ELTLM: Evaluation of Longitudinal Temporal Large Multimodal Models in Clinical Scenarios

Gengyuan Hu[1,2] Haoxiang Liu[3]* Chenhong Cao[1,2] Shilei Tan[1,2] Wei Gong[1,2]
[1]University of Science and Technology of China, China
[2]Ubiquitous Battery-free Internet of Things Lab, USTC, China
[3]City University of Hong Kong, China
hugengyuan@mail.ustc.edu.cn
lhx0612@gmail.com
{chcao, tanshilei, weigong}@ustc.edu.cn

Abstract

Large Multimodal Models (LMMs) have demonstrated significant potential in the medical domain, achieving impressive performance on tasks ranging from report generation to visual question answering. However, existing benchmarks predominantly focus on static evaluation, assessing models on isolated data points. This approach neglects a critical aspect of clinical practice: longitudinal analysis, where physicians interpret patient data as a dynamic trajectory to track disease progression and treatment response. To address this gap, we introduce ELTLM, the first benchmark specifically tailored to assess the temporal perception and reasoning capabilities of medical LMMs. Constructed from temporal chest X-rays, ELTLM features a hierarchical task taxonomy comprising Temporal Perception QA and Temporal Reasoning QA, requiring models to detect fine-grained visual changes and infer high-level clinical trends. Our evaluation of state-of-the-art models reveals that while they excel in static scenarios, they struggle significantly with temporal grounding and consistency. ELTLM serves as a vital resource to identify these limitations and guide the development of future time-aware medical AI systems. Our data is available at ELTLM.

1 Introduction

Reliable evaluation benchmarks serve as the compass for the advancement of Large Multimodal Models (LMMs) in the medical domain. As models like GPT-5 (OpenAI, 2025) and open-source medical adaptations like HuatuoGPT-Vision (Chen et al., 2024) demonstrate increasing proficiency in interpreting medical visual data, the community has established several benchmarks to gauge their capabilities. Existing benchmarks, such as VQA-RAD (Lau et al., 2018) and SLAKE (Liu et al., 2021), have provided standardized grounds for assessing visual perception and knowledge retrieval capabilities. These benchmarks typically present a model with a single medical image and a corresponding question, effectively testing the model's ability to perform single medical image analysis—interpreting a patient's status at a static point in time.

However, existing evaluation paradigms conspicuously overlook a fundamental dimension of real-world clinical practice: the *longitudinal* nature of patient care. A patient's health is rarely defined by a static snapshot; rather, it is a dynamic trajectory. Physicians routinely rely on temporal context, comparing current observations with historical records to track disease progression, evaluate treatment efficacy, and distinguish acute changes from chronic conditions. Consequently, benchmarks that focus solely on static image interpretation fail to capture the complexity of clinical reasoning, leaving a critical gap in our understanding of how LMMs would perform in realistic, time-dependent scenarios.

To bridge this gap, we introduce **ELTLM** (**E**valuation of **L**ongitudinal **T**emporal **L**arge **M**ultimodal Models in Clinical Scenarios), a novel benchmark specifically designed to assess the longitudinal reasoning capabilities of medical LMMs. The construction of ELTLM relies on a rigorous data curation pipeline that moves beyond isolated instances. We meticulously curate sequential patient records, aligning temporal chest X-rays with their corresponding temporal clinical reports. This construction ensures that the data retains high clinical relevance and temporal coherence, requiring models to process and correlate information across time rather than in isolation.

Within ELTLM, we design two tasks to provide a granular analysis of model capabilities:

- **Temporal Perception QA:** This task level focuses on fine-grained visual comparison. It requires the model to identify specific changes

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 15765–15779

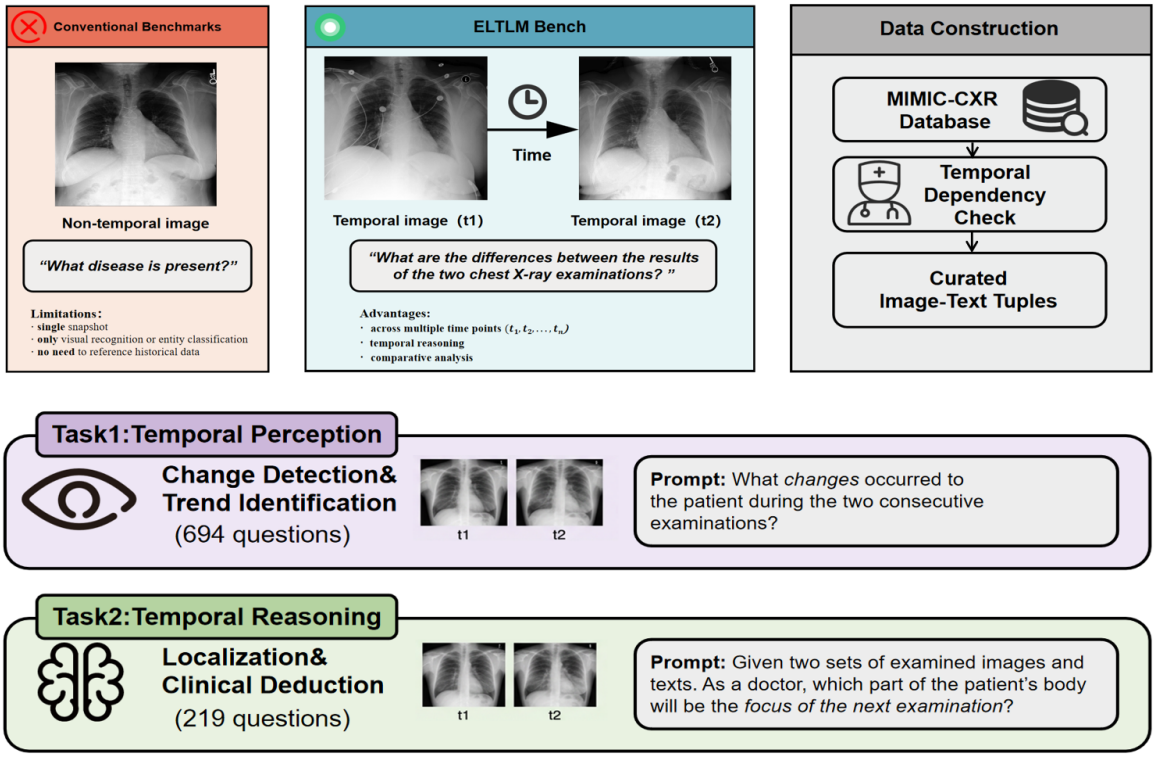


Figure 1: **Overview of ELTLM.** (Top) a temporal medical image-text benchmark for chest X-ray understanding. Unlike conventional single-snapshot benchmarks (left, limited to non-temporal image classification without historical reference), ELTLM constructs curated temporal image-text tuples via the MIMIC-CXR database (right). (Bottom) ELTLM supports two core task modules:Temporal Perception (694 samples): Tasks like change detection and trend identification ;Temporal Reasoning (219 samples): Clinical tasks like localization and follow-up deduction.

in anatomical features or lesion characteristics between two time points (e.g., "Compare the density of the opacity in the left upper lobe between the baseline and follow-up scans: has it increased, decreased, or remained stable?").

