Time-Grounded Clinical Prediction from Longitudinal Multimodal Patient Histories

Wanqi Yang[1,2], Ling Chen[2], Jianpeng Zhang[3], and Yutong Xie[1⋆]

[1] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
{wanqi.yang,yutong.xie}@mbzuai.ac.ae
[2] University of Technology Sydney, Sydney, NSW, Australia
Ling.Chen@uts.edu.au
[3] Zhejiang University, Hangzhou, China
jianpeng.zhang0@gmail.com

Abstract. Medical vision–language models (VLMs) have achieved strong performance on chest X-ray (CXR) interpretation, yet they are predominantly trained and evaluated under single-visit or limited pairwise comparison settings, with insufficient support for modeling richer longitudinal trajectories. We introduce MIMIC-MTE (Multi-Time Events), a comprehensive longitudinal dataset featuring multi-visit trajectories (3–10 events). Each event pairs a timestamped CXR and untruncated radiology reports, rich routine clinical tests, and explicit treatment plans. We define time-grounded clinical tasks spanning next-event forecasting (report generation and severity/progression selection), longitudinal question answering over specified time intervals, and treatment plan inference using temporally bracketing outcome evidence. To establish foundational baselines, we comprehensively benchmark a diverse suite of medical and general-purpose VLMs (e.g., LLaVA-Med, Hulu-Med, MedGemma, and Qwen2.5-VL) under zero-shot and supervised fine-tuning (SFT) settings. Our extensive evaluation reveals that off-the-shelf models exhibit substantial limitations in temporal grounding and multi-event evidence aggregation. Conversely, SFT on MIMIC-MTE significantly unlocks their longitudinal reasoning capabilities, establishing a robust benchmark to drive future research in time-aware clinical foundation models. The data preprocessing pipeline is available on MIMIC-MTE.

Keywords: Longitudinal clinical reasoning · Multimodal · VLMs.

1 Introduction

Recent advancements in medical vision-language models (VLMs) have demonstrated remarkable efficacy in chest X-ray interpretation [3,11,23,24] such as visual question answering [6,11,21] and radiology report generation or drafting [3,16,28]. However, these models are predominantly trained and evaluated on isolated single-visit inputs, under-representing the longitudinal context routinely used in clinical decision-making. A fundamental motivation for our work

⋆ Corresponding author: yutong.xie@mbzuai.ac.ae

stems from that chest and lung pathologies are highly non-monotonic. Conditions such as pneumonia, pulmonary edema, or pleural effusion frequently fluctuate, displaying irregular cycles of exacerbation, remission, and relapse [1,4,13,18,20]. Consequently, interpreting a prior or current chest radiograph in isolation is clinically insufficient. Accurate assessment requires a comprehensive long-horizon patient history to capture these undulating disease trajectories.

Despite growing interest in longitudinal modeling, existing "temporal" CXR resources predominantly emphasize short-horizon settings (e.g., pairwise change detection or consecutive prior-current comparisons) or narrow objectives such as report generation with limited history [30]. In parallel, multimodal EHR datasets often provide richer physiologic measurements but are typically aligned within a single admission window and do not capture multi-visit trajectories grounded in sequential imaging [17]. Consequently, core capabilities needed for real-world longitudinal care, such as time grounding, multi-visit evidence aggregation, and decision-relevant reasoning under irregular follow-up, remain insufficiently benchmarked and under-trained in current medical foundation models.

To address this critical gap, we introduce **MIMIC-MTE** (Multi-Time Events), a CXR-centered multimodal longitudinal dataset organized as irregular multi-event patient histories. Moving beyond pairwise deltas, each sample in MIMIC-MTE captures 3 to 10 clinical events for a specific patient and disease course. Crucially, each event tightly couples the chest radiograph and untruncated radiology reports, rich routine clinical tests (e.g., labs, vitals), and explicit interventional treatment plans. Our dataset reveals that over 32% of these trajectories exhibit explicit state inversions (both improving and worsening), validating the absolute necessity of multi-visit modeling. Built on this dataset, we define a suite of time-grounded clinical tasks that reflect realistic longitudinal workflows: (i) next-event forecasting (report generation and severity/progression selection), (ii) longitudinal question answering over explicit time intervals, and (iii) treatment plan inference using prior history and temporally bracketing outcome evidence.

