

648
649

APPENDIX CONTENTS

650
651

A THE USE OF LARGE LANGUAGE MODELS IN MINED 14

652
653

B MORE DETAILS ABOUT MINED 14

654

B.1 MINED 'S QUALITY AND EVOLVABILITY 14

655
656
657

B.2 MINED 'S DETAILED QUANTITY 14

658
659

C MORE EXPERIMENTAL RESULTS ABOUT MINED 15

660
661
662

C.1 MORE MAIN RESULTS ABOUT MINED 15

663
664

C.2 MORE MODEL SIZE RESULTS ABOUT MINED 16

665
666

D EXPERIMENT RESOURCES ABOUT MINED 17

667
668

E CASE STUDIES ABOUT MINED 17

669
670

F UPDATING TIME-SENSITIVE KNOWLEDGE VIA KNOWLEDGE EDITING 21

671
672

F.1 EDITING SETTING 21

673
674

F.2 KNOWLEDGE EDITING METHODS AND PARAMETERS 21

675
676

F.3 EDITING QUANTITY 22

677
678
679

G MORE DETAILS ABOUT CHAT TEMPLATES AND QUANTITATIVE EXAMPLES 23

680
681
682683
684
685686
687
688689
690691
692693
694695
696697
698699
700

701

702 A THE USE OF LARGE LANGUAGE MODELS IN MINED

704 In this section, we elaborate on the precise role of large language models within MINED, as detailed
 705 below.

706

- 707 • **Usage 1: MINED’s construction.** In the dimension of Awareness of Temporal Misalignment
 708 (Section 3.1), GPT-4o is employed to generate contextual content related to temporal misalignment.
 709 This approach is consistent with current academic research norms.
- 710 • **Usage 2: MINED’s evaluation.** In Section 4.2, we evaluate performance on MINED using Kimi-
 711 Latest, Gemini-2.5-Pro, Doubao-1.5-Vision-Pro, Seed-1.6-Vision and GPT-4.1, following standard
 712 benchmarking practices.
- 713 • **Usage 3: Paper grammar polishing.** The paper is initially drafted by human authors and
 714 subsequently polished for grammar using a large language model. It is not generated entirely by
 715 AI. This practice aligns with current academic norms.

716 B MORE DETAILS ABOUT MINED

718 B.1 MINED ’S QUALITY AND EVOLVABILITY

720 Owing to the time-sensitive nature of MINED, we will perform quarterly updates to endow the
 721 benchmark with evolvability. Unlike conventional benchmarks that merely replace outdated data,
 722 MINED offers a fundamentally distinct form of evolution. It not only evaluates model performance
 723 on time-sensitive knowledge but also probes models’ internal knowledge boundaries (in Section 4.3).
 724 To this end, we design an efficient pipeline to update the attribute list of each knowledge entry every
 725 quarter. This pipeline enables continuous renewal of knowledge, persistent evaluation of model
 726 knowledge boundaries, and provides the community with a dynamic and evolving evaluation resource.
 727 We outline MINED’s update pipeline:

728

- 729 • (1) Leveraging existing MINED subject S data, we retrieve corresponding Wikipedia text data
 730 offline (e.g., searching “Lionel Messi”).
- 731 • (2) For club affiliation information, we extract information from Wikipedia’s career sections
 732 using GPT-4o with strict parsing rules(the career field contains Lionel Messi’s club affiliation
 733 information).
- 734 • (3) Newly extracted club data is compared against MINED’s current records, triggering updates
 735 when discrepancies occur. This efficient pipeline ensures automated, continuous MINED updates,
 736 providing the community with an evolving evaluation resource.

737 Combined with this automated update pipeline, our proposed MINED benchmark can not only
 738 evaluate current state-of-the-art LMMs, **but also be used to evaluate newly emerging and more
 739 powerful LMMs in the future.**

740 B.2 MINED ’S DETAILED QUANTITY

741 Table 8: The detailed quantity of time-sensitive knowledge for each task

Cog.			Awa.		Tru.		Und.	Rea.	Rob.	Sum	
T.A	T.I.A	T.S.A	EM.C	P.M.C	P.U.D	F.U.D	I.T.C	R.K	C.A	A.T.E	
255	172	237	236	181	207	207	255	81	81	192	2104

756 C MORE EXPERIMENTAL RESULTS ABOUT MINED

758 C.1 MORE MAIN RESULTS ABOUT MINED

760 In this section, we present the complete experimental results on MINED. To further validate the
 761 reliability of our conclusions, we also employed the F1-Score as an additional evaluation metric.

762 The F1-Score is a metric for assessing model performance by quantifying the word-level similarity
 763 between a model’s output and the ground truth answer. It is the harmonic mean of Precision and
 764 Recall (Chan et al., 2024).

766 To calculate it, we first represent both the ground truth and the prediction as sets of words. Let the
 767 ground truth be $\mathcal{W}(y_q) = \{y_1, \dots, y_m\}$ and the model’s prediction be $\mathcal{W}(\hat{Y}) = \{\hat{y}_1, \dots, \hat{y}_n\}$. The
 768 number of common words between these sets, known as the overlap $\mathcal{U}(\hat{Y}, y_q)$, is computed using an
 769 indicator function $\mathbf{1}[\cdot]$:

$$770 \quad \mathcal{U}(\hat{Y}, y_q) = \sum_{t \in \mathcal{W}(y_q)} \mathbf{1}[t \in \mathcal{W}(\hat{Y})] \quad (2)$$

772 Precision, $\mathcal{P}(\hat{Y}, Y)$, is the fraction of relevant words among the predicted words. It is formally
 773 defined as:

$$774 \quad \mathcal{P}(\hat{Y}, Y) = \frac{\mathcal{U}(\hat{Y}, y_q)}{|\mathcal{W}(\hat{Y})|} \quad (3)$$