- **Temporal Reasoning QA:** Moving beyond perception, this task demands higher-level clinical logic. Models must synthesize the observed changes to infer disease progression trends or management implications (e.g., "Given the changes in the pleural effusion and consolidation over the past three visits, is the current treatment regimen likely effective?").

Our comprehensive evaluation of state-of-the-art generic and medically-adapted LMMs on ELTLM reveals significant challenges. While many models demonstrate proficiency in static benchmarks, they exhibit notable performance drops when tasked with maintaining temporal consistency and reasoning across longitudinal data.

In summary, our contributions are as follows:

- **Novel Benchmark for Longitudinal Analysis:** We propose ELTLM, the first comprehensive benchmark tailored to assess the temporal perception and reasoning capabilities of LMMs in the medical domain, shifting the evaluation focus from static perception to dynamic progression tracking.
- **High-Quality Temporal Dataset Construction:** We introduce a rigorous data curation pipeline that effectively aligns sequential multimodal medical data, ensuring the benchmark supports valid and clinically meaningful temporal evaluation.
- **Hierarchical Evaluation Taxonomy:** We design a structured task set comprising Temporal Difference QA and Temporal Reasoning QA, allowing for a nuanced assessment of how well models can perceive changes and reason about them over time.
- **Critical Insights into SOTA Limitations:** Our extensive experiments demonstrate that current top-tier LMMs lack robustness in longitudinal settings, highlighting specific weaknesses in temporal attention and reasoning that pave the way for future architectural improvements.

2 Related Work

2.1 Benchmarks for Medical Large Multimodal Models

The evaluation of Medical Large Multimodal Models (LMMs) has evolved from early VQA datasets (e.g., VQA-RAD, SLAKE) to comprehensive benchmarks like OmniMedVQA (Hu et al., 2024) and the medical subset of MMMU (Yue et al., 2024), which cover diverse modalities and tasks. Concurrently, model capabilities have advanced through both general foundation models (e.g., GPT-4 (OpenAI et al., 2023), Qwen (Bai et al., 2023), Gemini (Team et al., 2023)) and domain-specific adaptations (e.g., LLaVA-Med (Li et al., 2023), Med-Flamingo (Moor et al., 2023), Baichuan (Yang et al., 2023)). Despite efforts to address hallucinations (Liu et al., 2022), existing benchmarks predominantly remain *static*. In contrast to general vision benchmarks that extensively evaluate temporal understanding (e.g., MVBench (Li et al., 2024), Video-LLaVA (Lin et al., 2024)), medical assessment has traditionally treated clinical cases as dynamic snapshots at different time points. While existing temporal medical benchmarks focus primarily on high-frame-rate video analysis during surgical procedures (e.g., CholecT45 (Nwoye and Padoy, 2022), (Ghorbani et al., 2020)), they largely overlook the long-term pathological evolution that has only recently begun to be addressed by limited studies (Cui et al., 2025).

2.2 Longitudinal Analysis and Temporal Reasoning

Clinical decision-making is fundamentally longitudinal, requiring the correlation of current findings with historical records. Prior to LMMs, this was addressed by specialized models for lesion change detection (Yan et al., 2018) or risk prediction via EHR sequences (Choi et al., 2016; Li et al., 2020). In the realm of LMMs, general video benchmarks (e.g., Video-ChatGPT (Maaz et al., 2024)) focus on short-term frame-to-frame dynamics, which differ significantly from the discrete, long-term semantic shifts seen in patient visits. Furthermore, while datasets like MIMIC-CXR (Johnson et al., 2019) contain longitudinal metadata, standard evaluation protocols isolate individual studies, thereby neglecting inter-study temporal dependencies.

2.3 Bridging the Gap

To address the disconnect between static evaluation and dynamic clinical needs, we introduce ELTLM. Unlike prior benchmarks that focus on single-time interpretation, ELTLM necessitates reasoning about patient condition changes over time. By curating linked sequential data and defining tasks requiring cross-time reasoning, we provide a systematic framework to evaluate the capability of Medical LMMs in longitudinal patient management.

3 ELTLM

To rigorously evaluate the longitudinal capabilities of Medical LMMs, we construct a specialized dataset derived from the MIMIC-CXR database. Our dataset focuses on chest radiography, a domain chosen for its rich pathological diversity and the critical role of temporal monitoring in clinical workflows (e.g., tracking pneumonia progression or assessing edema resolution).

3.1 Data Sourcing and Processing

We curate patient records containing sequential imaging studies—spanning diagnosis to treatment response—to construct *temporal tuples*. Unlike standard datasets treating images as independent samples, each entry comprises a chest X-ray sequence $T = \{t_1, \ldots, t_n\}$, a query Q, and a ground-truth answer A. This longitudinal formulation necessitates reasoning over chronological dependencies across multiple visual inputs, rather than analyzing isolated static images. To ensure valid temporal correlations between sequential images, we adopt strict inclusion criteria: only sequences from records explicitly containing comparative terms (e.g., "compared to prior study", "interval change", "worsened", "improved") are included. Additionally, all queries are clinically certified to ensure that the required temporal information can be reliably derived from the corresponding image sequences.

