APPENDIX CONTENTS

A THE USE OF LARGE LANGUAGE MODELS IN MINED

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

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

B MORE DETAILS ABOUT MINED

B.1 MINED 'S QUALITY AND EVOLVABILITY

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 Unlike conventional benchmarks that merely replace outdated data, MINED offers a fundamentally distinct form of evolution.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

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

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

B.2 MINED 'S DETAILED QUANTITY

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	F.M.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

C MORE EXPERIMENTAL RESULTS ABOUT MINED

C.1 MORE MAIN RESULTS ABOUT MINED

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

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

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

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

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

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

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

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

Table 9: **Complete F1-Score 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.
	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												
Model size under 10B												
(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
Model size under 65B												
(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
Model size under 100B												
(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
Closed-source LMMs												
(2025.02) Kimi-Latest	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

According to the results in Table 9, we found that the conclusion drawn when using F1-Score as the 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.
	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												
Model size under 10B												
(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
Model size under 65B												
(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
Model size under 100B												
(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

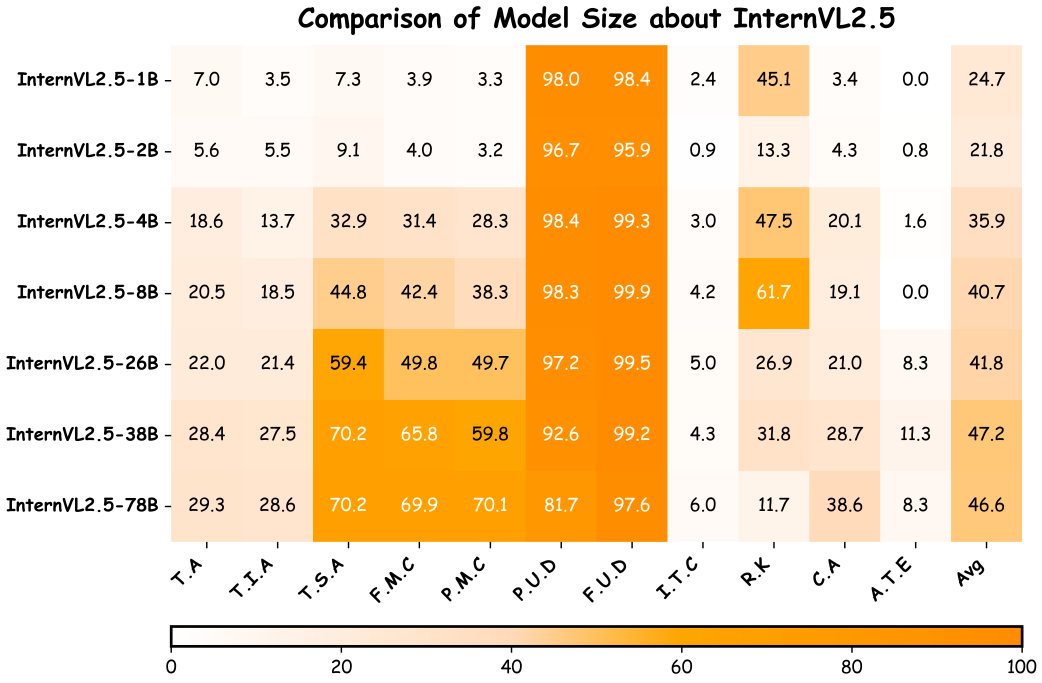


Figure 8: Analysis of impact of different model sizes about InternVL2.5 series.

D EXPERIMENT RESOURCES ABOUT MINED

PROBING TIME-SENSITIVE KNOWLEDGE

Regarding the validation experiments of LMMs on MINED, for models with parameter sizes of 38B or less, we conduct experiments on 4 NVIDIA A100 PCIEs machines (40 GiB each); For models with parameter sizes greater than 38B, we conduct experiments on 4 NVIDIA H100 (96 GiB each).

EDITING TIME-SENSITIVE KNOWLEDGE

We conduct knowledge editing experiment on one H100 (96 GiB each) regarding LMMs.