776 Recall, $\mathcal{R}(\hat{Y}, Y)$, is the fraction of ground truth words that the model successfully identified. It is
 777 defined as:

$$778 \quad \mathcal{R}(\hat{Y}, Y) = \frac{\mathcal{U}(\hat{Y}, y_q)}{|\mathcal{W}(y_q)|} \quad (4)$$

781 **Table 9: Complete F1-Score Performance Comparison (%) on MINED.** The top two and worst
 782 results are highlighted in red (1st), yellow (2nd) and blue (bottom) backgrounds, respectively.
 783 Subscripts L , M , V and I stand for LLaMA3-8B, Mistral-7B, Vicuna-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	E.U.D	I.T.C	R.K	C.A	A.T.E		
<i>Open-source LMMs</i>													
<i>Model size under 10B</i>													
(2023.04) LLaVA-v1.5 (7B)	7.89	11.44	16.88	10.60	9.49	53.99	50.00	1.95	15.33	6.38	0.39	16.76	
(2023.08) Qwen-VL (7B)	14.56	20.30	47.09	7.66	8.81	80.00	69.40	4.94	23.13	18.96	0.00	26.80	
(2023.11) mPLUG-Owl2 (7B)	13.40	17.05	50.94	48.26	44.21	11.19	44.20	3.34	43.40	16.59	6.12	27.15	
(2024.01) LLaVA-Next _L (8B)	9.39	16.68	46.39	47.51	38.20	99.64	99.88	3.47	36.08	10.85	0.13	37.11	
(2024.01) LLaVA-Next _M (7B)	13.37	18.74	46.59	37.34	32.05	96.74	90.22	4.43	38.85	24.23	0.00	36.60	
(2024.01) LLaVA-Next _V (7B)	13.89	18.34	39.15	27.60	22.54	81.16	87.92	3.99	32.23	15.25	31.25	33.94	
(2024.08) LLaVA-OV (7B)	14.22	15.24	31.91	35.12	34.84	39.61	76.21	4.86	52.56	14.73	2.21	29.23	
(2024.08) mPlug-Owl3 (8B)	9.94	14.07	33.09	21.87	20.86	97.60	99.76	3.27	41.53	7.62	3.65	32.11	
(2024.08) MiniCPM-V2.6 (8B)	24.11	25.91	58.78	41.37	34.63	81.52	97.83	5.81	53.67	27.74	14.45	42.35	
(2024.09) Qwen2-VL _I (7B)	19.20	21.34	37.49	21.92	14.71	99.52	99.76	6.09	50.27	18.40	9.90	36.24	
(2024.12) InternVL2.5 (1B)	4.53	2.65	4.86	3.48	3.06	97.95	98.43	1.19	42.35	3.85	0.00	23.85	
(2024.12) InternVL2.5 (2B)	6.67	7.29	10.21	5.96	4.98	96.74	95.89	2.04	13.77	5.27	0.78	22.69	
(2024.12) InternVL2.5 (4B)	21.02	17.35	35.32	34.06	31.36	98.43	99.28	4.26	47.74	22.07	1.56	37.50	
(2024.12) InternVL2.5 (8B)	21.71	23.29	49.14	47.38	42.64	98.31	99.88	6.00	62.11	24.52	0.00	43.18	
(2025.02) Qwen2.5-VL _I (3B)	19.55	16.39	25.16	15.20	14.61	40.10	57.25	5.28	50.58	16.46	9.38	24.54	
(2025.02) Qwen2.5-VL _I (7B)	21.59	22.29	47.47	45.77	38.83	99.64	99.76	5.74	39.22	28.35	22.29	42.81	
<i>Model size under 65B</i>													
(2024.12) InternVL2.5 (26B)	23.85	26.20	62.74	54.07	52.18	97.22	99.52	6.52	27.71	25.33	8.33	43.97	
(2024.12) InternVL2.5 (38B)	29.71	32.50	73.72	68.91	62.41	92.63	99.15	5.48	32.83	32.82	11.33	49.23	
<i>Model size under 100B</i>													
(2024.12) InternVL2.5 (78B)	30.44	35.91	75.35	74.59	73.79	81.16	97.58	7.75	12.80	43.09	8.33	49.16	
(2025.02) Qwen2.5-VL _I (72B)	32.42	36.97	76.21	75.32	73.56	91.67	97.95	7.78	11.91	38.07	5.73	49.78	
<i>Closed-source LMMs</i>													
(2025.02) Kimi-Lates	28.55	31.63	76.34	73.19	71.16	72.10	85.27	8.45	46.48	47.12	6.38	49.70	
(2025.03) Doubao-1.5-Vision-Pro	36.87	34.33	76.52	78.39	74.61	93.12	100.00	6.21	19.71	38.63	12.24	51.88	
(2025.03) Gemini-2.5-Pro	35.21	58.86	87.06	86.37	86.67	75.50	93.77	17.39	39.72	81.21	31.94	63.07	
(2025.04) GPT-4.1	37.26	43.42	84.93	82.47	82.02	64.44	91.30	10.11	16.77	62.03	17.58	53.85	
(2025.08) Seed-1.6-Vision	38.50	48.55	82.83	79.85	83.59	74.15	96.86	9.22	22.00	62.55	31.05	57.20	

808 According to the results in Table 9, we found that the conclusion drawn when using F1-Score as the
 809 evaluation metric is consistent with the conclusion drawn when using CEM as the evaluation metric,
 highlighting the reliability of our results and observations.

Table 10: **Complete CEM Performance Comparison (%) on MINED.** The top two and worst results are highlighted in red (1st), yellow (2nd) and blue (bottom) backgrounds, respectively. Subscripts L , M , V and I stand for LLaMA3-8B, Mistral-7B, Vicuna-7B and Instruct, respectively.