3.2 Task Taxonomy and VQA Pair Construction

The dataset comprises a total of 913 open-ended question-answer pairs, involving 913 patients, categorized into two distinct capability levels: **Temporal Perception** and **Temporal Reasoning**.Clinically, it is an important task to perceive the changes in the patient's condition over different periods of time and to make reasoning diagnoses

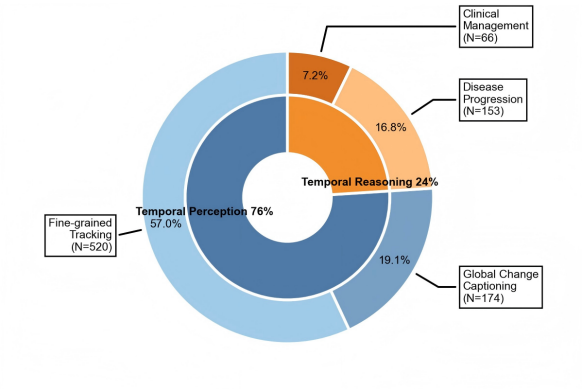


Figure 2: Distribution of tasks within the ELTLM benchmark. The inner ring represents the two primary cognitive levels: **Temporal Perception** and **Temporal Reasoning** . The outer ring details the hierarchical subtasks, highlighting the dataset's focus on fine-grained attribute tracking and disease progression assessment.

based on the patient's condition. These two tasks provide a fine-grained assessment of LLM's ability to identify chest diseases from medical information. We have created open-ended questions for these two tasks.

Temporal Perception Tasks (694 samples): These tasks assess the model's fundamental ability to perform visual subtraction and state comparison between time points t_i and t_j. We define two sub-categories based on the scope of the question:

- **Global Change Captioning:** The model is presented with open-ended prompts (e.g., *"What are the differences between the results of the two chest X-ray examinations?"*). This requires the model to scan the entire image pair, identifying multiple concurrent changes such as the expansion of effusions, the appearance of new opacities, or shifts in cardiac silhouette, and summarize them in a coherent narrative.

- **Fine-grained Attribute Tracking:** These questions focus on specific anatomical regions or medical devices (e.g., *"What changes were observed in the position of the Swan-Ganz catheter and the intra-aortic balloon pump?"* or *"Compare the left apical pneumothorax size between the two visits."*). The model must ignore irrelevant visual noise and ground its attention solely on the specified target to detect subtle evolutions in size, density, or position.

Temporal Reasoning Tasks (219 samples): Building upon perception, these tasks require the model to synthesize visual changes with clinical knowledge to derive high-level conclusions. This simulates the cognitive process of a physician determining the "clinical trajectory." We categorize these into:

- **Disease Progression Assessment:** The model must interpret the directionality of pathological changes to judge the patient's status. Unlike simple change detection, this requires semantic understanding of "improvement" versus "deterioration." For instance, a question might ask: *"Given the changes in bibasilar opacities, has the patient's pneumonia improved or worsened? Provide reasons."*

- **Clinical Management Prediction:** These tasks test the model's ability to predict future clinical actions based on historical trends. Questions involve deducing necessary interventions or diagnostic steps (e.g., *"Based on the worsening pleural effusion and new consolidation, propose the next immediate treatment plan."* or *"Predict what tests the patient will undergo next to evaluate the new pericardial findings."*). This represents the highest level of longitudinal intelligence, bridging the gap between visual perception and clinical decision-making.

3.3 Quality Assurance and Human Verification

To ensure benchmark reliability, we implemented a rigorous human-in-the-loop verification with board-certified radiologists. The process validates three criteria: (1) **Factuality**, ensuring ground truth aligns with radiological findings; (2) **Clinical Relevance**, confirming questions mimic authentic physician inquiries; and (3) **Temporal Dependency Check**, a critical step verifying that answers strictly necessitate longitudinal comparison. This explicitly precludes responses derived from single images or static priors, ensuring the evaluation reflects genuine temporal reasoning.

Table 1: Summarizes the statistics of the constructed dataset.

Task Type	Count	Focus
Perception	694	Change Detection
Reasoning	219	Disease diagnosis
Total		**913**

Table 2: **Main Results on Longitudinal Capabilities.** Comparison of general-purpose and medical-specific LLMs. **Temporal Perception** focuses on change detection, while **Temporal Reasoning** involves higher-order logic. Note that ROUGE-L is represented by the F1 score. Accuracy (Acc) and Completeness (Comp) scores are evaluated by GPT-4o-mini to ensure consistency. The best results in each column are highlighted in **bold**.

Model	Temporal Perception			Temporal Reasoning		
	ROUGE-L	Acc	Comp	ROUGE-L	Acc	Comp
Generalist Models						
GPT-5	0.145	**0.359**	0.271	0.171	**0.639**	**0.699**
GPT-4o-2024-05-13	0.158	0.066	0.247	0.179	0.375	0.390
Gemini-2.5-Pro	0.175	0.111	0.220	0.206	0.360	0.370
Claude-3.7-Sonnet	0.172	0.146	0.206	0.221	0.505	0.420
Qwen-VL-Max	0.177	0.098	0.233	**0.231**	0.500	0.585
Qwen2.5-VL-72B-Instruct	0.137	0.057	0.254	0.177	0.380	0.402
Doubao-1.5-Vision-250328	0.099	0.022	0.265	0.139	0.225	0.245
Medical Models						
Lingshu-7B	0.147	0.058	0.274	0.180	0.350	0.300
HuatuoGPT-Vision-7B	0.116	0.038	0.309	0.116	0.325	0.670
MedGemma-4B-it	**0.186**	0.042	**0.368**	0.221	0.350	0.525

3.4 Evaluation Metrics:

To provide a holistic assessment, we employ two types of metrics:

- **ROUGE-L** (Lin, 2004): This metric measures the lexical overlap based on the Longest Common Subsequence (LCS) between the generated response and the ground truth. Unlike n-gram based metrics, ROUGE-L captures sentence-level structure by identifying the longest co-occurring sequence of words in their relative order.

- **LLM-based Assessment (GPT-4o-mini):** Given the open-ended nature of our tasks, standard lexical metrics often fail to capture semantic validity in medical reasoning. To address this, we utilize GPT-4o-mini as an expert evaluator $\mathcal{E}$ to assess the quality of the generated response $\hat{y}$ against the ground truth reference y, conditioned on the input query x. The evaluator assigns binary scores along two distinct dimensions:

  Accuracy (S_{acc}): Evaluates factual accuracy and absence of hallucinations.

$$S_{acc}(\hat{y}, y) = \begin{cases} 1, & \text{if } \hat{y} \text{ aligns with key facts in } y \text{ without major errors;} \\ 0, & \text{if } \hat{y} \text{ contains factual errors or misinterprets data.} \end{cases} \quad (1)$$

  Completeness (S_{comp}): Assesses the recall of essential clinical details.

$$S_{comp}(\hat{y}, y) = \begin{cases} 1, & \text{if } \hat{y} \text{ covers all critical points present in } y; \\ 0, & \text{if } \hat{y} \text{ misses core information or fails the prompt.} \end{cases} \quad (2)$$

  We report the percentage of samples achieving a score of 1 for each metric. The detailed prompt template used for this automated evaluation is provided in the Appendix.