We conduct a comprehensive benchmarking study on MIMIC-CXR-MTE. We first evaluate a diverse suite of state-of-the-art medical and general-purpose VLMs (including LLaVA-Med [11], Hulu-Med-7B [8], MedGemma-1.5-4B [19], and Qwen2.5-VL variants [25]) under zero-shot settings. Furthermore, to establish strong baselines, we perform supervised fine-tuning (SFT) on two powerful open-weight backbones (Qwen2.5-VL-7B and MedGemma-1.5-4B) using our formulated longitudinal tasks. Our extensive experiments demonstrate that while current off-the-shelf models struggle significantly with temporal grounding, non-monotonic course understanding, and multi-event evidence integration, direct instruction tuning on our dataset effectively bridges this gap.

2 Method

2.1 Dataset Collection and Construction

Data sources and multimodal integration. MIMIC-MTE is constructed by integrating complementary clinical signals from four robust resources. Chest Im-

aGenome (silver) [26] serves as the foundational scaffold, providing study-level disease/anatomical finding annotations and radiology report text. We source chest radiographs from the established MIMIC-CXR-JPG dataset [10]. MIMIC-IV [9] augments each imaging event with comprehensive treatment context (medications and procedures) and a diverse set of routine clinical tests, including laboratory panels, blood chemistry, coagulation profiles, blood gas measurements, cardiac markers, and vital signs. MIMIC-IV-ECG [5] supplements synchronous 12-lead ECG interpretations when available. All sources are rigorously aligned across modalities using a composite key comprising patient identity, disease category, event identifier, and event timestamp.

Longitudinal event history construction. To construct multi-time event histories, we first restrict the concept space to progression-relevant entities by retaining only ImaGenome entries with category $\in \{\texttt{disease}, \texttt{finding}\}$. We then group records by `(patient_id, study_type)`, where `study_type` denotes the target disease/anatomical finding. For each patient–disease pair (p, d), we retrieve all associated events and sort them chronologically by their timestamps:

$$\mathcal{H}_{p,d} = \{e_1, e_2, \ldots, e_T\}, \quad \text{time}(e_1) \leq \text{time}(e_2) \leq \cdots \leq \text{time}(e_T). \tag{1}$$

Each event explicitly stores a date field (`time`), which defines temporal ordering and supports time-aware learning objectives such as forecasting and interval-based querying. We enforce longitudinal constraints to emphasize multi-visit progression: each sample contains $3 \leq T \leq 10$ events and spans at least three distinct dates; multiple events on the same date are allowed to reflect repeated imaging and reassessment in real practice. Each event e_i represents a clinical snapshot at time t_i and contains: (i) identifiers (e.g., `eventID` and image identifier `gid`); (ii) radiology evidence as free-text report (`report`); (iii) progression descriptors (`severity` and `comparison`, when available); (iv) `treatment` describing medications/procedures; and (v) routine examinations (`clinical_tests`) spanning labs, chemistry, coagulation, blood gas, cardiac markers, vitals, and ECG-derived context when present (see middle part of Fig. 1). This design yields a time-indexed multimodal record suitable for learning both localized radiographic cues and longitudinal clinical context across irregular time gaps.

Data split and leakage prevention. We inherit the official train/validation/test splits defined by Chest ImaGenome to ensure consistent comparability with prior benchmarks. To strictly prevent data leakage and ensure reliable downstream evaluation, we implement these verification protocols: 1) Sample-level Isolation: We compute hash values for all JSON samples to verify a strict zero-intersection across splits; 2) Patient-level Separation: We guarantee mutually exclusive `patient_id` sets between the training and testing sets.

