mao, Weixun Luo, Junpeng Jing, Anlan Qiu, Krystian Mikolajczyk

https://matchlab-imperial.github.io/Hypo3D/

As humans, we can reason through scenarios even with incomplete perceptual knowledge, using imagination to bridge knowledge gaps.



Mental Imagery: Creating or manipulating sensory-like representations in the mind without direct external stimuli.

IMPERIAL

ICML International Conference On Machine Learning

MatchLab

Case Study: In 3D reasoning, real-time access to an accurate scene is often limited by:

- Specialized equipment requirements.
- Prolonged scanning times.
- Complex reconstruction processes.

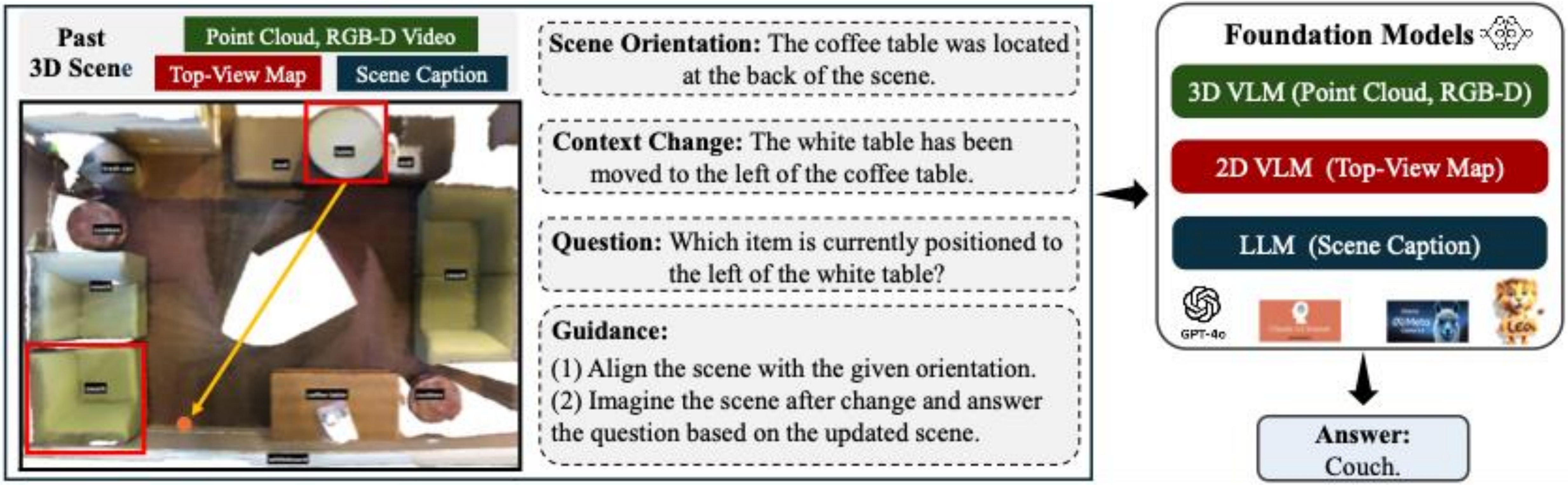
Hypothetical 3D Reasoning (Hypo3D): Use an easy-to-access context change description to *imagine* the updated scene and adjust reasoning results accordingly.

IMPERIAL

ICML
International Conference
On Machine Learning

MatchLab

Example of Hypothetical 3D Reasoning



IMPERIAL

ICML International Conference On Machine Learning

MatchLab

Context Change

1. Each object mentioned in change must have a uniquely specified location if it appears multiple times.
2. Each change should be spatially feasible within the scene layout.
3. Each change must be relied on the original scenes.