(Release Time) Models	Cog.			Awa.		Tru.		Und.		Rea.		Rob.	Avg.
	TA	T.LA	T.SA	E.M.C	P.M.C	P.U.D	E.U.D	L.T.C	R.K	C.A	A.T.E		
Open-source LMMs													
<i>Model size under 10B</i>													
(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 _L (8B)	8.24	12.21	39.03	41.10	31.63	99.64	99.88	2.35	35.19	8.33	0.13	34.34	
(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	
(2024.01) LLaVA-Next _V (7B)	11.47	14.83	34.39	23.62	17.82	81.16	87.92	2.55	31.17	10.80	31.25	31.54	
(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 (1B)	6.96	3.49	7.28	3.92	3.31	97.95	98.43	2.35	45.06	3.40	0.00	24.74	
(2024.12) InternVL2.5 (2B)	5.59	5.52	9.07	4.03	3.18	96.74	95.89	0.88	13.27	4.32	0.78	21.75	
(2024.12) InternVL2.5 (4B)	18.63	13.66	32.91	31.36	28.31	98.43	99.28	3.04	47.53	20.06	1.56	35.89	
(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 (3B)	17.65	13.66	21.41	12.08	11.88	40.10	57.25	3.73	50.31	13.58	9.38	22.82	
(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	
<i>Model size under 65B</i>													
(2024.12) InternVL2.5 (26B)	21.96	21.37	59.39	49.79	49.72	97.22	99.52	5.00	26.85	20.99	8.33	41.83	
(2024.12) InternVL2.5 (38B)	28.43	27.47	70.15	65.78	59.81	92.63	99.15	4.31	31.79	28.70	11.33	47.23	
<i>Model size under 100B</i>													
(2024.12) InternVL2.5 (78B)	29.31	28.63	70.25	69.92	70.86	81.16	97.58	5.98	11.73	38.58	8.33	46.58	
(2025.02) Qwen2.5-vL _I (72B)	29.22	31.10	71.41	70.44	69.34	91.67	97.95	6.18	11.42	34.88	5.73	47.21	
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	

C.2 MORE MODEL SIZE RESULTS ABOUT MINED

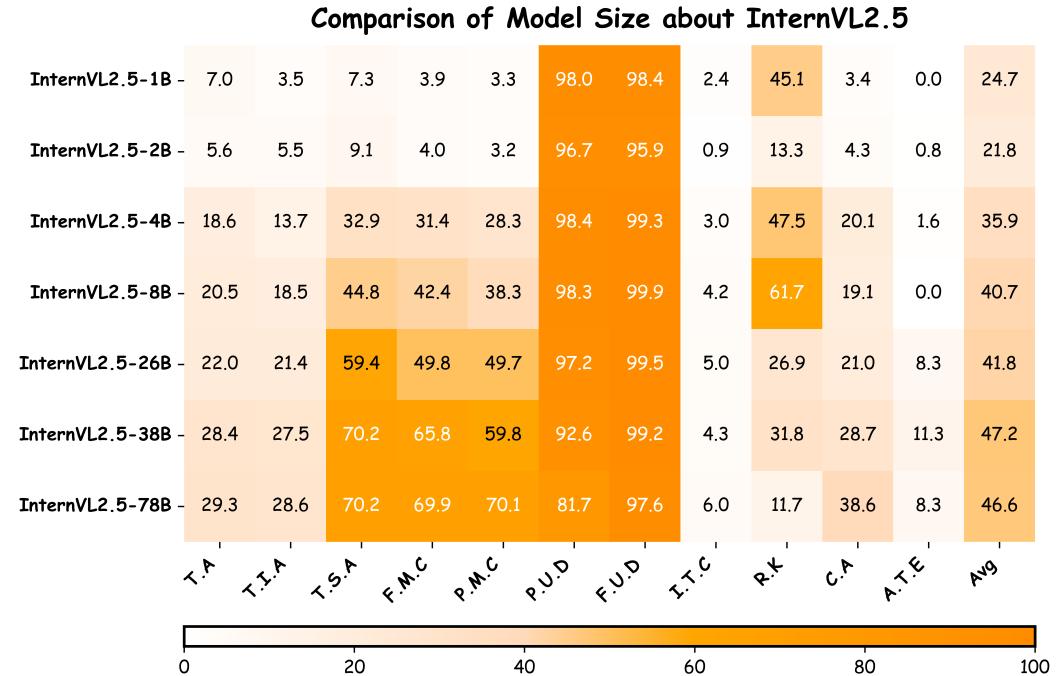
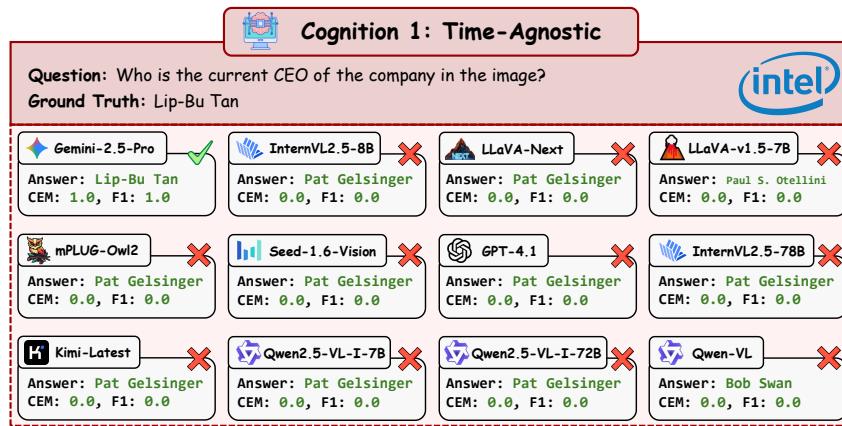
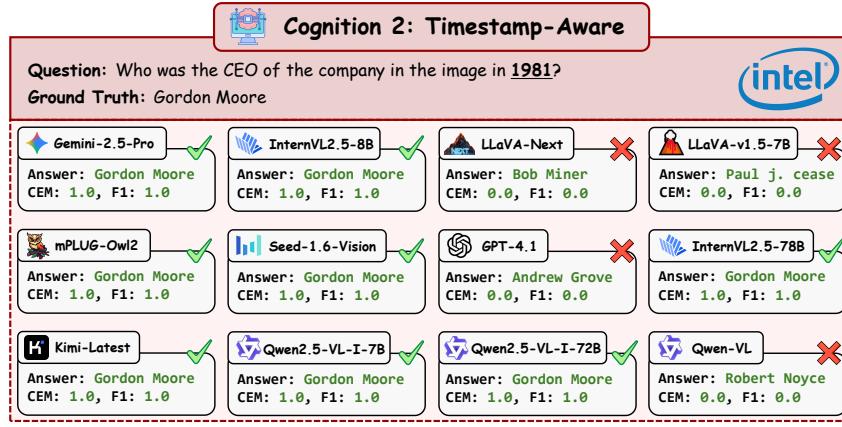


Figure 8: Analysis of impact of different model sizes about InternVL2.5 series.