2.2 Tasks for Longitudinal Event Modeling

To systematically evaluate the capability of vision-language models in reasoning over multi-visit histories, we formulate a suite of temporal modeling tasks based on the established longitudinal trajectory $\mathcal{H}_{p,d} = \{e_1, e_2, \ldots, e_T\}$. We design

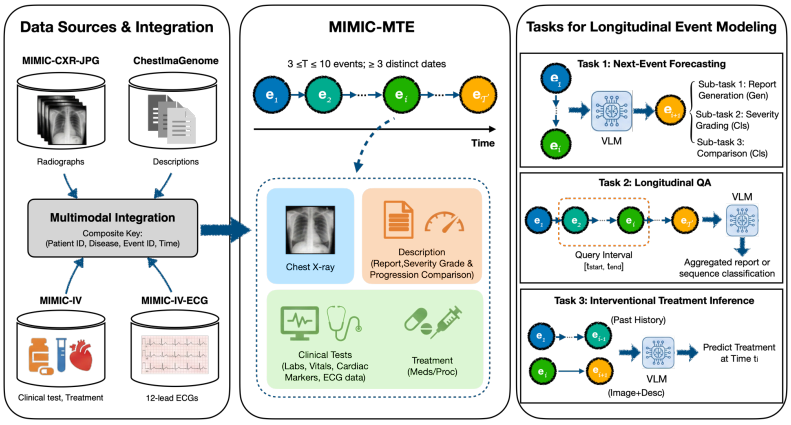

Fig. 1. Framework overview of the MIMIC-MTE dataset construction and associated temporal modeling tasks.

three primary clinical objectives, which inherently encompass distinct generative and discriminative sub-tasks.

Task 1: Next-Event Forecasting. Given a patient's histories up to the i-th event, $\{e_1, \ldots, e_i\}$, the model must forecast the clinical status of the upcoming $(i+1)$-th event, e_{i+1}. This task family assesses whether a model can extrapolate near-future status from irregularly spaced, multi-visit multimodal evidence. The objective is evaluated through three sub-tasks: 1) Report Generation: Autoregressively generating the comprehensive radiological description for e_{i+1}. 2) Severity Grading: A discriminative task predicting the severity level of e_{i+1} from predefined categorical terms. 3) Progression Comparison: A discriminative task classifying changes in state at e_{i+1} relative to the previous visit e_i.

Task 2: Longitudinal QA. In practice, physicians often query a patient's record over a specific time. This task explicitly probes temporal grounding beyond short-horizon heuristics. The model is provided with the full event history $\mathcal{H}_{p,d}$ (with target reports, severities, and comparisons masked) and a time-bounded query specifying an interval $[t_{start}, t_{end}]$. The model must temporally localize all relevant events falling within this window and aggregate their findings. Depending on the query formulation, the model must output either a synthesized narrative report covering multiple visits, or a sequence of severity/comparison classifications tracking the disease fluctuation during that exact timeframe.

Task 3: Interventional Treatment Inference. This task models the clinical decision-making process by inferring the treatment administered based on state transitions. The model receives the patient's prior history $\{e_1, \ldots, e_{i-1}\}$, the clinical state at the current visit e_i (specifically, the image and image's description), and the subsequent outcome state at e_{i+1} (image and image's description). The objective is to accurately predict the interventional treatment plan implemented at time t_i. This setting reflects realistic care where treatment decisions are informed by prior trajectory and present status, and where the subsequent

Table 1. Comprehensive statistics of the MIMIC-MTE dataset.

