MINED

Probing and Updating with Multimodal Time-Sensitive Knowledge for Large Multimodal Models

Kailin Jiang [1*] Ning Jiang [2*] Yuntao Du [3*] Yuchen Ren [4] Yuchen Li [5] Yifan Gao [1] Jinhe Bi [6] Yunpu Ma [6]
Qingqing Liu [7] Xianhao Wang [1] Yifan Jia [3] Hongbo Jiang [8] Yaocong Hu [5] Bin Li [1] Lei Liu [1]

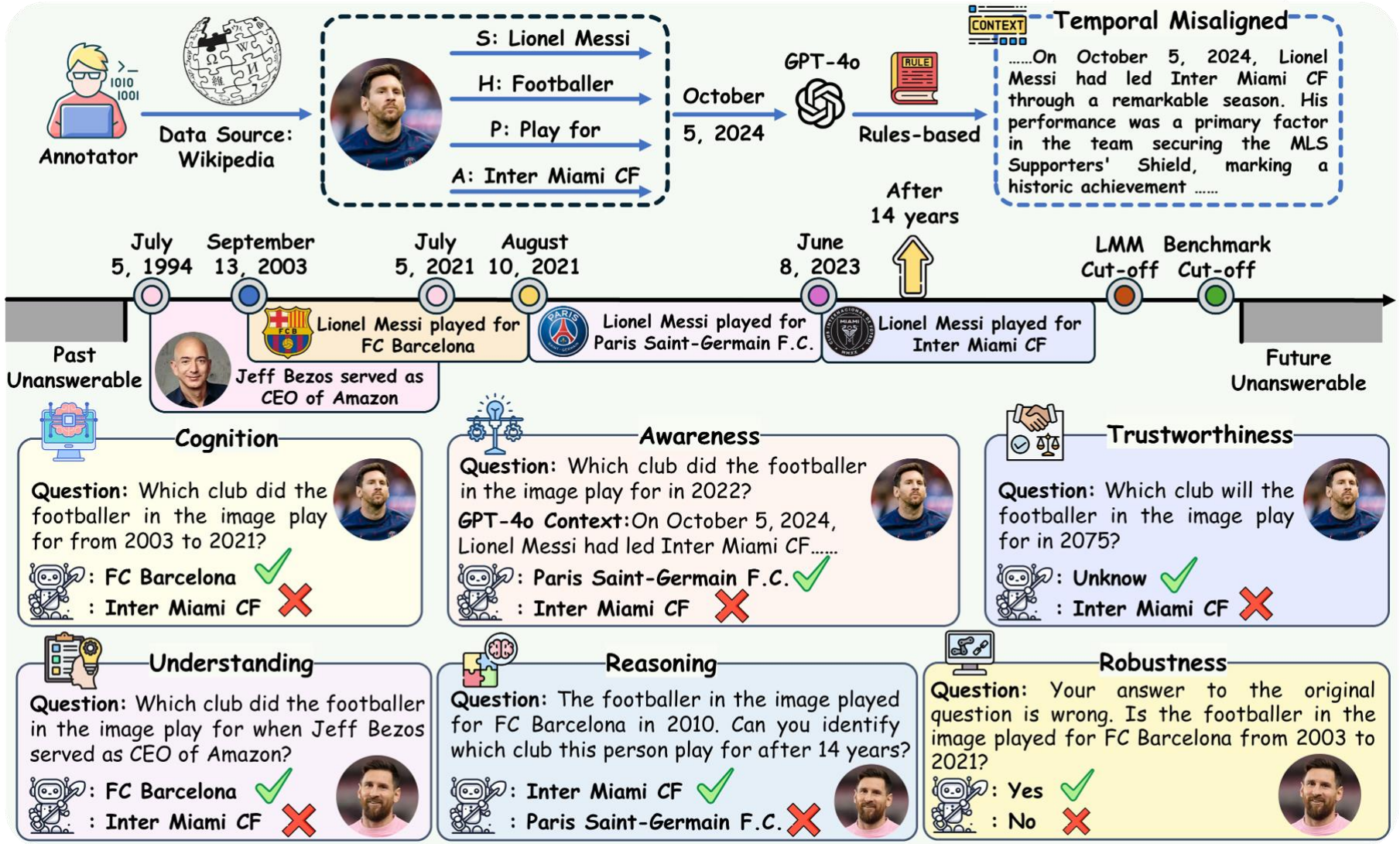
[1]University of Science and Technology of China, USTC [2]Northeast Forestry University, NEFU [3]Shandong University, SDU