864 **D EXPERIMENT RESOURCES ABOUT MINED**
865866 **PROBING TIME-SENSITIVE KNOWLEDGE**
867868 Regarding the validation experiments of LMMs on MINED, for models with parameter sizes of 38B
869 or less, we conduct experiments on 4 NVIDIA A100 PCIEs machines (40 GiB each); For models
870 with parameter sizes greater than 38B, we conduct experiments on 4 NVIDIA H100 (96 GiB each).
871872 **EDITING TIME-SENSITIVE KNOWLEDGE**
873874 We conduct knowledge editing experiment on one H100 (96 GiB each) regarding LMMs.
875876 **E CASE STUDIES ABOUT MINED**
877892 Figure 9: Case study of Time-Agnostic.
893908 Figure 10: Case study of Timestamp-Aware.
909

918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971

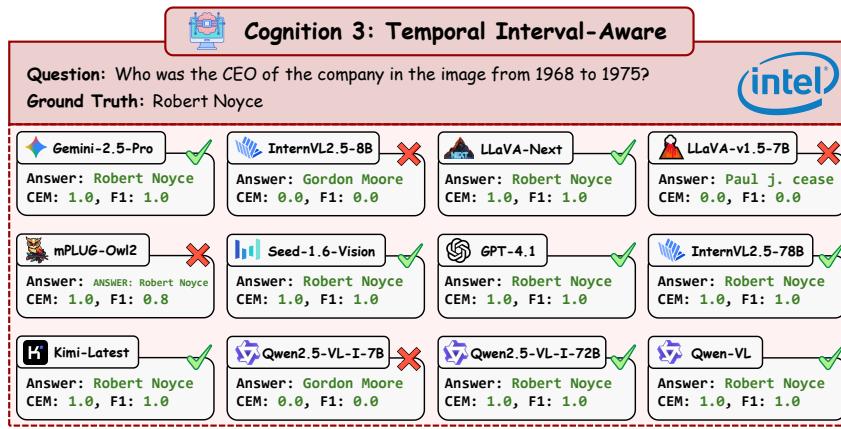


Figure 11: Case study of Temporal Interval-Aware.

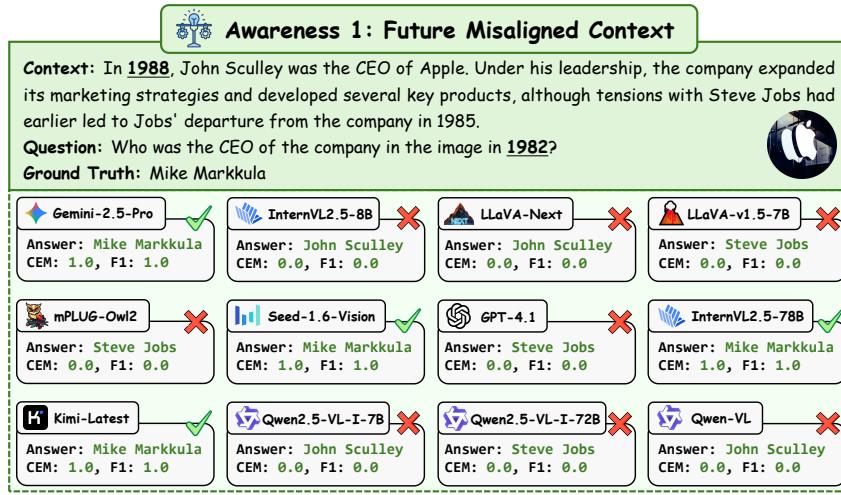


Figure 12: Case study of Future Misaligned Context.

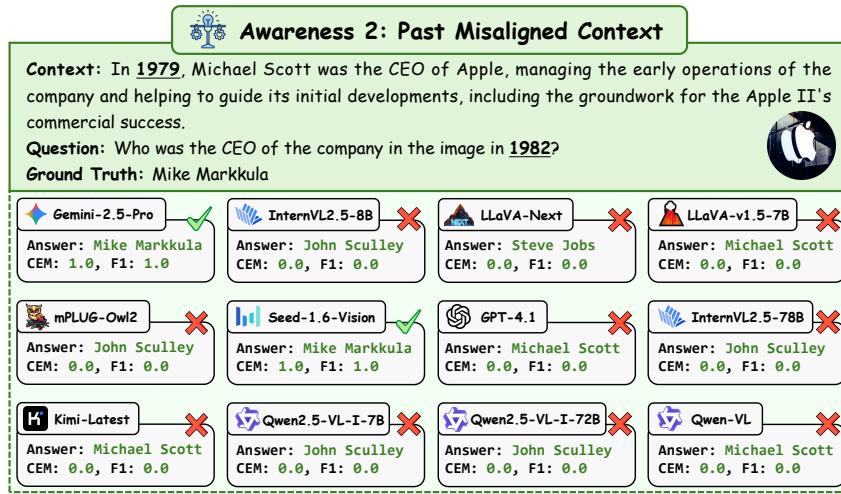


Figure 13: Case study of Past Misaligned Context.