Implementation Details: All experiments are conducted in a **Zero-shot** setting. We provide the models with the sequential images and the question without any few-shot examples or chain-of-thought demonstrations, strictly testing their innate capability to generalize to medical longitudinal contexts.

4 Experiments and Analysis

4.1 Experimental Setup

Models Evaluated: We evaluate a comprehensive suite of state-of-the-art Large Multimodal Models (LMMs). The experimental environment and hyperparameters for each model are strictly aligned with their official source code. First, we select leading general-purpose models, including: **GPT-5** (OpenAI, 2025), **GPT-4o-2024-05-13** (Hurst et al., 2024), **Gemini-2.5-pro** (Comanici et al., 2025), **Claude-3.7-Sonnet-Latest** (Anthropic, 2025), **Qwen-VL-Max** (Bai et al., 2025), **Qwen2.5-VL-72B-Instruct**, and **Doubao-1.5-Vision-Pro-250328** (ByteDance, 2025). In addition, to assess domain-specific adaptation, we eval-

uate a series of specialized medical LMMs, including **Lingshu-7B** (Xu et al., 2025), **HuatuoGPT-Vision-7B**, and **MedGemma-4B-it** (Sellergren et al., 2025). Across all evaluations, we employ zero-shot prompting and encourage the models to provide explicit reasoning steps before delivering their final answers.

4.2 Main Results

Table 2 presents the performance of various Large Multimodal Models on the proposed ELTLM benchmark. The overall results clearly demonstrate that processing longitudinal medical data poses significant challenges to current state-of-the-art models.

4.2.1 Task 1: Temporal Perception

In the Temporal Perception task, models are required to identify and describe the evolution of pathologies across multiple images. As shown in Table 2, this task proves to be extremely difficult for existing models.

- **Significant Accuracy Bottleneck:** With the notable exception of **GPT-5**, which achieves an accuracy of 35.9%, most models struggle to surpass a 15% accuracy threshold. For instance, open-weights models like **Lingshu-7B** and **HuatuoGPT-Vision-7B** achieve only 5.8% and 3.8% accuracy, respectively. This indicates a severe deficiency in fine-grained visual change detection.
- **Hallucination:** A crucial observation is the discrepancy between Accuracy (Acc) and Completeness (Comp). While models like **MedGemma-4B-it** achieve a relatively high completeness score of 0.368 (surpassing GPT-5's 0.271), their accuracy remains negligible at 4.2%. This aligns with the trend illustrated in Figure 4, suggesting that models tend to generate lengthy, generic, and structurally complete medical descriptions that are statistically plausible but factually incorrect regarding the specific temporal changes. They fail to ground their generation in the actual visual trajectory.

4.2.2 Task 2: Temporal Reasoning

The Temporal Reasoning task challenges the models to deduce clinical implications based on the trajectory.

- **Performance Gap: GPT-5** dominates this task with scores of 0.639 (Acc) and 0.699 (Comp), demonstrating a robust capability to infer high-level clinical logic. Other proprietary models like **Claude-3.7-Sonnet** and **Qwen-VL-Max** also show competitive performance (around 50% accuracy), significantly outperforming the perception task. This suggests that once visual information is (even partially) captured, strong LLM backbones can effectively leverage medical knowledge for reasoning.
- **General vs. Medical Models:** Surprisingly, general-purpose models generally outperform medical-specific models (Lingshu, HuatuoGPT) in temporal reasoning. This implies that the current "medical" adaptation of LMMs primarily focuses on static image-text alignment, lacking the temporal pre-training necessary to understand disease progression over time.

4.2.3 Summary

In conclusion, excluding the most performant model GPT-5, the perception and reasoning capabilities of most LMMs regarding temporal issues remain at a relatively low level. The results highlight that while models can generate logically coherent medical text (High Completeness), they fundamentally lack the ability to precisely track visual changes (Low Accuracy). There is substantial room for improvement in integrating temporal dynamics into medical AI systems.

4.3 Ablation Studies

To investigate the factors influencing model performance, we conduct three ablation studies using the GPT-4o-mini evaluation metric.

4.3.1 Impact of Sequence Length

We analyze how the number of time points affects model performance. We subset the data to include sequences of lengths 2, 3, 4, and 5. As shown in Figure 3, we aim to determine if increased temporal complexity leads to performance degradation.

The experimental results presented in Figure 3 refute the hypothesis that increased temporal complexity leads to performance degradation. Contrary to expectations, we observe that extending the sequence length from 2 to 5 time points generally enhances model performance, particularly in terms of content coverage.

Specifically, our analysis yields two key observations:

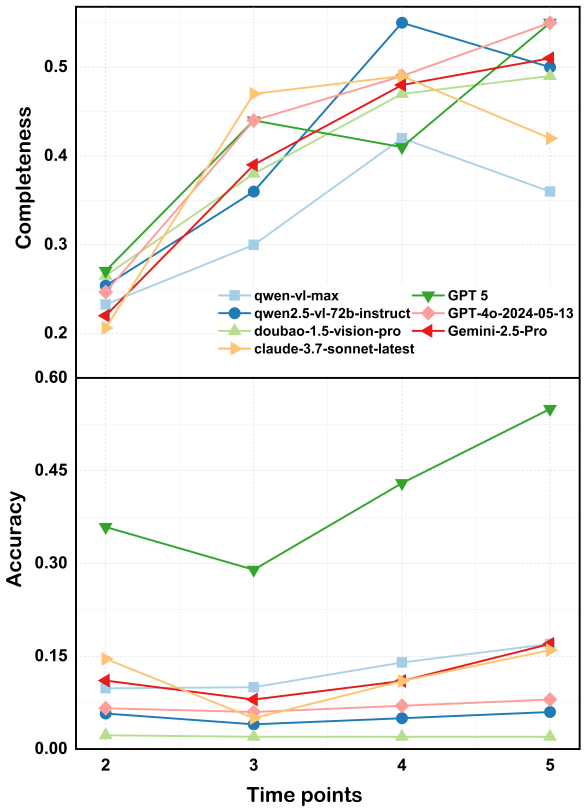


Figure 3: **Ablation Study on Sequence Length.** We evaluate model performance (Accuracy & Completeness) across varying numbers of historical visits. The scores indicate how well models maintain context as the temporal window expands.