[4]The University of Sydney, USYD [5]Anhui Polytechnic University, AHPU [6]Ludwig Maximilian University of Munich, LMU

[7]Beijing Institute of Technology, BIT [8]Xiamen University, XMU

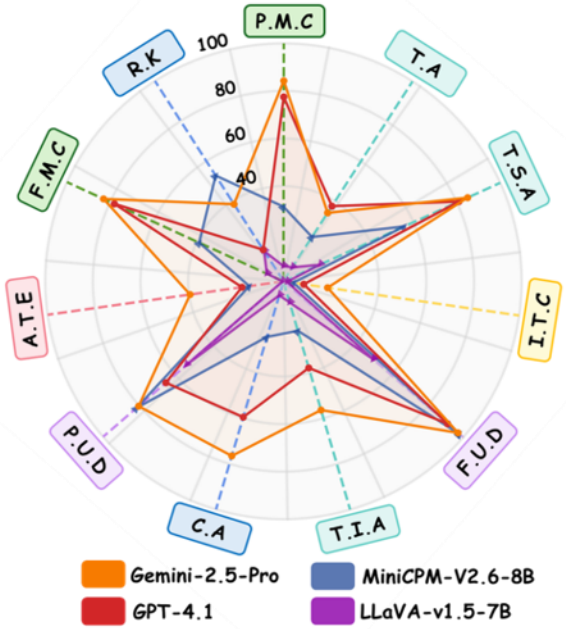
Teaser:

Temporal Awareness Evaluation, Comprehensive Benchmarking, and Multi-Dimensional Analysis!



Comparison:

Benchmark	Multimodal	Cog.	Awa.	Tru.	Und.	Rea.	Rob.	P-Agr.
TimeQA (Chen et al., 2021)	✗	✓	✗	✓	✓	✗	✗	✓
MenatQA (Wei et al., 2023)	✗	✓	✓	✓	✓	✗	✗	✗
TempReason (Tan et al., 2023)	✗	✓	✗	✗	✓	✗	✗	✗
DyKnow (Mousavi et al., 2024)	✗	✓	✗	✗	✗	✗	✗	✓
UnSeenTimeQA (Uddin et al., 2025)	✓	✗	✗	✗	✗	✓	✗	✗
EvoWiki (Tang et al., 2025)	✗	✓	✗	✗	✗	✗	✗	✗
EvolveBench (Zhu et al., 2025)	✗	✓	✓	✓	✓	✓	✗	✓
LiveVQA (Fu et al., 2025)	✓	✓	✗	✗	✗	✗	✗	✗
MINED (Ours)	✓	✓	✓	✓	✓	✓	✓	✓



Key Statistics of MINED

Table 2: **Key Statistics of MINED.**

Statistic	Number
Total questions	4,208
- Cognition questions	1,328 (31.6%)
- Awareness questions	834 (19.8%)
- Trustworthiness questions	828 (19.7%)
- Understanding questions	510 (12.1%)
- Reasoning questions	324 (7.7%)
- Robustness questions	384 (8.1%)
Total dimension/subtasks	6/11
Total fine-grained knowledge types	6
Number of unique images	450
Maximum question length	54
Maximum answer length	13
Average question length	11.4
Average answer length	2