972

973

974

975

976

977

978

979

980

981

982

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

1000

1001

1002

1003

1004

1005

1006

1007

1008

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018

1019

1020

1021

1022

1023

1024

1025

Trustworthiness 1: Past Unanswerable Date			
Question: Who was the CEO of the company in the image in 1889 ?		SONY	
Ground Truth: Unknown			
Gemini-2.5-Pro	✓	InternVL2.5-8B	✓
Answer: Unknown CEM: 1.0, F1: 1.0		Answer: Unknown CEM: 1.0, F1: 1.0	
mPLUG-Owl2	✗	Seed-1.6-Vision	✓
Answer: Masaru Ibuka CEM: 0.0, F1: 0.0		Answer: Unknown CEM: 1.0, F1: 1.0	
Kimi-Latest	✓	GPT-4.1	✓
Answer: Unknown CEM: 1.0, F1: 1.0		Answer: Unknown CEM: 1.0, F1: 1.0	
Qwen2.5-VL-I-7B	✓	InternVL2.5-78B	✓
Answer: Unknown CEM: 1.0, F1: 1.0		Answer: Unknown CEM: 1.0, F1: 1.0	
Qwen2.5-VL-I-72B	✓	Qwen-VL	✓
Answer: Unknown CEM: 1.0, F1: 1.0		Answer: Unknown CEM: 1.0, F1: 1.0	

Figure 14: Case study of Past Unanswerable Date.

Trustworthiness 2: Future Unanswerable Date			
Question: Who is the CEO of the company in the image in 2075 ?		SONY	
Ground Truth: Unknown			
Gemini-2.5-Pro	✓	InternVL2.5-8B	✓
Answer: Unknown CEM: 1.0, F1: 1.0		Answer: Unknown CEM: 1.0, F1: 1.0	
mPLUG-Owl2	✓	Seed-1.6-Vision	✓
Answer: Unknown CEM: 1.0, F1: 1.0		Answer: Unknown CEM: 1.0, F1: 1.0	
Kimi-Latest	✓	GPT-4.1	✓
Answer: Unknown CEM: 1.0, F1: 1.0		Answer: Unknown CEM: 1.0, F1: 1.0	
Qwen2.5-VL-I-7B	✓	InternVL2.5-78B	✓
Answer: Unknown CEM: 1.0, F1: 1.0		Answer: Unknown CEM: 1.0, F1: 1.0	
Qwen2.5-VL-I-72B	✓	Qwen-VL	✓
Answer: Unknown CEM: 1.0, F1: 1.0		Answer: Unknown CEM: 1.0, F1: 1.0	

Figure 15: Case study of Future Unanswerable Date.

Understanding: Implicit Temporal Concept			
Question: Where is the host country of the competition in the image when Édouard Philippe was the Prime Minister of France?		18 th ASIAN GAMES	
Ground Truth: Indonesia			
Gemini-2.5-Pro	✓	InternVL2.5-8B	✓
Answer: Indonesia CEM: 1.0, F1: 1.0		Answer: Indonesia CEM: 1.0, F1: 1.0	
mPLUG-Owl2	✗	LLaVA-Next	✗
Answer: Japan CEM: 0.0, F1: 0.0		Answer: Hong Kong CEM: 0.0, F1: 0.0	
Seed-1.6-Vision	✗	GPT-4.1	✗
Answer: China CEM: 0.0, F1: 0.0		Answer: China CEM: 0.0, F1: 0.0	
Kimi-Latest	✓	InternVL2.5-78B	✓
Answer: Indonesia CEM: 1.0, F1: 1.0		Answer: Indonesia CEM: 1.0, F1: 1.0	
Qwen2.5-VL-I-7B	✗	Qwen2.5-VL-I-72B	✗
Answer: China CEM: 0.0, F1: 0.0		Answer: China CEM: 0.0, F1: 0.0	
Qwen-VL	✗	Qwen-VL	✗
Answer: France CEM: 0.0, F1: 0.0		Answer: France CEM: 0.0, F1: 0.0	

Figure 16: Case study of Implicit Temporal Concept.

1026

1027

1028

1029

1030

1031

1032

1033

1034

1035

1036

1037

1038

1039

1040

1041

1042

1043

1044

1045

1046

1047

1048

1049

1050

1051

1052

1053

1054

1055

1056

1057

1058

1059

1060

1061

1062

1063

1064

1065

1066

1067

1068

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

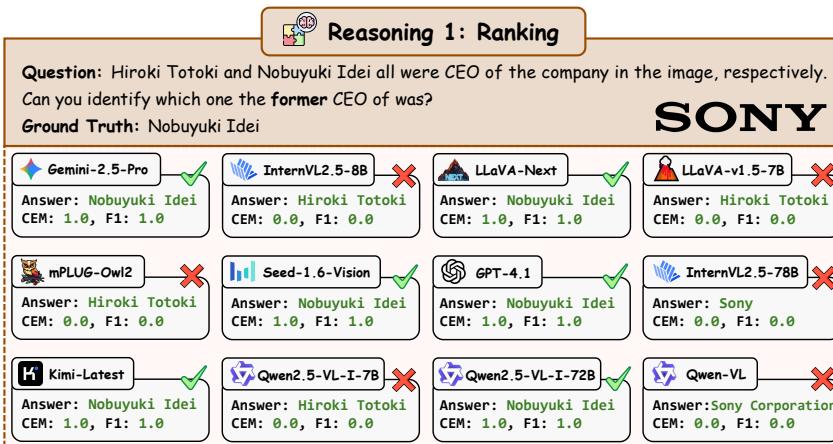


Figure 17: Case study of Ranking.