Completeness improves with sequence length. As shown in the upper panel of Figure 3, completeness exhibits a distinct upward trend for nearly all evaluated models as the number of time points increases, which indicates that longer temporal sequences provide models with richer contextual information to support more comprehensive information retrieval and generation. However, this expanded temporal context also leads to more hallucinations, as the models' insufficient temporal reasoning ability fails to effectively discern and utilize the extended sequential information accurately. While some models (e.g., *Qwen-VL-Max*) show a slight completeness dip at the maximum sequence length ($T = 5$), this further reflects the challenge of temporal reasoning under extreme long-sequence settings, where the trade-off between information enrichment and hallucination risk becomes more prominent.

Accuracy exhibits stability or improvement. The lower panel demonstrates that Accuracy remains largely stable for most models across varying time points, indicating robust temporal reasoning capabilities. Notably, the top-performing model, *GPT 5*, shows a significant positive correlation between sequence length and accuracy, improving from approximately 34% at $T = 2$ to over 51% at $T = 5$. This indicates that state-of-the-art models can effectively leverage the additional dependencies found in longer sequences to refine their predictions, rather than suffering from hallucination or context loss.

In conclusion, within the scope of our experiments (sequences of length 2 to 5), increasing the number of time points does not hinder performance; instead, it tends to facilitate a more complete understanding of the visual-temporal data.

4.3.2 Robustness to Temporal Permutation

Current MLMMs often rely on the heuristic that "input order equals chronological order." To test true temporal understanding, we shuffle the input order of the images while explicitly providing correct timestamps for each image in the prompt. The specific timestamp will be generated by GPT-5-mini and it is required that he ensure it conforms to the clinical facts regarding the time interval and duration.If a model truly understands the temporal metadata, its performance should remain stable.

To decouple temporal reasoning from sequence input, we evaluate model performance under an input-shuffling protocol with explicit timestamp integration. In the prompt, it is emphasized that the model must understand the content of the image in the order of timestamps. The empirical results reveal a systemic deficiency in contemporary MLMMs; state-of-the-art models, including Claude-3.7-sonnet and GPT-4o, exhibit a performance collapse (Accuracy $< 15\%$) when positional cues are disrupted. This strongly validates the hypothesis that current architectures rely on superficial positional pattern matching rather than intrinsic semantic comprehension of temporal metadata. Conversely, **GPT-5** emerges as a significant outlier, achieving an accuracy of approximately 28%. While this demonstrates a qualitative leap in robustness against permutation, the modest absolute performance underscores that rigorous cross-modal temporal grounding remains a challenging frontier. Ultimately, the experiment confirms that most existing MLMMs fail to align textual temporal constraints with visual progressions without the aid of linear input structures.

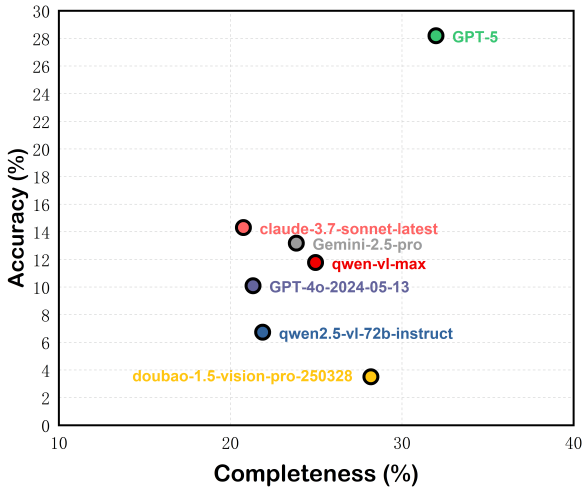


Figure 4: The performance of the models after the input order of the pictures is disrupted.

4.3.3 Modality Ablation: Multimodal vs. Text-Only

To verify the necessity of visual processing and multimodal integration, we compare the standard Multimodal integration approach against a text-only pipeline. In the text-only setting, we utilize **GPT-5** to generate detailed textual descriptions for each image in the sequence, as well as a summary of the series. The inputs to the text-only pipeline include the query, image descriptions, and image information summaries. In addition to generating image descriptions for each raw chest X-ray image, GPT-5 produces structured, temporally focused image information summaries by integrating sequential image metadata. These image information summaries are entirely distinct from the radiology reports in the MIMIC-CXR database: the latter are unstructured raw clinical notes, and crucially, they were not used as model inputs due to their strong correlation with the standard answers. These textual descriptions are then fed to the models to answer the questions, bypassing direct visual input. Figure 5 highlights the performance gap, demonstrating the value of visual features in capturing subtle pathological changes that text descriptions may miss.

Figure 5 presents a counter-intuitive finding: the **text-only pipeline outperforms standard multimodal integration** across the majority of models and metrics. Notably, strong baselines like GPT-5 and Qwen-VL-Max exhibit superior Accuracy and Completeness when relying solely on textual descriptions rather than direct visual inputs.

This performance gap highlights a critical bottleneck in the **visual-temporal encoding capabilities**

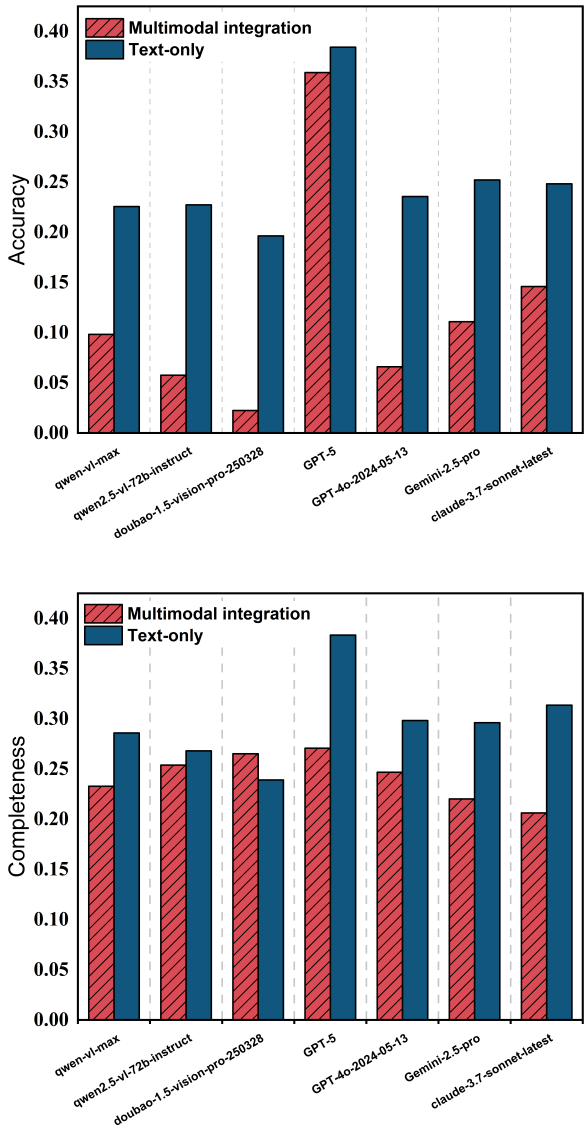


Figure 5: Performance comparison between the standard **Multimodal integration** and the **Text-only pipeline** across different models .

of current LMMs. The results suggest that raw spatiotemporal data introduces noise, which complicates the extraction of subtle pathological changes. In contrast, the text-only approach acts as a generic "semantic filter," where the captioning model abstracts complex visual dynamics into structured textual representations, reducing the cognitive load on the reasoning backbone. Consequently, we conclude that current performance is limited by the visual perception module's inability to effectively align long-context visual features with textual instructions, rather than a deficiency in reasoning logic.