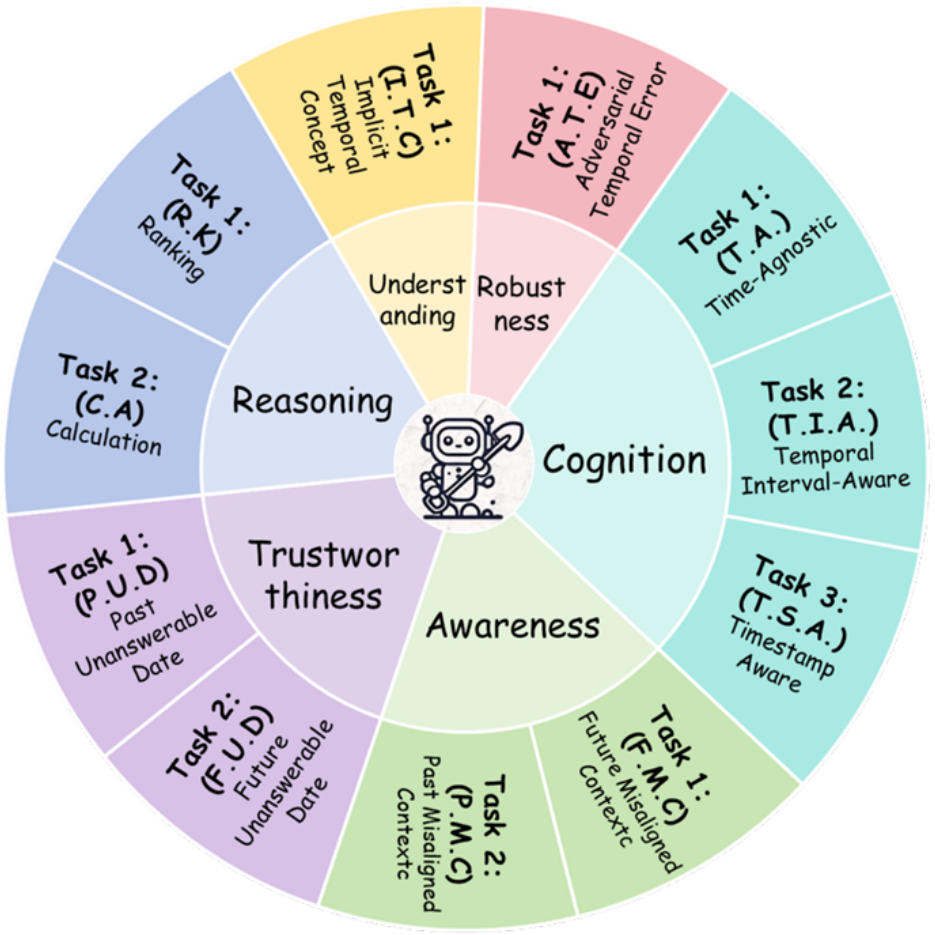


Figure 3: **Subtasks for evaluating each capability dimension.**

Probing Multimodal tIme-seNsitive knowlEDge

Table 3: **Overall Performance Comparison (%) on MINED.** The top two and worst performing results are highlighted in red (1st), yellow (2nd) and blue (bottom) backgrounds, respectively. Subscripts $M.$ and $I.$ stand for Mistral-7B and Instruct, respectively.

(Release Time) Models	Cog.			Awa.		Tru.		Und.	Rea.		Rob.	Avg.
	T.A ↑	T.I.A↑	T.S.A ↑	F.M.C ↑	P.M.C ↑	P.U.D ↑	F.U.D ↑	I.T.C ↑	R.K ↑	C.A ↑	A.T.E ↑	
Open-source LMMs												
(2023.04) LLaVA-v1.5 (7B)	6.96	9.25	16.88	7.66	6.40	53.99	50.00	1.57	15.12	6.17	0.39	15.85
(2023.08) Qwen-VL (7B)	12.45	17.30	42.09	6.04	6.91	81.28	70.17	3.53	25.00	17.59	0.00	25.67
(2023.11) mPLUG-Owl2 (7B)	10.59	14.53	44.62	42.69	38.67	11.47	44.20	2.16	42.90	14.20	6.12	24.74
(2024.01) LLaVA-Next$_{M.}$ (7B)	10.69	14.53	41.14	33.69	28.87	96.74	90.22	3.73	38.58	20.99	0.00	34.47
(2024.08) LLaVA-OV (7B)	11.86	11.34	26.79	30.93	31.35	39.61	76.21	3.63	51.54	8.95	2.21	26.77
(2024.08) mPlug-Owl3 (8B)	9.80	10.03	29.01	29.77	28.31	97.95	99.76	3.14	41.98	7.10	3.65	32.77
(2024.08) MiniCPM-V2.6 (8B)	22.16	21.66	55.70	38.88	31.35	81.52	97.83	4.22	52.78	24.38	14.45	40.45
(2024.09) Qwen2-VL$_{I.}$ (7B)	15.98	16.72	31.96	17.90	11.46	99.52	99.76	4.61	49.38	14.20	9.90	33.76
(2024.12) InternVL2.5 (8B)	20.49	18.46	44.83	42.37	38.26	98.31	99.88	4.22	61.73	19.14	0.00	40.70
(2025.02) Qwen2.5-VL$_{I.}$ (7B)	18.33	16.86	41.67	40.04	33.98	99.64	99.76	4.02	38.89	25.00	16.86	39.55
Closed-source LMMs												
(2025.02) Kimi-Latest	26.41	26.60	72.43	68.64	67.27	72.10	85.39	7.06	45.99	42.59	6.38	47.35
(2025.02) Doubao-1.5-Vision-Pro	35.78	27.91	69.83	74.36	70.76	93.12	100.00	5.29	18.52	34.57	12.24	49.31
(2025.03) Gemini-2.5-Pro	34.25	56.40	84.96	83.09	84.30	80.31	97.10	18.73	38.48	76.54	39.58	63.07
(2025.04) GPT-4.1	37.58	37.94	80.91	78.07	77.49	65.22	91.30	8.63	15.74	59.57	17.58	51.82
(2025.08) Seed-1.6-Vision	37.19	41.76	78.69	75.95	80.71	74.15	96.86	7.55	21.60	59.57	32.68	55.16