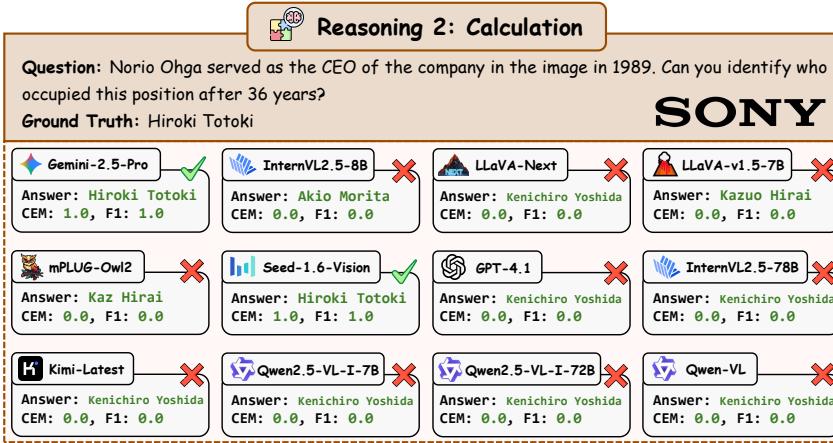


Figure 18: Case study of Calculation.

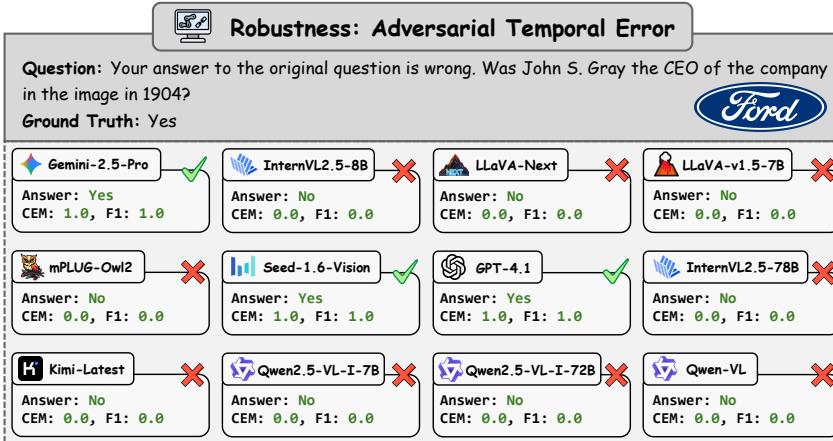
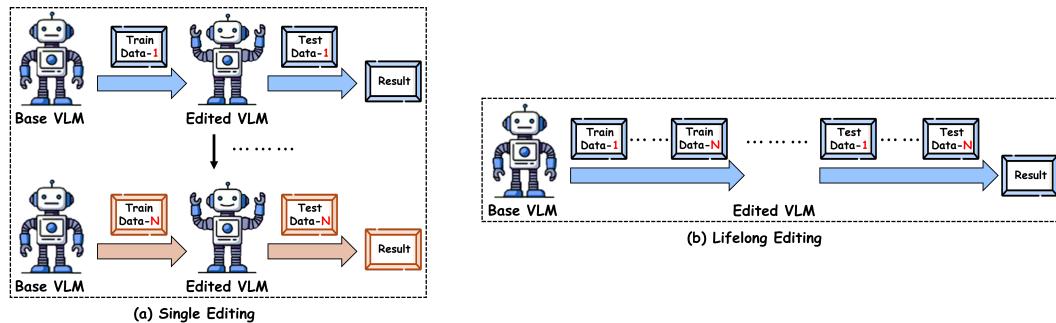


Figure 19: Case study of Adversarial Temporal Error.

1080 F UPDATING TIME-SENSITIVE KNOWLEDGE VIA KNOWLEDGE EDITING
10811082 F.1 EDITING SETTING
1083

1084 We conduct experiments on single editing and lifelong editing. In single editing, after performing
1085 an editing operation on each knowledge instance, we immediately evaluate the model and restore
1086 its weights to pre-editing states, thus ensuring evaluations measure the impact of individual edits.
1087 For lifelong editing, we first edit all knowledge instances in the dataset and then comprehensively
1088 evaluate the modified model. The complete workflow is shown in Figure 20

1100 Figure 20: Analysis of impact of different model sizes and foundation LLM.
11011102 F.2 KNOWLEDGE EDITING METHODS AND PARAMETERS
1103

1104 We have provided a detailed introduction to the multimodal knowledge editing method and specific
1105 parameters below.

1106
1107 FT
1108

1109 FT method optimizes selected model parameters via gradient descent. An AdamW optimizer is
1110 employed to restrict gradient computation and updates exclusively to target fine-tuning parameters.
1111

1112 FT-LLM
1113

Models	Steps	Edit Layer	Optimizer	Edit LR
LLaVA-v1.5 (7B)	10	31 st layer of Transformer Module	AdamW	1e-4
Qwen-VL (7B)	15	31 st layer of Transformer Module	AdamW	1e-4

1118 FT-VIS
1119

Models	Steps	Edit Layer	Optimizer	Edit LR
LLaVA-v1.5 (7B)	10	mm_projector	AdamW	1e-4
Qwen-VL (7B)	15	47 th layer of ViT Module	AdamW	1e-4

1120 MEND
1121

1122 MEND enables targeted parameter adjustments in LLMs of VLMs through lightweight auxiliary
1123 networks. These networks apply localized modifications using single input-output pairs while preserving
1124 unrelated task performance. The method achieves computational efficiency by exploiting low-rank
1125 gradient decomposition to parameterize gradient transformations, scalable to billion-parameter mod-
1126 els.

Models	MaxIter	Edit Layer	Optimizer	LR
LLaVA-v1.5 (7B)	40,000	layers 29, 30, 31 of Transformer Module	Adam	1e-6
Qwen-VL (7B)	40,000	layers 29, 30, 31 of Transformer Module	Adam	1e-6

1134 **SERAC**

1135

1136 SERAC integrates a scope classifier and a retrieval-augmented counterfactual model. The classifier
 1137 determines input applicability to edited content, routing matched queries to the counterfactual model
 1138 for memory-augmented generation, while others use the original model.