Data Scale	**Value**	**Progression Labels**	**Value**
Total patient timelines	36,928	Severity: Mild / Hedge	44.7% / 21.4%
Unique patients	14,203	Severity: Mod. / Severe	18.0% / 15.8%
Total multimodal events	169,412	Comparison: No change	52.9%
Events per timeline (mean / range)	4.59 (3–10)	Comparison: Worsened / Improved	31.2% / 15.9%
Unique study types (diseases)	50	Non-monotonic timelines	32.29%
Temporal Statistics (Days)	**Value**	**Multimodal Completeness**	**Empty (%)**
Timeline duration median (IQR)	103 (12–549)	Radiology Report	0 (0.0%)
Timeline duration mean (min/max)	337.4 (1 / 1,846)	Severity Grade	74747 (44.1%)
Adjacent gap Δt_i median (IQR)	5 (1–61)	Progression Comparison	76304 (45.0%)
Adjacent gap Δt_i mean (min/max)	94.0 (0 / 1,844)	All Clinical Tests	20,385 (12.0%)
Timelines with 0-day gaps	7,223	Treatment	32,934 (19.4%)
Task Evaluation Splits	**Train Set**	**Validation Set**	**Test Set**
Task 1: Next-Event Forecasting	35,998	273	657
Task 2: Time-Bounded Longitudinal QA	30,957	241	551
Task 3: Interventional Treatment Inference	35,998	273	657

Table 2. Comparison to our MIMIC-MTE and existing longitudinal chest X-ray works. Here, "Tem." denotes Temporal, "Clin." refers to routine Clinical tests, "Treat." indicates Interventional Treatments, and "longi. QA" stands for Longitudinal QA.

Benchmarks	**Tem. Granularity**	**Visits**	**Image**	**Clin. Tests**	**Treat.**	**Evaluation Paradigms & Output Formats**
MIMIC-Diff-VQA (2023) [6]	Pairwise	2	✓	✗	✗	Diff-VQA (short template answers)
MS-CXR-T (2023) [2]	Pairwise	2	✓	✗	✗	Progression classification / similarity
TempA-VLP [27] (2025)	Pairwise	2	✓	✗	✗	Temporal grounding / representation
Longitudinal-MIMIC (2023) [30]	Pairwise	2	✓	✗	✗	Pairwise report pre-fill / generation
HERGen/MLRG (2024) [22]	Multi-visit history	>2	✓	✗	✗	Report generation
SYMILE-MIMIC (2025) [17]	Single admission window	1	✓	✓	✗	Cross-modal retrieval/alignment
MultiTimeSurv (2025) [29]	Multi-time/Irregular	>2	Optional	✓	✗	Survival/risk prediction (numeric curves)
EHR2Path (2025) [15]	Long-horizon trajectory	>2	✗	✓	✓	Structured sequence simulation (Tokenized)
MIMIC-MTE	Multi-time/Irregular	>2	✓	✓	✓	Multi-Task (forecasting, longi. QA, Treat.)

event provides weak outcome evidence. The task thus evaluates whether models can map temporally evidence to clinically plausible therapeutic actions.

All targets are deterministically extracted from Chest ImaGenome and MIMIC-IV records (no LLM-generated labels), and we use diverse prompt templates to reduce template overfitting.

2.3 Dataset Statistics

MIMIC-MTE contains 36,928 patient timelines from 14,203 patients with 169,412 multimodal events (sequence length 3–10; mean 4.59). The dataset captures long-horizon and highly irregular follow-up: timeline duration has a median of 103 days (IQR 12–549), and adjacent gaps have a median of 5 days (IQR 1–61), including 7,223 timelines with same-day repeats. Progression comparison highlight non-monotonicity: among timelines with valid comparisons, 32.29% contain both "improved" and "worsened" events. Multimodal completeness and task splits are summarized in Table 1.

2.4 Comparison to Existing Longitudinal CXR Works

Prior work has explored temporal modeling in medical vision-language learning, yet existing benchmarks predominantly capture short-horizon evaluations [2,6,27],

single-task applications [22,30], and structured predictions lacking open-ended visual reasoning [15,17,29]. We contextualize MIMIC-MTE against representative resources across three categories (Table 2). Compared with prior datasets, MIMIC-MTE offers broader modality coverage, supports a larger number of longitudinal visits per patient, and encompasses a more comprehensive suite of evaluation tasks for longitudinal event modeling. Together, these characteristics enable systematic assessment of time-grounded reasoning, multi-visit evidence aggregation, and clinically relevant decision-making in realistic care trajectories.