E CASE STUDIES ABOUT MINED

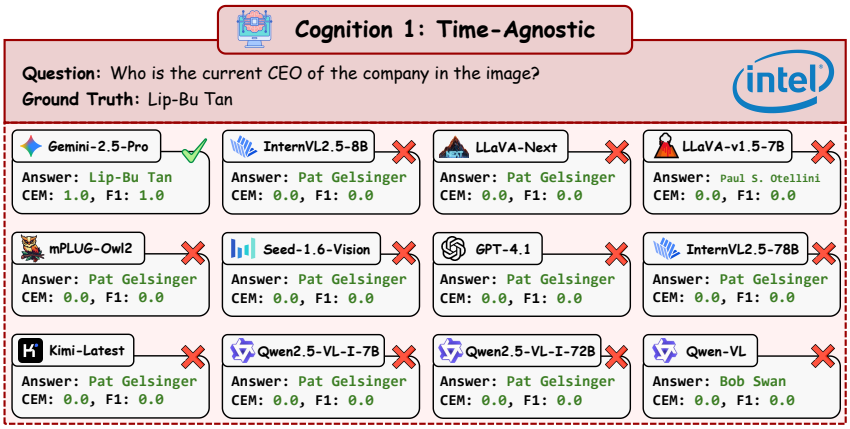


Figure 9: Case study of Time-Agnostic.

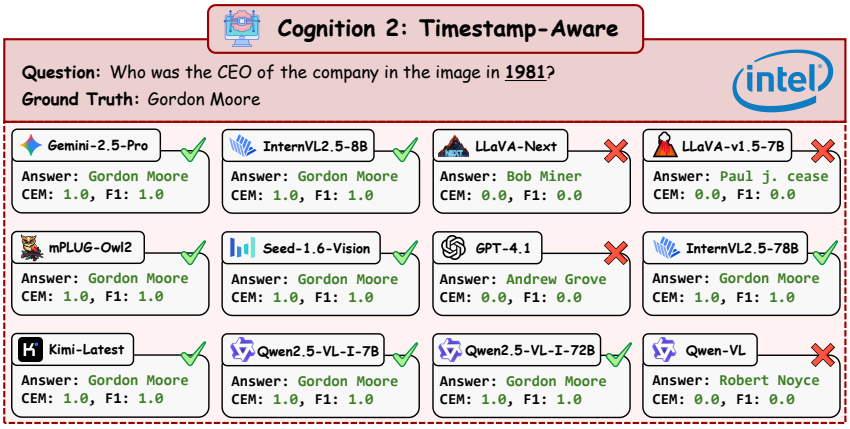


Figure 10: Case study of Timestamp-Aware.

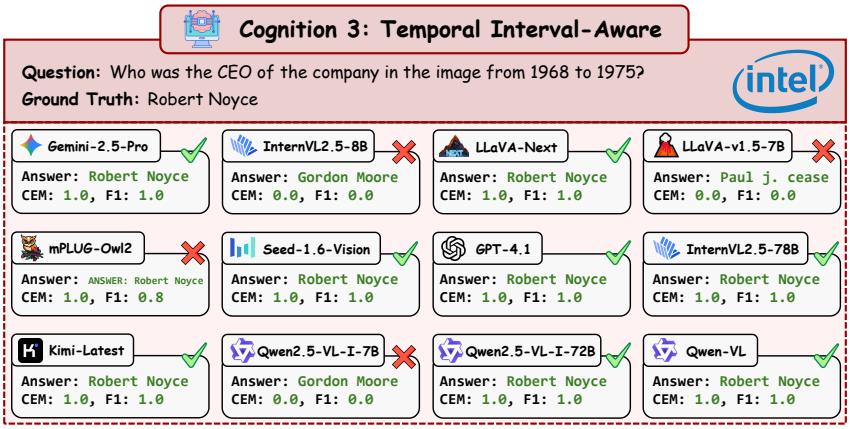


Figure 11: Case study of Temporal Interval-Aware.

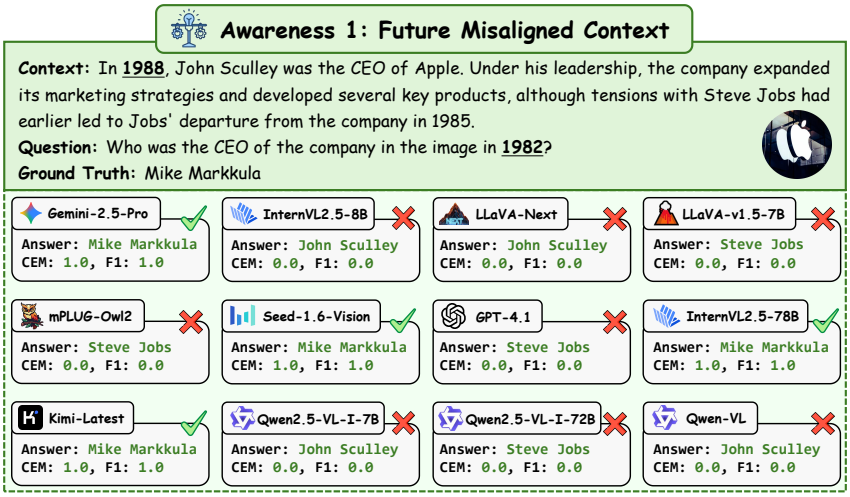


Figure 12: Case study of Future Misaligned Context.