IMPERIAL

ICML International Conference On Machine Learning

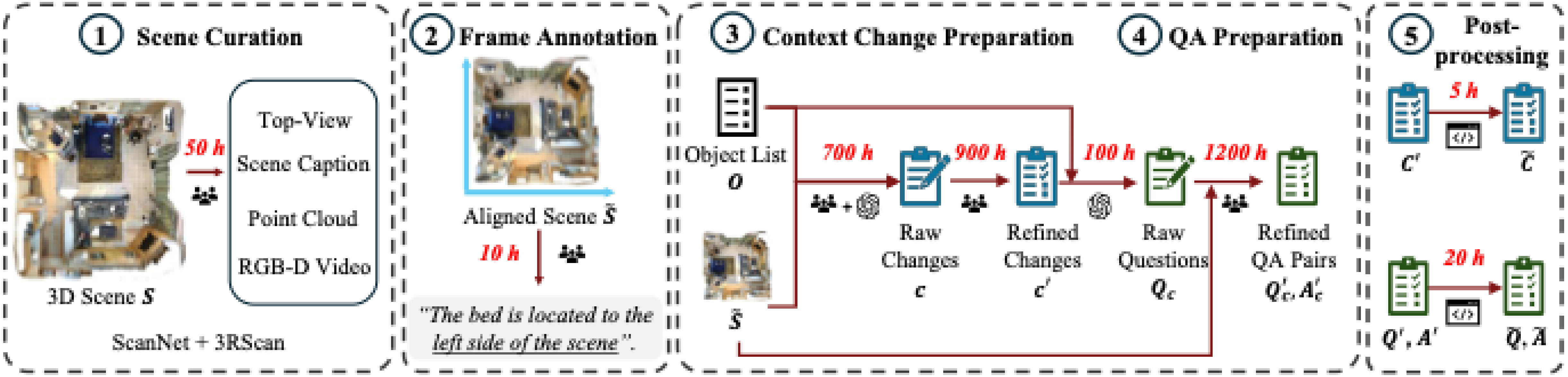
MatchLab

Questions & Answers

Scale	Is the laptop closer to the couch than the backpack now? **Proximity**	What is the largest item by the couch now? **Size-based Recognition**	Which object is encountered first along the direct path from the lamp to the bed? **Path Reasoning**
Direction	What is the position of the guitar relative to the TV after relocation? **Relative Position**	What object is now to the right of the laptop? **Direction-based Recognition**	From the chair, which direction should you take to reach the new location of the guitar? **Navigation**
Semantic	What furniture now can still hold items like the removed coffee table? **Functionality**	How many lamps are in the room now? **Counting**	Are there any curtains in the room that remain completely drawn? **Attribute**

1. Each question can only be answered using both the scene and context change, as neither is sufficient on its own.
2. Answers cannot be inferred from commonsense knowledge (e.g., bed is larger than pillow).
3. Each question has a unique and unambiguous answer.