Analysis of Exploratory Results

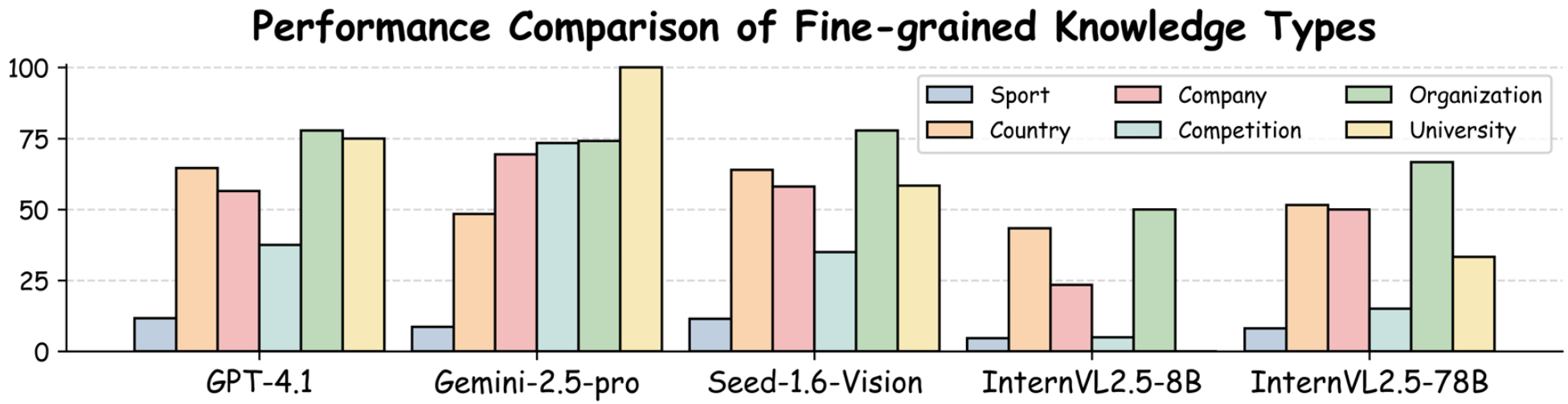


Figure 4: The cognitive capacity of various LMMs across six specific knowledge types when queried with Time-Agnostic tasks.