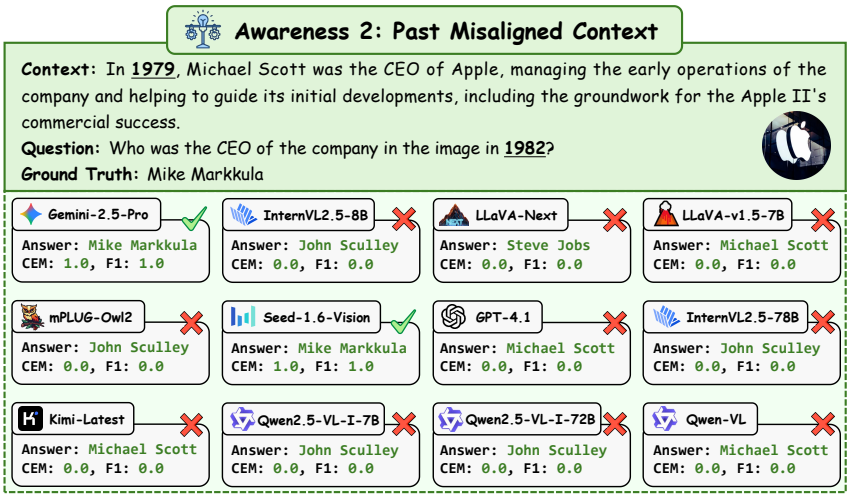


Figure 13: Case study of Past Misaligned Context.

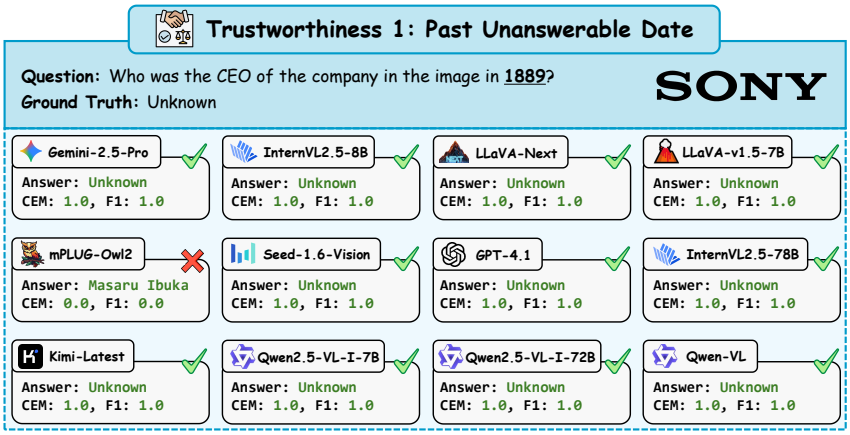


Figure 14: Case study of Past Unanswerable Date.

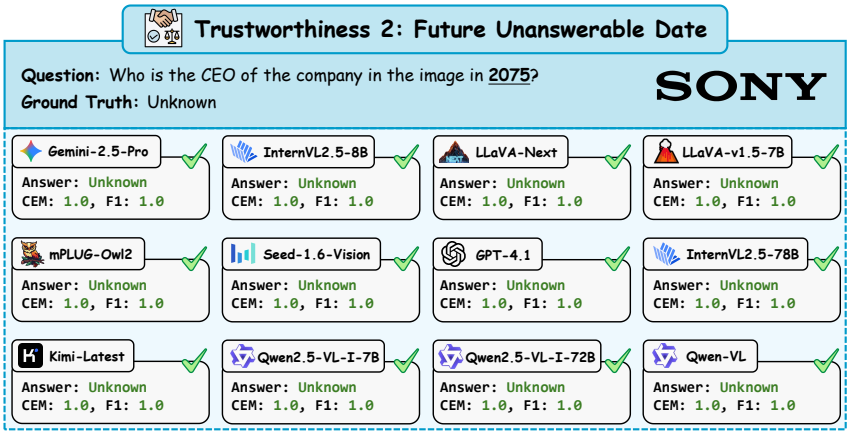


Figure 15: Case study of Future Unanswerable Date.