5 Conclusion

This paper addresses the gap in evaluating the longitudinal capabilities of Medical Large Multi-

modal Models by introducing a physician-verified benchmark derived from MIMIC-CXR. Our experiments with state-of-the-art models, including GPT-5 and GPT-4o, reveal a significant disparity between static perception and temporal reasoning: while models excel at identifying isolated pathologies, they struggle to track disease progression accurately. Furthermore, our ablation studies demonstrate that text-only summaries are insufficient, underscoring the need for end-to-end multimodal processing. We hope that this benchmark catalyzes the development of medical AI agents that are visually proficient, temporally intelligent, and clinically reliable.

Limitations

This study acknowledges limitations primarily regarding data scope and evaluation methodology. First, our benchmark relies exclusively on 2D chest radiography from MIMIC-CXR; thus, the generalizability of our findings to 3D volumetric modalities (e.g., CT, MRI) and diverse anatomical regions remains unverified. Second, utilizing GPT-4o-mini as a proxy evaluator, while efficient, may introduce biases toward verbosity or overlook subtle domain-specific hallucinations detectable by board-certified radiologists. Future research should therefore aim to encompass broader imaging modalities and integrate rigorous human-in-the-loop validation.

Ethical Considerations

This study utilizes the publicly available MIMIC-CXR dataset, strictly adhering to PhysioNet's Data Use Agreement and HIPAA Safe Harbor provisions to ensure patient anonymity. We emphasize that the evaluated Large Multimodal Models are intended solely for research purposes and are not cleared for clinical deployment given risks such as hallucination; they should not support diagnostic decision-making without expert supervision. Furthermore, dataset construction involved AI assistance subject to manual verification, ensuring data quality without ethical concerns regarding crowd-sourced labor.

References

Anthropic. 2025. Claude 3.7 sonnet and claude code. https://www.anthropic.com/news/claude-3-7-sonnet. Accessed: 2025-2-25.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *ArXiv preprint*, abs/2309.16609.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *ArXiv preprint*, abs/2502.13923.

ByteDance. 2025. Doubao-1.5-pro. https://seed.bytedance.com/zh/special/doubao_1_5_pro. Accessed: 2025-1-22.

Junying Chen, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. 2024. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *ArXiv preprint*, abs/2406.19280.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *ArXiv preprint*, abs/2507.06261.

Hejie Cui, Alyssa Unell, Bowen Chen, Jason Alan Fries, Emily Alsentzer, Sanmi Koyejo, and Nigam H Shah. 2025. Timer: Temporal instruction modeling and evaluation for longitudinal clinical records. *npj Digital Medicine*, 8(1):577.

Amirata Ghorbani, David Ouyang, Abubakar Abid, Bryan He, Jonathan H Chen, Robert A Harrington, David H Liang, Euan A Ashley, and James Y Zou. 2020. Deep learning interpretation of echocardiograms. *NPJ digital medicine*, 3(1):10.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024. Omnimed-vqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22170–22183.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *ArXiv preprint*, abs/2410.21276.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng.

2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. 2020. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 5971–5984.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. *ArXiv preprint*, abs/2102.09542.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and William B Dolan. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6723–6737.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR.

Chinedu Innocent Nwoye and Nicolas Padoy. 2022. Data splits and metrics for method benchmarking on surgical action triplet datasets. *ArXiv preprint*, abs/2204.05235.

OpenAI. 2025. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/. Accessed: 2025-8-7.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2023. Gpt-4 technical report. *ArXiv preprint*, abs/2303.08774.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *ArXiv preprint*, abs/2507.05201.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *ArXiv preprint*, abs/2312.11805.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, and 1 others. 2025. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *ArXiv preprint*, abs/2506.07044.

Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. 2018. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of medical imaging*, 5(3):036501–036501.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, and 1 others. 2023. Baichuan 2: Open large-scale language models. *ArXiv preprint*, abs/2309.10305.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

A Evaluation Prompts

To ensure reproducibility, we provide the exact prompt templates used for our automated evaluation pipeline. We use GPT-4o-mini as the judge.

Table 3: The specific prompt template used for GPT-4o-mini based evaluation.

System Prompt:
You are an expert in electronic health records (EHR) analysis. Your task is to evaluate responses generated by AI models based on given instructions and EHR. You will assess the quality of the responses based on specific criteria, comparing them to provided reference answers (ground truth). Aim to be fair and balanced in your evaluation, recognizing both strengths and limitations in the model's response.

User Input Template:
`[Question]`
{question}
`[Model Answer]`
{model_answer}
`[Reference Answer]`
{reference_answer}
Evaluation Criteria:
1. Accuracy (0/1):
- Score 1 if the response is generally accurate and aligns with the key points in the reference answer, even if there are minor discrepancies or omissions.
- Score 0 only if the response contains significant factual errors or clearly misinterprets the EHR data.
2. Completeness (0/1):
- Score 1 if the response addresses the main aspects of the given instruction and covers the essential points present in the reference answer.
- Score 0 only if the response misses critical information or fails to address the core of the instruction.
Only output the two labeled lines.

B Evaluation Metrics

We have clearly defined Accuracy and Completeness in Section 3.4, with each metric focusing on distinct evaluation dimensions. Many questions in the experiments require binary Yes/No judgments, and an incorrect answer to the binary judgment alone results in a score of 0 for Accuracy, after which the evaluation of Completeness is still conducted independently. Completeness assesses the recall of critical clinical and temporal details required to answer the query, and it is not a measure of response length. A response is assigned a Completeness score of 1 only if it covers all core temporal changes and clinical inferences specified in the standard answer; a score of 0 is given if key details are omitted, regardless of the response's length. This distinction is reflected in the responses generated by LLMs, and we provide illustrative examples for the same question to demonstrate this in Table 4.