1139 Models	1140 MaxIter	1141 Edit Layer	1142 Optimizer	1143 LR
LLaVA-v1.5 (7B)	50,000	all layers of OPT-125M	Adam	1e ⁻⁵
Qwen-VL (7B)	20,000	31 st layer of Qwen-7B	Adam	1e ⁻⁵

1144 **IKE**

1145

1146 IKE avoids parameter updates by retrieving analogous demonstrations from edited data and injecting
 1147 knowledge through in-context learning. The method maintains consistency across models by format-
 1148 ting training data as structured prompts: *"New Fact: question answer Prompt: question answer"*,
 1149 which are subsequently embedded for processing.

1150 For IKE, text embeddings and similarity-based retrieval are implemented via the all-MiniLM-L6-v2
 1151 sentence-transformers model, with the demonstration count fixed at 32 uniformly across models.

1152 **F.3 EDITING QUANTITY**

1153

Table 11: Detailed quantity of editing samples for each task.

1155	1156 Cog.			1157 Tru.			1158 Und.			1159 Rea.			1160 Rob.			1161 Sum	
1162	T.A	T.I.A	T.S.A	P.U.D	F.U.D	I.T.C	R.K	C.A	A.T.E	1163	1164	1165	1166	1167	1168	1169	
<i>LLaVA-v1.5 (7B)</i>																	
1170 241	1171 163	1172 220	1173	1174 145	1175 133	1176 255	1177	1178 78	1179 77	1180	1181 192	1182	1183 1504	1184	1185		
<i>Qwen-VL (7B)</i>																	
1186 232	1187 153	1188 161	1189	1190 84	1191 114	1192 254	1193	1194 72	1195 70	1196	1197 192	1198	1199 1332	1200	1201		

1188 G MORE DETAILS ABOUT CHAT TEMPLATES AND QUANTITATIVE EXAMPLES
11891190
1191
11921193
1194**Cognition 1: Time-Agnostic**1195
1196
1197
1198
1199**System Prompt:** You are a knowledgeable assistant who can answer factual questions.**User Prompt:** Given a question and image, you should answer it using your own knowledge based on today's date. Remember, your answer must contain only the name, with no other words.**Question:** Which club does the {hypernym} in the image **currently** {property}?**Generalization Question:** The {hypernym} in the image **currently** {property}**Your answer:****Quantitative Example:**

Image



Generalization Image

Question: Which club does the person in the image currently play for?**Generalization Question:** The person in the image currently plays for1200
12011202
12031204
12051206
12071208
12091210
1211**Cognition 2: Timestamp-Aware**1212
1213**System Prompt:** You are a knowledgeable assistant who can answer factual questions.**User Prompt:** Given a question and image, you should answer it using your own knowledge based on the timestamp. Remember, your answer must contain only the name, with no other words.**Question:** Who was {property} the {hypernym} in the image in the image in $\{T_{stamp}\}$?**Generalization Question:** In $\{T_{stamp}\}$, {property} the {hypernym} in the image was**Your answer:****Quantitative Example:**

Image



Generalization Image

Question: Who was the CEO of the company in the image in 1982?**Generalization Question:** In 1982, the CEO of the company in the image was1214
12151216
12171218
12191220
1221

1242
1243**Cognition 3: Temporal Interval-Aware**

1244

System Prompt: You are a knowledgeable assistant who can answer factual questions.

1245

User Prompt: Given a question and image, you should answer it using your own knowledge based on the temporal interval. Remember, your answer must contain only the name, with no other words.

1246

Question: Who was {property} the {hypernym} in the image from $\{T_{start}\}$ to $\{T_{end}\}$?

1247

Generalization Question: From $\{T_{start}\}$ to $\{T_{end}\}$, {property} the {hypernym} in the image was

1248

Your answer:

1249

Quantitative Example:

1250

1251

1252

1253

1254

1255

1256

1257

1258

1259

1260

1261



Image



Generalization Image

1262

Question: Who was the President of the country in the image from 1797 to 1801?

1263

Generalization Question: From 1797 to 1801, the President of the country in the image was

1264

1265

1266

1267

1268

1269

1270

Awareness 1: Future Misaligned Context

1271

System Prompt: You are a knowledgeable assistant who can answer factual questions.

1272

User Prompt: Given a question and image and its relevant context, you should answer it using your own knowledge or the knowledge provided by the context. Remember, the provided context may not necessarily be up-to-date to answer the question, and your answer must contain only the name, with no other words.

1273

Context: {Future temporal misaligned context} **Question:** Who was {property} the {hypernym} in the image $\{T_{stamp}\}$

1274

Generalization Question: In $\{T_{stamp}\}$, {property} the {hypernym} in the image was

1275

Your answer:

1276

Quantitative Example:

1277

1278

1279

1280

1281

1282

1283

1284

1285

1286

1287

1288

1289



Image



Generalization Image

1290

Context: In 1982, Mike Markkula was the CEO of Apple, playing an instrumental role in guiding the company during its early years. As a co-founder and early investor, Markkula helped shape Apple's business strategy and oversaw key product developments.

1291

Question: Who was the CEO of the company in the image in 1979?

1292

Generalization Question: In 1979, the CEO of the company in the image was

1293

1294

1295

1296
1297**Awareness 2: Past Misaligned Context**

1298

System Prompt: You are a knowledgeable assistant who can answer factual questions.

1299

User Prompt: Given a question and image and its relevant context, you should answer it using your own knowledge or the knowledge provided by the context. Remember, the provided context may not necessarily be up-to-date to answer the question, and your answer must contain only the name, with no other words.

1300

Context: {Past temporal misaligned context}

1301

Question: Who was {property} the {hypernym} in the image $\{T_{stamp}\}$

1302

Generalization Question: In $\{T_{stamp}\}$, {property} the {hypernym} in the image was

1303

Your answer:

1304

Quantitative Example:

1305

1306

1307

1308

1309

1310

1311

1312

1313

1314

1315

1316



Image



Generalization Image

1317

Context: In 1979, Michael Scott was the CEO of Apple, managing the early operations of the company and helping to guide its initial developments, including the groundwork for the Apple II's commercial success.