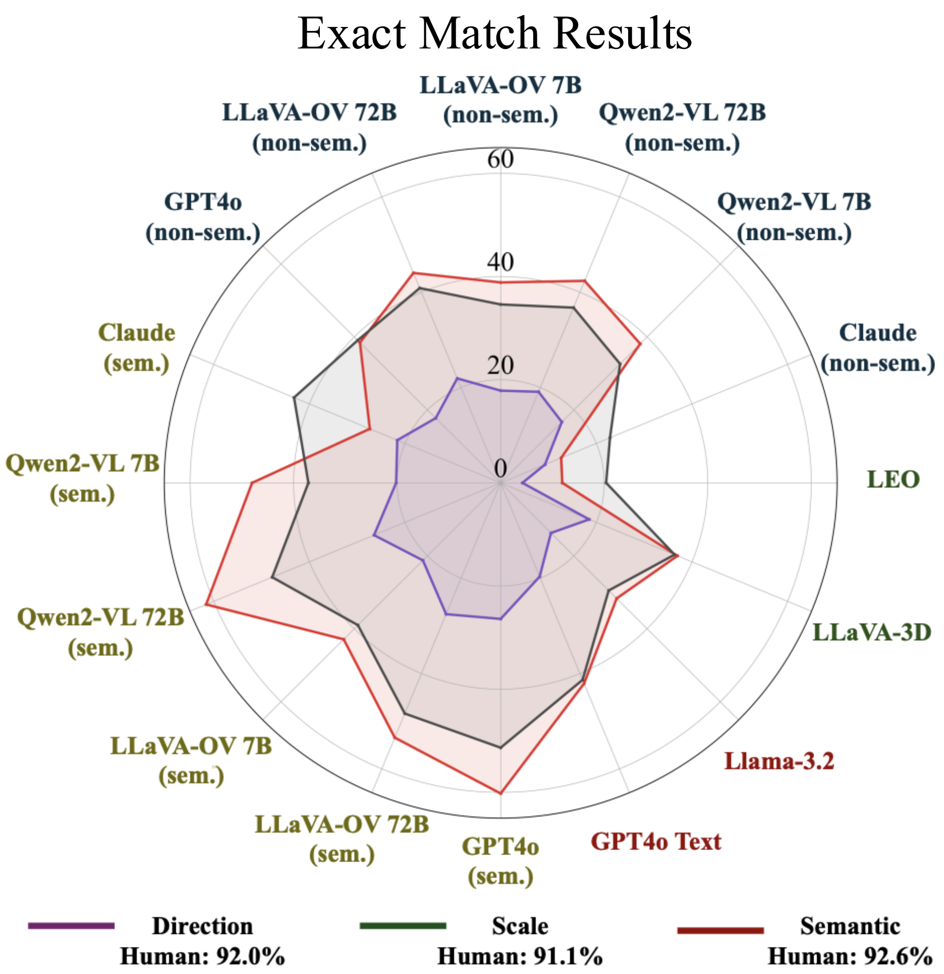
Data Generation Pipeline

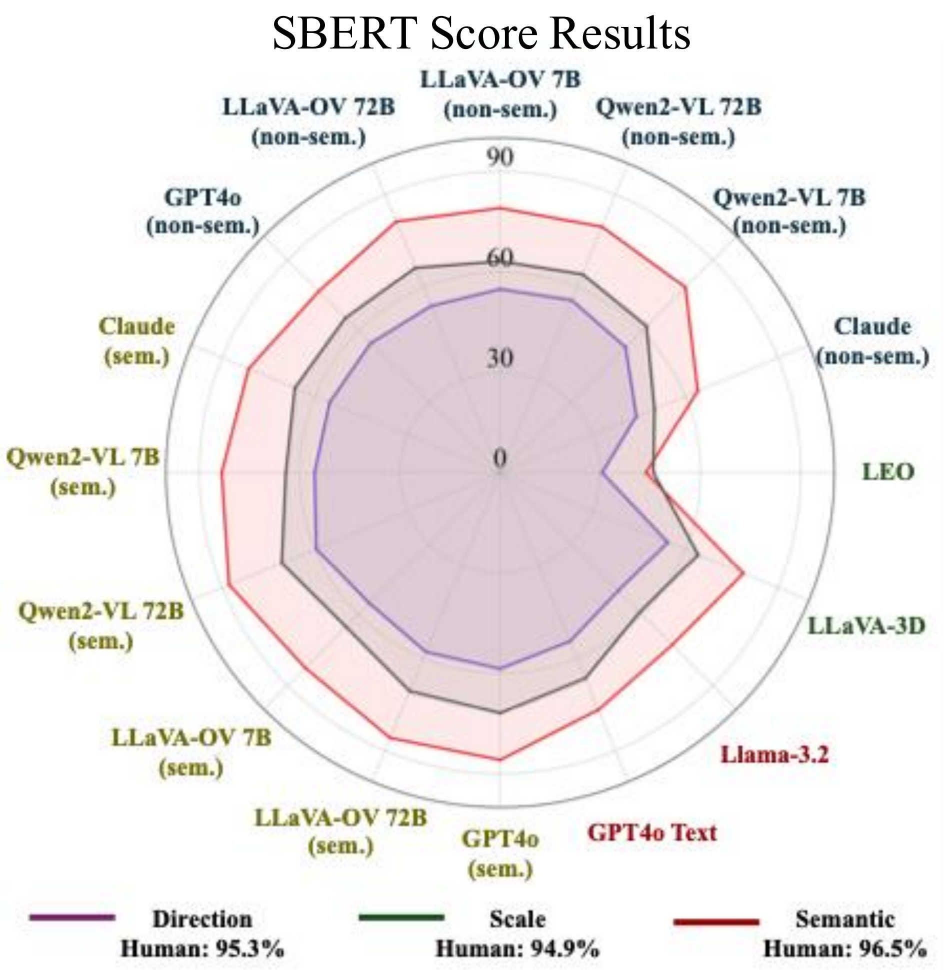


- Substantial performance gap between human and models.
- Models perform particularly poor on movement and replacement changes.
- Closed-source models does not always outperform open-source counterparts.