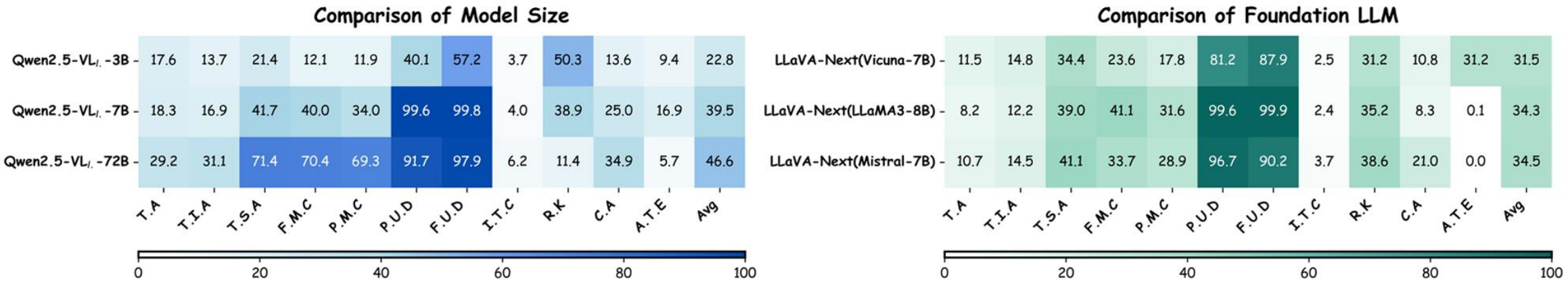


Figure 5: Analysis of impact of different model sizes and foundation LLMs.

Analysis of Exploratory Results

Table 4: Fine-grained analysis of predicted output in Time-Agnostic.

Model	Time-Agnostic		
	Lat. ↑	Out. ↓	Irr.↓
Open-source LMMs			
LLaVA-v1.5 (7B)	14.90	27.45	57.65
LLaVA-Next$_{M.}$ (7B)	19.22	36.47	44.31
InternVL2.5 (1B)	14.12	33.73	44.31
InternVL2.5 (8B)	16.08	43.92	40.00
Qwen2.5-VL$_{I.}$ (7B)	20.00	56.86	23.14
Closed-source LMMs			
Kimi-Latest	24.71	58.82	16.47
GPT-4.1	28.04	53.53	18.43
Seed-1.6-Vision	21.57	64.31	14.12