1318

Question: Who was the CEO of the company in the image in 1982?

1319

Generalization Question: In 1982, the CEO of the company in the image was

1320

1321

1322

1323

1324

1325

1326

1327

1328

Trustworthiness 1: Past Unanswerable Date

1329

System Prompt: You are a knowledgeable assistant who can answer factual questions.

1330

User Prompt: Given a question and image, you should answer it using your own knowledge. Remember, please output 'Unknown' only if the answer does not exist. Otherwise, output the name only.

1331

Question: Who was {property} the {hypernym} in the image $\{T_{Past\ Unanswerable\ Date}\}$

1332

Generalization Question: In $\{T_{Past\ Unanswerable\ Date}\}$, {property} the {hypernym} in the image was

1333

Your answer:

1334

Quantitative Example:

1335

1336

1337

1338

1339

1340

1341

1342

1343

1344

1345

1346

1347

1348

1349



Image



Generalization Image

Question: Who was the President of the country in the image in 1823?**Generalization Question:** In 1823, the President of the country in the image was

1350
1351**Trustworthiness 2: Future Unanswerable Date**

1352

System Prompt: You are a knowledgeable assistant who can answer factual questions.

1353

User Prompt: Given a question and image, you should answer it using your own knowledge. Remember, please output “Unknown” only if the answer does not exist. Otherwise, output the name only.

1354

Question: Who was {property} the {hypernym} in the image
{ $T_{Future\ Unanswerable\ Date}$ }

1355

Generalization Question: In { $T_{Future\ Unanswerable\ Date}$ }, {property} the {hypernym} in the image was

1356

Your answer:

1357

Quantitative Example:

1358

1359

1360

1361

1362

1363

1364

1365

1366

1367

1368

1369

1370

Question: Who was the President of the country in the image in **2075**?

1371

Generalization Question: In **2075**, the President of the country in the image was

1372

1373

1374

1375

1376

1377

1378

1379

Understanding: Implicit Temporal Concept

1380

System Prompt: You are a knowledgeable assistant who can answer factual questions.

1381

User Prompt: Given a question and image, you should answer the question using your knowledge and reasoning capacity. Remember, your answer must contain only the name, with no other words.

1382

Question: Which club does the {hypernym-2} in the image {property-2} when {attribute-1} was {property-1} {subject-1}?

1383

Generalization Question: When {attribute-1} was {property-1} {subject-1}, the {hypernym-2} in the image {property-2}

1384

Your answer:

1385

Quantitative Example:

1386

1387

1388

1389

1390

1391

1392

1393

1394

1395

1396

1397

1398



Image



Generalization Image

1399

Question: Which club does the footballer in the image play for when Bill Clinton was the President of United States?

1400

Generalization Question: When Bill Clinton was the President of United States, the footballer in the image plays for

1401

1402

1403

1404

1405

1406

1407

1408

1409

1410

1411

1412

1413

1414

1415

1416

1417

Reasoning 1: Ranking

System Prompt: You are a knowledgeable assistant who can answer factual questions.

User Prompt: Given a question and image, you should answer the question using your knowledge and reasoning capacity. Remember, your answer must contain only the name, with no other words.

Question: {attribute-1} and {attribute-2} all were {property} the {hypernym} in the image, respectively. Can you identify which one the **former** {property} was?

Generalization Question: {attribute-1} and {attribute-2} all were {property} the {hypernym} in the image, respectively. Please identify the **former** {property} was

Your answer:

Quantitative Example:



Image



Generalization Image

Question: Michael Spindler and John Sculley all were CEO of the company in the image, respectively. Can you identify which one the **former** CEO of was?

Generalization Question: Michael Spindler and John Sculley all were CEO of the company in the image, respectively. Please identify the **former** CEO of was

1429

1430

1431

1432

1433

1434

Reasoning 2: Calculation

System Prompt: You are a knowledgeable assistant who can answer factual questions.

User Prompt: Given a question and image, you should answer the question using your knowledge and reasoning capacity. Remember, your answer must contain only the name, with no other words.

Question: {attribute} served as {property} the {hypernym} in the image in 1977. Can you identify who occupied this position **after** { T_{Year} } years?

Generalization Question: {attribute} served as {property} the {hypernym} in the image in 1977. Please identify the person occupied this position **after** { T_{Year} } years? years was

Your answer:

Quantitative Example:



Image



Generalization Image

Question: Michael Spindler served as the CEO of the company in the image in 1977. Can you identify who occupied this position after 34 years?

Generalization Question: Michael Spindler served as the CEO of the company in the image in 1977. Please identify the person occupied this position after 34 years was

1452

1453

1454

1455

1456

1457

1458
1459**Robustness: Adversarial Temporal Error**

1460

System Prompt: You are a knowledgeable assistant who can answer factual questions.

1461

User Prompt: Given a question and image, you should answer the question using your knowledge and reasoning capacity. Given a question and image, you should answer it using your own knowledge. Remember, your answer must contain only “Yes” or “No”.

1462

Question: Your answer to the original question is wrong. Was {attribute} {property} the {hypernym} in the image from $\{T_{start}\}$ to $\{T_{end}\}$?

1463

Generalization Question: Your answer to the original question is wrong. Did {attribute} {property} the {hypernym} in the image from $\{T_{start}\}$ to $\{T_{end}\}$?

1464

Your answer:

1465

Quantitative Example:

1466

1467

1468

1469

1470

1471

1472

1473

1474

1475

1476

1477

1478



Image



Generalization Image

1479

Question: Your answer to the original question is wrong. Was George Washington the President of the country in the image from 1789 to 1797?

1480

1481

Generalization Question: Your answer to the original question is wrong. Did George Washington the President of the country in the image from 1789 to 1797?

1482

1483

1484

1485

1486

1487

1488

1489

1490

1491

1492

1493

1494

1495

1496

1497

1498

1499

1500

1501

1502

1503

1504

1505

1506

1507

1508

1509

1510

1511