Model	Movement		Removal		Attribute		Addition		Replacement		Overall	
	EM	PM	EM	PM	EM	PM	EM	PM	EM	PM	EM	PM
LLM (Scene Caption)												
Llama-3.2 3B	25.31	28.37	29.85	33.65	24.95	29.59	26.78	30.78	23.75	27.68	26.08	29.91
GPT-4o API (Text)	35.76	38.66	36.88	41.71	34.05	39.58	39.74	43.28	31.33	35.24	35.54	39.65
2D VLM (Non-Semantic Top-View Map)												
Qwen2-VL 7B	29.23	35.08	30.71	34.69	29.04	33.94	31.48	35.17	28.41	33.10	29.68	34.47
Qwen2-VL 72B	33.02	37.38	33.88	37.57	33.48	37.62	35.95	40.29	30.66	34.64	33.39	37.51
LLaVA-OV 7B	30.34	34.17	29.81	33.24	31.37	36.13	33.12	35.64	28.41	31.81	30.62	34.34
LLaVA-OV 72B	36.46	39.83	36.45	40.22	35.70	40.46	39.64	42.25	33.83	37.85	36.38	40.13
Claude 3.5 Sonnet API	17.49	30.24	19.90	27.34	22.96	33.47	22.90	31.61	20.35	27.70	20.42	30.29
GPT-4o API	34.49	37.69	32.85	36.53	31.23	35.38	38.09	40.70	30.04	33.22	33.58	36.75
2D VLM (Semantic Top-View Map)												
Qwen2-VL 7B	31.26	36.41	38.09	41.90	34.83	39.41	37.64	41.41	31.86	36.62	34.40	38.91
Qwen2-VL 72B	38.42	42.56	**47.36**	**51.05**	46.76	51.10	47.63	50.87	**44.43**	48.78	44.25	48.25
LLaVA-OV 7B	33.32	36.80	34.34	37.84	34.98	39.50	38.96	41.98	33.93	38.33	34.81	38.60
LLaVA-OV 72B	39.39	42.99	43.44	46.87	44.57	49.37	46.12	49.06	44.10	48.18	43.01	46.83
Claude 3.5 Sonnet API	30.92	42.98	40.26	48.54	42.29	**52.72**	43.16	51.59	43.28	**50.73**	38.86	48.65
GPT-4o API	**40.77**	**43.79**	47.36	50.40	**47.42**	51.39	**50.59**	**53.77**	44.24	47.68	**45.50**	**48.82**
3D VLM (RGB-D Video, Point Cloud)												
LEO 7B	14.40	22.96	18.54	22.82	14.35	21.56	14.64	24.83	11.76	19.50	14.83	22.40
LLaVA-3D 7B	31.63	35.11	30.60	33.91	31.60	36.16	33.67	36.70	30.42	34.16	31.56	35.23
Human	95.00	96.00	93.00	95.00	93.00	94.83	89.00	90.67	85.00	86.00	91.00	92.50

Models struggle with direction-based questions.





IMPERIAL

ICML International Conference On Machine Learning

MatchLab

Reasoning in hypothetically changed scenes is more challenging than in unchanged scenes.

w/o change: Current Scene + Question → Answer
w. change: Past Scene + Context Change + Question → Answer

Table 3. Comparison of model performance when using and not using context change, where the changes **affect** the answer.

Model	**w/o change**		**w. change**	
	EM	PM	EM	PM
LLaMA-3.2 3B	19.00	23.25	20.50 (+1.50)	24.50 (+1.25)
Qwen2-VL 72B	37.00	41.50	31.50 (-5.50)	36.00 (-5.50)
GPT-4o API	38.00	40.25	33.00 (-5.00)	36.00 (-4.25)
Claude 3.5 Sonnet API	33.00	39.75	29.00 (-4.00)	35.50 (-4.25)
LLaVA-3D 7B	27.00	31.00	20.50 (-6.50)	24.00 (-7.00)

IMPERIAL

ICML International Conference On Machine Learning

MatchLab

Models hallucinate when changes are irrelevant.

Example：

Context Change: The cup is moved from table to the chair.

Question: What is the color of the cup?

Table 4. Comparison of model performance when using and not using context change, where the changes **do not affect** the answer.

Model	**w/o change**		**w. change**	
	EM	PM	EM	PM
LLaMA-3.2 3B	27.50	31.42	29.00 (+1.50)	33.25 (+1.83)
Qwen2-VL 72B	56.50	60.17	51.50 (-5.00)	55.17 (-5.00)
GPT-4o API	57.00	60.00	52.50 (-4.50)	56.92 (-3.08)
Claude 3.5 Sonnet API	52.50	59.00	49.00 (-3.50)	53.25 (-5.75)
LLaVA-3D 7B	37.50	40.17	37.00 (-0.50)	40.17 (0.00)