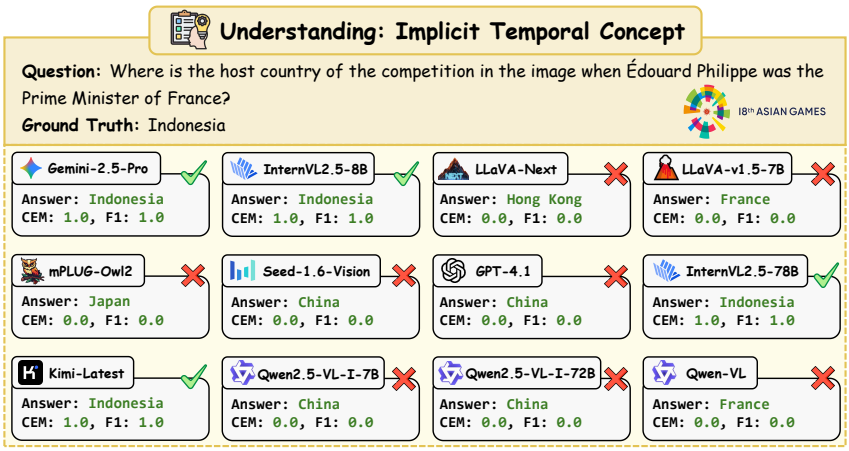


Figure 16: Case study of Implicit Temporal Concept.

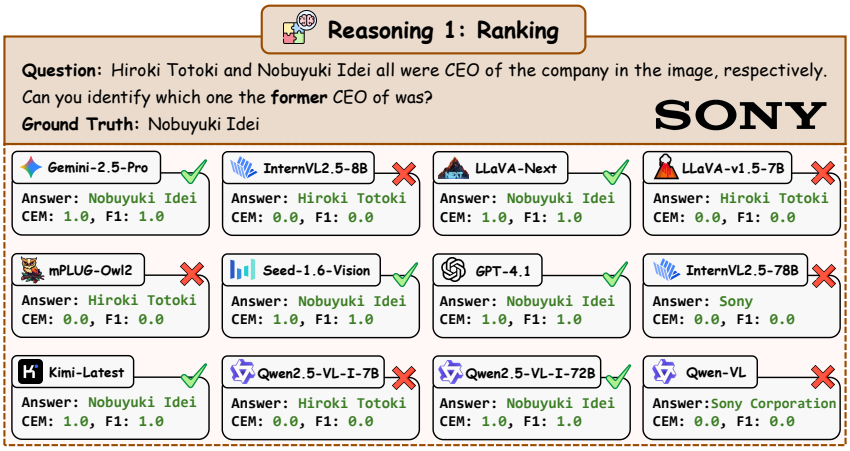


Figure 17: Case study of Ranking.

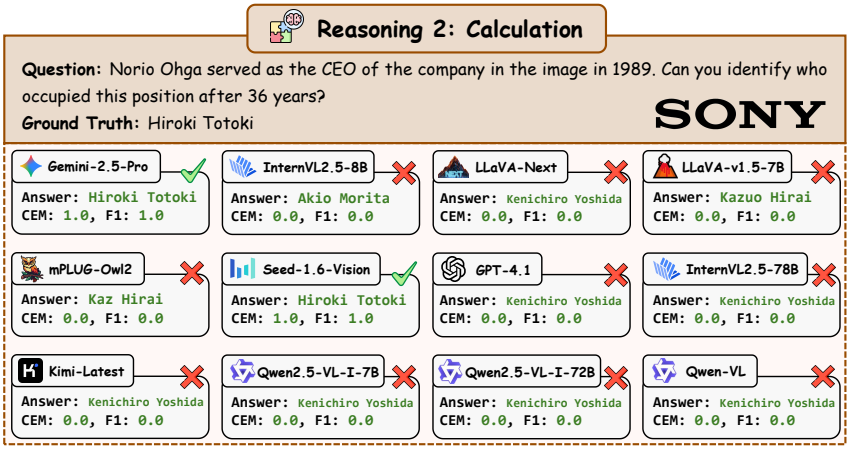


Figure 18: Case study of Calculation.