3 Experiments and Results

3.1 Experimental Setup

Models. We evaluate a diverse suite of medical and general-purpose vision-language models, including LLaVA-Med [11], Hulu-Med-7B [8], MedGemma-1.5-4B [19], and Qwen2.5-VL (3B/7B) [25], to establish foundational baseline capabilities for long-horizon clinical reasoning under zero-shot settings. To demonstrate the utility of our dataset and establish strong longitudinal baselines, we also perform SFT on two powerful open-weight backbones: Qwen2.5-VL-7B and MedGemma-1.5-4B.

We use AdamW with a learning rate of 1e-4 (and 1e-5 for the multimodal merger), weight decay 0.1, and a cosine scheduler with 3% warmup. Training is conducted for 1 epoch with an effective batch size of 16 on 4 A40 GPUs. We perform LoRA fine-tuning (rank=32, $\alpha = 64$, dropout=0.05), with the base model frozen and only the adapters and multimodal merger updated. Image resolution is constrained between $256 \times 28 \times 28$ and $768 \times 28 \times 28$. Training uses bfloat16 and DeepSpeed ZeRO-3.

Evaluation metrics. We evaluate all tasks using metrics matched to the output format. The missing labels reported in Table 1 are only a dataset completeness statistic; they are not used as evaluation targets. (i) For the generative variants of Task1 (Next-Event Forecasting) and Task2 (Longitudinal QA), we report BLEU-avg [14] (average of BLEU-1 through BLEU-4)) and ROUGE-L [12] ((F1 and Precision)) for linguistic fluency, together with RadGraph-F1 [7] (overall, Entity-F1, Relation-F1) to assess factual correctness. (ii) For the discriminative/selection variants of Task1 and Task2, we report the Macro-F1 score to robustly account for the imbalanced distribution of clinical states. (iii) For Task3 (Interventional Treatment Inference), we evaluate with BLEU-avg, ROUGE-L, and entity-level metrics (micro-F1, precision, recall) to reflect treatment accuracy. We also report Parse Rate, defined as the percentage of model outputs successfully parsed by our treatment entity extraction pipeline. (iv) Finally, for additional discriminative benchmarks (MS-CXR-T and Medical-Diff-VQA), we report macro-F1 for consistency.

3.2 Main Results

Table 3 summarizes the performance of representative medical and general-purpose VLMs on three longitudinal tasks. The results reveal a consistent pat-

Table 3. Performance of state-of-the-art vision-language models on the MIMIC-MTE test set and existing pairwise longitudinal benchmarks.