Table 4: Comparison of Answer Quality Metrics.

Standard Answer:
Yes, there is evidence of pneumonia. The second radiograph shows interval worsening of multifocal lung opacities, particularly in the right upper lobe, which suggests progression of presumed Cryptogenic Organizing Pneumonia or superimposed acute infectious pneumonia. The first radiograph indicated small consolidation that could represent pneumonia or atelectasis, while the third radiograph showed no evidence of pneumonia, indicating a change in the patient's condition over time.

High Accuracy, Low Completeness:
Yes, there is evidence of pneumonia. The second chest radiograph shows worsened multifocal lung opacities in the right upper lobe, which is a key sign of pneumonia progression.

High Completeness, Low Accuracy:
Yes, there is evidence of pneumonia. The first radiograph shows large bilateral lung consolidation typical of severe pneumonia, the second radiograph presents massive pleural effusion with diffuse lung opacities indicating pneumonia exacerbation, and the third radiograph still has mild opacification suggesting incomplete resolution of pneumonia. The patient's acute myelocytic leukemia background also increases the risk of bacterial pneumonia, which is consistent with the imaging changes of infectious pneumonia rather than Cryptogenic Organizing Pneumonia.

Low on Both Metrics:
No obvious pneumonia is found in the three radiographs, and the lung imaging changes are related to the patient's leukemia.

High on Both Metrics:
Yes, there is evidence of pneumonia. The first radiograph shows small lung consolidation that is suspicious for pneumonia or atelectasis; the second radiograph reveals interval worsening of multifocal lung opacities, especially in the right upper lobe, which indicates the progression of presumed Cryptogenic Organizing Pneumonia or superimposed acute infectious pneumonia; the third radiograph has no imaging signs of pneumonia, reflecting the dynamic change of the patient's condition over the three time points.

C Data Construction Pipeline

We define the specific inclusion and exclusion criteria for the image-text pairs from MIMIC-CXR as follows:

- **Inclusion:** Patients with at least two sequential chest X-ray studies; Reports containing explicit comparisons (e.g., "compared to previous study...").
- **Exclusion:** Images with severe artifacts; Reports with fewer than 10 words.

We used GPT-4o-mini to generate initial questions with guided prompts based on sequential

chest X-ray reports and image metadata from MIMIC-CXR. The model was strictly constrained to produce temporally relevant questions focusing on temporal perception/reasoning, with initial standard answers derived from explicit comparative descriptions in the reports. Templates for questions centered on temporal perception are relatively fixed and can be automatically generated with minor lexical adjustments, while those for temporal reasoning are more diverse.

The curation and verification process was conducted by three board-certified radiologists from a tertiary academic hospital, each with more than 5 years of specialized clinical experience in chest imaging interpretation. We provided the radiology team with a detailed, standardized guideline for screening and verification, which they followed to review all initial questions and answers to ensure strict clinical accuracy and temporal relevance of the dataset. Notably, the radiologists had no prior exposure to the chest X-ray images and clinical reports from the MIMIC-CXR database used in this work, eliminating potential biases from pre-existing familiarity with the data.

We provided a detailed guideline for screening and verification to the clinical radiologist team, who screened all initial questions and answers to ensure clinical accuracy and temporal relevance. A total of 1,247 initial QA pairs were generated in the first step of the pipeline. After screening and verification by the clinical radiologist team, 334 QA pairs failed to meet the three verification criteria (112 for Factuality, 98 for Clinical Relevance, and 124 for Temporal Dependency). All failed QA pairs were completely filtered out to ensure the high quality of the final dataset, which eliminates any low-quality samples that might affect model evaluation. The final QA pairs all comply with the three criteria upon verification. Regarding the usage of GPT-5-mini mentioned in Examples in ELTLM, the model is only used for timestamp generation and is not involved in the actual construction of QA pairs. Timestamps serve to distinguish input order from chronological order, simulating real clinical scenarios.

D Examples in ELTLM

We present additional examples of the Temporal Perception and Temporal Reasoning tasks in Figure 6.

E Evaluation Models

ELTLM focuses on seven general MLLMs, including GPT-5, GPT-4o, Gemini-2.5-pro, Claude-3.7-Sonnet, Qwen-VL-Max, Qwen2.5-VL and Doubao-1.5-Vision along with three specialized Med-MLLMs, i.e. Lingshu, HuatuoGPT-Vision and MedGemma, as shown in Table 6. We evaluate the GPT-5, GPT-4o, Gemini-2.5-pro, Claude-3.7-Sonnet, Qwen-VL-Max, Qwen2.5-VL and Doubao-1.5-Vision through the official API. We test the rest of the specialized Med-MLLMs through the released code and pre-trained model. The prompt for these MLLMs are listed in Table 7, where the prompts for five Med-MLLMs are recommended in their papers. Sometimes Med-MLLMs will refuse to generate responses because there are too many input images.

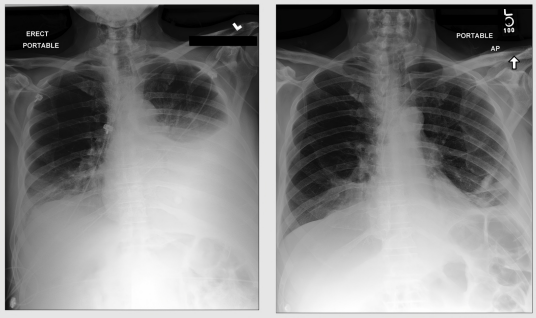
F Prompt for Timestamps Generation

Table 5 shows the prompts used in the Robustness to Temporal Permutation experiment, where we utilizes GPT-5-mini to generate explicit timestamps for each image.

Table 5: The specific prompt template used for generating timestamps for longitudinal image sequences.

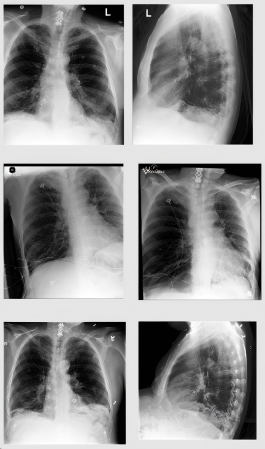
System Prompt:
You are an expert radiologist/data assistant. For the provided short question, and the ordered image descriptions (Image1..ImageN) and a temporal summary, generate a reasonable timestamp for each image in ISO format YYYY-MM-DD.