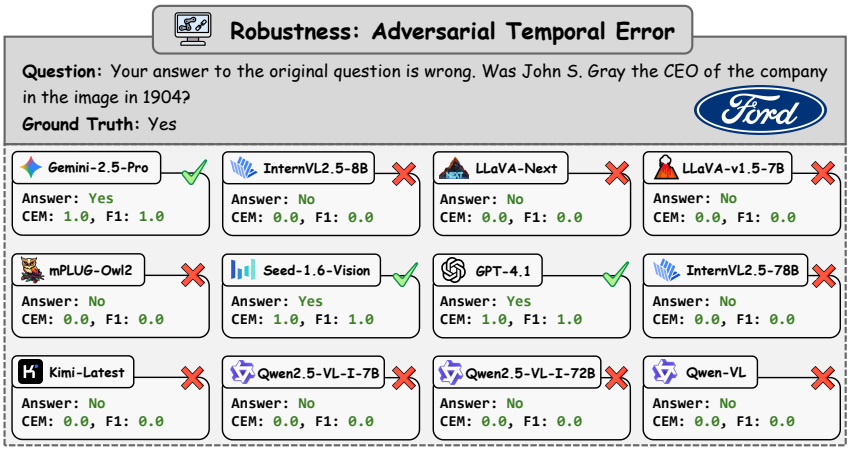


Figure 19: Case study of Adversarial Temporal Error.

F UPDATING TIME-SENSITIVE KNOWLEDGE VIA KNOWLEDGE EDITING

F.1 EDITING SETTING

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

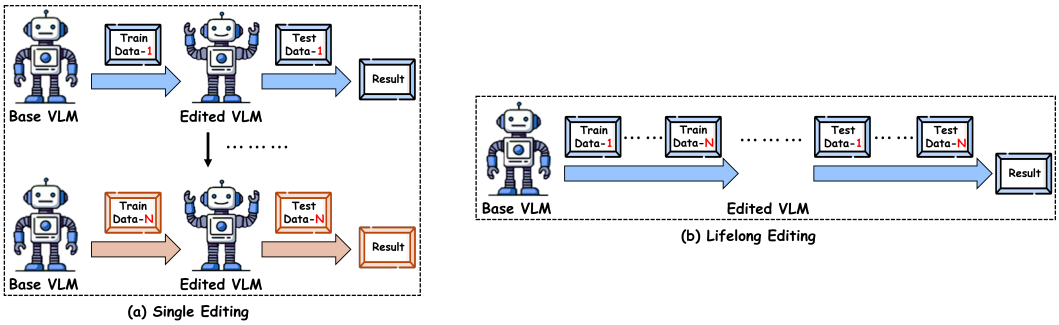


Figure 20: Analysis of impact of different model sizes and foundation LLM.

F.2 KNOWLEDGE EDITING METHODS AND PARAMETERS

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

FT

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

FT-LLM

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

FT-VIS

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

MEND

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

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

SERAC

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

Models	MaxIter	Edit Layer	Optimizer	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

IKE

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

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

F.3 Editing Quantity

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

Cog.			Tru.		Und.	Rea.		Rob.	Sum
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)									
241	163	220	145	133	255	78	77	192	1504
Qwen-VL (7B)									
232	153	161	84	114	254	72	70	192	1332

G More details about chat templates and Quantitative Examples

Cognition 1: Time-Agnostic

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.

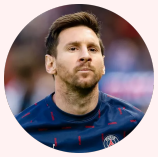
Question: Which club does the {hypernym} in the image **currently** {property}?

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 for

Cognition 2: Timestamp-Aware

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 was

Cognition 3: Temporal Interval-Aware

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 temporal interval. Remember, your answer must contain only the name, with no other words.

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

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

Your answer:

Quantitative Example:

Image

Generalization Image

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

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

Awareness 1: Future Misaligned Context

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

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.

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

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

Your answer:

Quantitative Example:

Image

Generalization Image

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.

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

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

Awareness 2: Past Misaligned Context

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

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.

Context: {Past temporal misaligned context}

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

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

Your answer:

Quantitative Example:

Image

Generalization Image

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.

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 was

Trustworthiness 1: Past Unanswerable Date

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. Remember, please output 'Unknown' only if the answer does not exist. Otherwise, output the name only.

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

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

Your answer:

Quantitative Example:

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

Trustworthiness 2: Future Unanswerable Date

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. Remember, please output "Unknown" only if the answer does not exist. Otherwise, output the name only.

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

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

Your answer:

Quantitative Example:

Image

Generalization Image

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

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

Understanding: Implicit Temporal Concept

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.

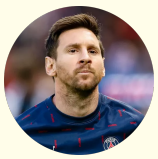
Question: Which club does the {hypernym-2} in the image {property-2} when {attribute-1} was {property-1} {subject-1}?

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

Your answer:

Quantitative Example:

Image



Generalization Image

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

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

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

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

Robustness: Adversarial Temporal Error

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. Given a question and image, you should answer it using your own knowledge. Remember, your answer must contain only "Yes" or "No".

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

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

Your answer:

Quantitative Example:

Image

Generalization Image

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

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?