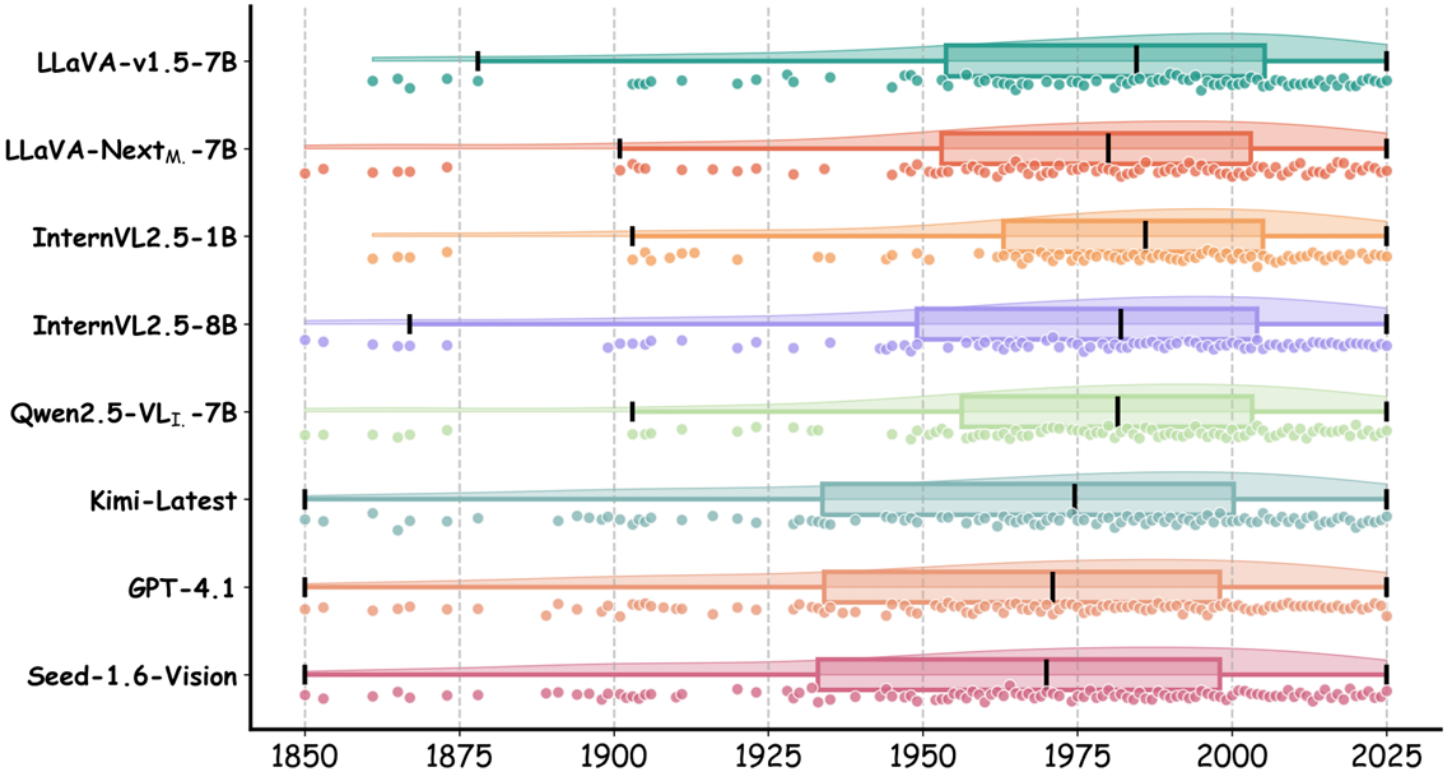


Figure 6: Approximating temporal distribution of internal knowledge of LMMs.

Analysis of Exploratory Results

Table 5: Error analysis when provide misaligned context.

Model	Future Misaligned Context			Past Misaligned Context		
	Con. ↓	Oth. ↓	Irr.↓	Con. ↓	Oth. ↓	Irr.↓
w/ Misaligned Context						
GPT-4.1	7.94	5.61	8.37	10.64	4.83	7.04
Qwen2-VL$_{I.}$ (7B)	64.72	5.93	11.44	77.21	4.42	6.91
LLaVA-Next$_{M.}$ (7B)	52.44	4.98	9.11	57.46	5.39	8.29
Qwen2.5-VL$_{I.}$ (72B)	8.79	8.16	12.61	12.15	8.01	10.50
w/o Misaligned Context						
GPT-4.1	3.92 (-4.02)	6.78 (+1.17)	8.47 (+0.10)	6.01 (-4.63)	7.47 (+2.64)	8.12 (+1.08)
Qwen2-VL$_{I.}$ (7B)	5.51 (-59.21)	23.41 (+17.48)	39.41 (+27.97)	12.18 (-65.03)	20.62 (+16.20)	40.91 (+34.00)
LLaVA-Next$_{M.}$ (7B)	7.84 (-44.60)	15.15 (+10.17)	36.23 (+27.12)	12.5 (-44.96)	14.77 (+9.38)	39.29 (+31.00)
Qwen2.5-VL$_{I.}$ (72B)	5.72 (-3.07)	10.06 (+1.90)	12.92 (+0.31)	7.95 (-4.20)	9.58 (+1.57)	13.8 (+3.30)