User Input Template:
`Question:`
{question}
`Image descriptions in time order:`
{formatted_image_list}
(e.g., Image1: Description... Image2: Description...)
`Temporal Summary:`
{temporal_summary}
Generation Constraints:
1. Output Format:
- Return ONLY a JSON array of strings (dates) or a JSON object mapping Image1..ImageN to dates.
2. Temporal Logic:
- There must be at least 1 day between adjacent image timestamps.
- Use ascending time order (older -> newer).
- Make timestamps reasonable given the Temporal Summary and question (e.g., matching 'days/weeks/months/years').
- Prefer dates within the past 10 years and not in the future.
3. Handling Missing Data:
- If an image description is empty, set the corresponding entry to an empty string.
`Output example (only JSON):`
"2023-01-01"



Question: What are the notable changes in the pleural effusions and lung aeration between the two time points?

Standard Answer: There has been a development of a large left pleural effusion filling the hemithorax in the first report, while the second report indicates that the left pleural catheter has been removed with a persistent small left pleural effusion and an improvement in the right pleural effusion. Additionally, lung volumes are slightly greater and there is improved aeration at both lung bases with residual atelectasis in the second report compared to the first.



Question: Given the findings from the three chest radiographs, does the patient have pneumonia, and what evidence supports your conclusion?

Standard Answer: Yes, the patient has left lower lobe pneumonia as indicated by increasing opacification at the left base in the first report, which is consistent with the consolidation seen on the CT examination. The subsequent reports show no evidence of pneumonia, but the initial finding confirms the diagnosis.

Figure 6: Sample cases from ELTLM. Top: A perception task requiring the model to identify changes in pleural effusion. Bottom: A reasoning task requiring the model to infer disease type.

Table 6: Comparison of different MLLMs included in our evaluation. The symbol "/" indicates that the detailed architecture is not publicly disclosed due to the proprietary nature of the model. Parameters for GPT-5, GPT-4o, Gemini-2.5-pro, and Claude-3.7-Sonnet are estimated to be greater than 100B .

MLLMs	**Vision Encoder**	**LLM**	**Parameters**
Proprietary General MLLMs			
GPT-5	/	/	>100B
GPT-4o	/	/	>100B
Gemini-2.5-pro	/	/	>100B
Claude-3.7-Sonnet	/	/	>100B
Qwen-VL-Max	/	/	>100B
Doubao-1.5-Vision	/	/	>100B
Open-Source General MLLMs			
Qwen2.5-VL	ViT	Qwen2.5	72B
Specialized Med-MLLMs			
Lingshu	ViT	Qwen2.5-VL	7B
HuatuoGPT-Vision	ViT-L/14	Qwen2	7B
MedGemma	MedSigLIP	Gemma 3	4B

Table 7: The prompts used to evaluate different MLLMs in our benchmark. General MLLMs utilize a detailed instruction to regulate output length and prevent hallucination, while specialized Med-MLLMs use a concise instruction to align with their training paradigms.

Models	**Prompt**
GPT-5, GPT-4o, Gemini-2.5-pro, Claude-3.7-Sonnet, Qwen-VL-Max, Qwen2.5-VL, Doubao-1.5-Vision	"You are a helpful radiologist assistant. Use only the provided images and the question to produce a clear, concise answer. Answer as briefly as possible. Answer within 30 words. Do NOT attempt to guess unseen clinical details. If uncertain, say you are uncertain and what would help (e.g., more imaging, clinical info)."
Lingshu, HuatuoGPT-Vision, MedGemma	"You are a helpful medical assistant. Please answer the following question based on the image provided."

G Impact of Image Description Models on Reasoning Performance

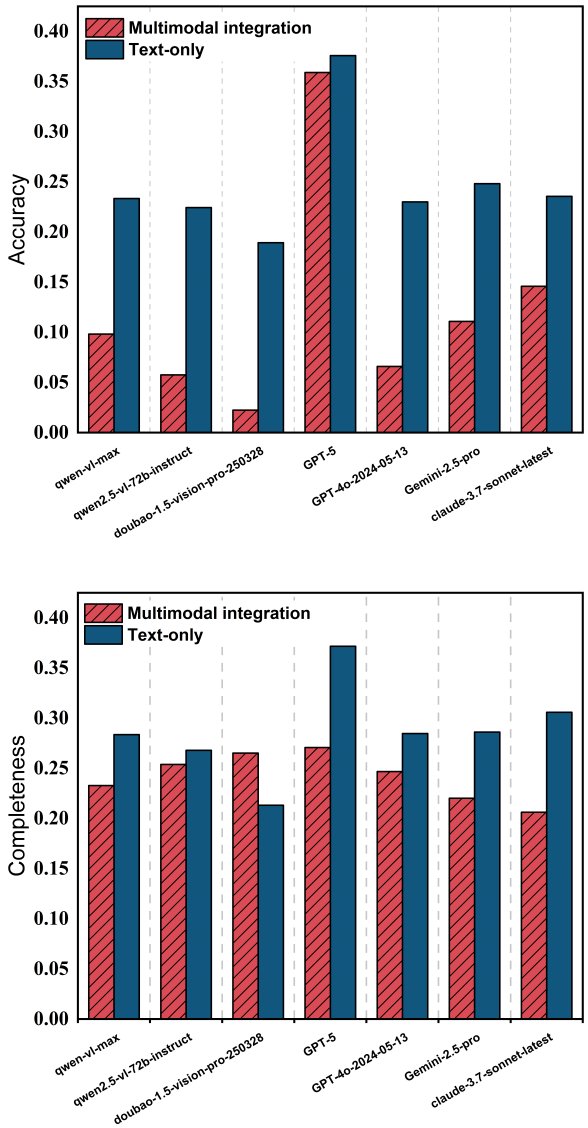


Figure 7: Performance comparison between the standard **Multimodal integration** and the **Text-only pipeline** across different models .

To verify the robustness of the text-only processing pipeline illustrated in Figure 5, we evaluated how the capacity of the upstream visual encoder influences downstream inference. Specifically, we conducted a comparative analysis using **GPT-5-mini** and the flagship **GPT-5** as the image captioning modules. As demonstrated in Figure 7, upgrading the prompt generator to **GPT-5** results in a marginal performance variation of approximately 5% across accuracy and completeness metrics. This observation provides compelling support for the **"semantic filtering"** hypothesis discussed in the main text. The fact that **GPT-5-mini** effectively performs this semantic abstraction and saturates the reasoning capabilities of downstream models suggests that the primary performance bottleneck is not the captioner's model size, but rather the inherent challenge within the visual perception module to effectively align long-context textual instructions with visual features.