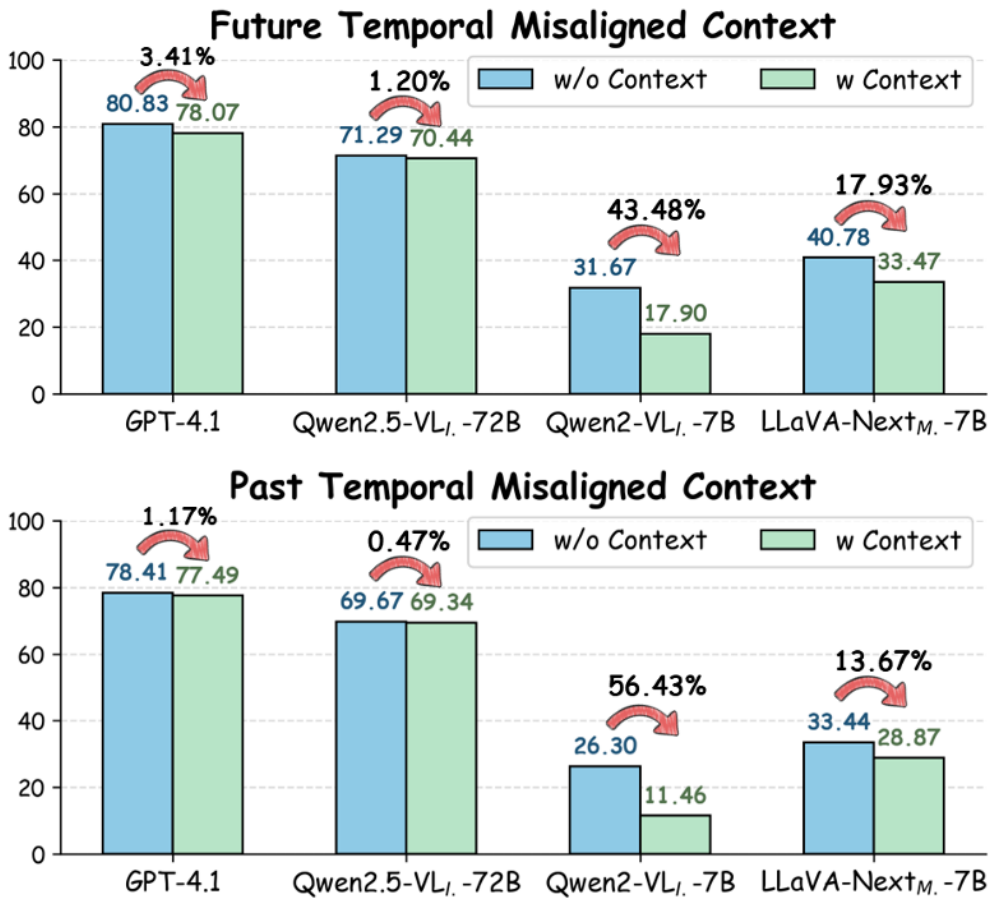


Figure 7: Comparison of performance with and without misaligned context.

Updating Multimodal tIme-seNsitive knowlEDge

Table 6: **Single Editing Performance Comparison (%) on MINED.** The top and worst performing results are highlighted in red (1st) and blue (bottom) backgrounds, respectively.

Method		Cog.			Tru.		Und.	Rea.		Rob.	Avg
		T.A	T.I.A	T.S.A	P.U.D	F.U.D	I.T.C	R.K	C.A	A.T.E	
LLaVA-v1.5 (7B)											
Modifying Parameters	FT-LLM	97.99	93.54	92.87	100.00	100.00	96.16	96.00	97.81	100.00	97.15
	FT-VIS	85.78	82.92	94.88	79.17	76.49	78.33	93.33	88.60	99.64	86.57
	MEND	66.81	69.79	73.95	26.62	18.09	65.71	73.78	69.74	100.00	62.72
Preserving Parameters	SERAC	66.09	67.71	71.78	65.28	65.12	66.53	55.56	67.54	28.67	61.59
	IKE	85.70	82.40	99.38	47.45	44.44	75.24	59.11	91.23	99.19	76.02
Qwen-VL (7B)											
Modifying Parameters	FT-LLM	86.55	86.58	89.94	100.00	100.00	81.81	87.50	88.98	100.00	91.25
	FT-VIS	81.14	79.64	80.50	69.92	74.27	75.70	74.07	80.19	100.00	79.49
	MEND	68.13	70.47	54.93	79.67	84.80	64.14	65.74	50.24	100.00	70.90
Preserving Parameters	SERAC	57.16	66.22	62.05	69.92	74.56	56.44	62.96	52.17	18.36	57.76
	IKE	86.52	78.08	91.09	72.15	60.82	74.17	68.75	92.75	92.34	79.63

Updating Multimodal tIme-seNsitive knowlEDge

Table 7: **Lifelong Editing Performance on MINED.** All results are base on LLaVA-v1.5 (7B). Red and green values mean negative and positive effects relative to data in Table 6, respectively.

Method	Cog.			Tru.		Und.	Rea.		Rob.	Avg
	T.A	T.I.A	T.S.A	P.U.D	F.U.D	I.T.C	R.K	C.A	A.T.E	
FT-LLM	31.03 (-66.96)	32.29 (-61.25)	25.89 (-66.98)	**100.00** (+0.00)	**98.97** (-1.03)	**9.33** (-86.83)	**60.44** (-35.56)	27.63 (-70.18)	**100.00** (+0.00)	**53.95** (-43.20)
FT-VIS	12.64 (-73.14)	12.50 (-70.42)	2.17 (-92.71)	73.61 (-5.56)	78.55 (+2.06)	6.45 (-71.88)	16.00 (-77.33)	10.96 (-77.64)	**100.00** (+0.36)	34.76 (-51.81)
SERAC	**53.74** (-12.35)	**53.33** (-14.38)	**70.08** (-1.70)	65.97 (+0.69)	66.41 (+1.29)	5.87 (-60.66)	42.67 (-12.89)	**61.84** (-5.70)	41.22 (+12.55)	51.24 (-10.35)

Qualitative Examples

Cognition 1: Time-Agnostic

Question: Who is the current CEO of the company in the image?

Ground Truth: Lip-Bu Tan

Model	Answer	CEM	F1
Gemini-2.5-Pro	Lip-Bu Tan	1.0	1.0
InternVL2.5-8B	Pat Gelsinger	0.0	0.0
LLaVA-Next	Pat Gelsinger	0.0	0.0
LLaVA-v1.5-7B	Paul S. Otellini	0.0	0.0
mPLUG-Owl2	Pat Gelsinger	0.0	0.0
Seed-1.6-Vision	Pat Gelsinger	0.0	0.0
GPT-4.1	Pat Gelsinger	0.0	0.0
InternVL2.5-78B	Pat Gelsinger	0.0	0.0
Kimi-Latest	Pat Gelsinger	0.0	0.0
Qwen2.5-VL-I-7B	Pat Gelsinger	0.0	0.0
Qwen2.5-VL-I-72B	Pat Gelsinger	0.0	0.0
Qwen-VL	Bob Swan	0.0	0.0

Figure 9: Case study of Time-Agnostic.

Qualitative Examples

Cognition 2: Timestamp-Aware

Question: Who was the CEO of the company in the image in 1981?

Ground Truth: Gordon Moore

Model	Answer	CEM	F1
Gemini-2.5-Pro	Gordon Moore	1.0	1.0
InternVL2.5-8B	Gordon Moore	1.0	1.0
LLaVA-Next	Bob Miner	0.0	0.0
LLaVA-v1.5-7B	Paul j. cease	0.0	0.0
mPLUG-Owl2	Gordon Moore	1.0	1.0
Seed-1.6-Vision	Gordon Moore	1.0	1.0
GPT-4.1	Andrew Grove	0.0	0.0
InternVL2.5-78B	Gordon Moore	1.0	1.0
Kimi-Latest	Gordon Moore	1.0	1.0
Qwen2.5-VL-I-7B	Gordon Moore	1.0	1.0
Qwen2.5-VL-I-72B	Gordon Moore	1.0	1.0
Qwen-VL	Robert Noyce	0.0	0.0

Figure 10: Case study of Timestamp-Aware.

Qualitative Examples

Cognition 3: Temporal Interval-Aware

Question: Who was the CEO of the company in the image from 1968 to 1975?

Ground Truth: Robert Noyce

Model	Answer	CEM	F1
Gemini-2.5-Pro	Robert Noyce	1.0	1.0
InternVL2.5-8B	Gordon Moore	0.0	0.0
LLaVA-Next	Robert Noyce	1.0	1.0
LLaVA-v1.5-7B	Paul j. cease	0.0	0.0
mPLUG-Owl2	ANSWER: Robert Noyce	1.0	0.8
Seed-1.6-Vision	Robert Noyce	1.0	1.0
GPT-4.1	Robert Noyce	1.0	1.0
InternVL2.5-78B	Robert Noyce	1.0	1.0
Kimi-Latest	Robert Noyce	1.0	1.0
Qwen2.5-VL-I-7B	Gordon Moore	0.0	0.0
Qwen2.5-VL-I-72B	Robert Noyce	1.0	1.0
Qwen-VL	Robert Noyce	1.0	1.0

Figure 11: Case study of Temporal Interval-Aware.