Task	Model (Setting)	BL-avg	R-L-F1	R-L-P	RadG-F1	RadG-E-F1	RadG-R-F1	Macro-F1
Next-Event Forecasting	Llava-Med-zs	2.73	3.63	4.64	1.15	1.97	1.61	2.34
	Hulu-Med-7B-zs	7.33	12.00	12.38	8.84	13.86	12.06	26.94
	Medgemma1.5-4B-zs	6.55	10.50	7.75	9.83	15.43	13.49	24.59
	Qwen2.5-VL-3B-zs	6.04	9.68	9.49	7.63	11.58	10.20	16.00
	Qwen2.5-VL-7B-zs	8.20	13.06	9.75	11.22	17.73	15.18	25.24
	Medgemma1.5-4B-sft	**14.86**	**23.58**	33.17	**17.32**	**22.89**	**20.98**	58.06
	Qwen2.5-VL-7B-sft	12.42	22.11	**35.81**	16.79	21.83	19.81	**59.78**
Longitudinal QA	Llava-Med-zs	1.53	3.73	6.94	0.24	0.43	0.42	3.53
	Hulu-Med-7B-zs	0.46	4.35	4.36	0.98	1.28	0.90	44.15
	Medgemma1.5-4B-zs	4.99	10.34	13.07	5.57	8.97	8.01	25.10
	Qwen2.5-VL-3B-zs	0.18	1.78	37.55	0.08	0.38	0.09	13.71
	Qwen2.5-VL-7B-zs	1.05	5.80	34.86	0.50	1.13	0.47	38.59
	Medgemma1.5-4B-sft	6.48	**22.01**	**49.01**	**19.40**	**26.02**	**24.18**	62.29
	Qwen2.5-VL-7B-sft	**8.65**	21.02	34.57	16.57	21.27	19.83	**65.09**
		BL-avg	**R-L-F1**	**R-L-P**	**Ent-Micro-F1**	**Ent-P**	**Ent-R**	**Parse-Rate**
Interventional Treatment Inference	Llava-Med-zs	2.47	1.95	2.55	0.00	0.00	0.00	0.00
	Hulu-Med-7B-zs	18.81	26.52	29.22	23.89	27.59	21.07	65.34
	Medgemma1.5-4B-zs	2.15	2.76	1.96	8.04	27.71	4.70	16.52
	Qwen2.5-VL-3B-zs	10.44	13.87	14.81	15.10	26.99	10.48	34.48
	Qwen2.5-VL-7B-zs	11.10	14.88	14.34	9.60	28.84	5.75	19.42
	Medgemma1.5-4B-sft	14.45	17.27	16.08	19.20	25.11	15.54	45.37
	Qwen2.5-VL-7B-sft	**30.43**	**40.83**	**39.62**	**31.61**	**28.77**	**35.08**	**100.00**

Pairwise Longitudinal Benchmarks

Task	Model	Macro-F1	Task	Model	Macro-F1
MS-CXR-T	Llava-Med-zs	4.17	Med-Diff-VQA	Llava-Med-zs	7.33
	Hulu-Med-7B-zs	41.82		Hulu-Med-7B-zs	34.17
	Medgemma1.5-4B-zs	39.93		Medgemma1.5-4B-zs	36.33
	Qwen2.5-VL-3B-zs	23.32		Qwen2.5-VL-3B-zs	27.67
	Qwen2.5-VL-7B-zs	30.03		Qwen2.5-VL-7B-zs	33.33
	Medgemma1.5-4B-sft	42.78		Medgemma1.5-4B-sft	**40.17**
	Qwen2.5-VL-7B-sft	**43.55**		Qwen2.5-VL-7B-sft	38.00

tern: off-the-shelf models exhibit a pronounced gap between "single-visit radiology competence" and "multi-visit longitudinal reasoning ability", while direct instruction tuning on MIMIC-MTE substantially narrows this gap.

The Catastrophic Failure of Medical VLMs in Temporal Grounding. A primary finding from our zero-shot evaluation is that while off-the-shelf medical VLMs (e.g., LLaVA-Med, Hulu-Med-7B) possess strong single-image perceptual capabilities, they experience profound degradation when subjected to longitudinal constraints. This is most starkly illustrated in Task 2 (Longitudinal QA). When tasked with localizing and aggregating findings within a specific time bracket $[t_{start}, t_{end}]$, zero-shot models largely fail to retrieve clinically factual evidence, yielding near-zero RadGraph-F1 scores.

This phenomenon suggests that, without explicit temporal alignment, base models tend to generate generic, single-visit radiology narratives that entirely ignore interval constraints. Even Hulu-Med-7B, which achieves a moderate Macro-F1 (44.15) on categorical selection in Task 2, completely collapses on generative factuality (RadGraph-F1 of 0.98), indicating that it guesses classification labels without genuinely grounding the reasoning in the longitudinal visual evidence.

Instruction Tuning Unlocks Multi-Visit Reasoning. Applying SFT to the MIMIC-MTE corpus yields transformative improvements across all architectures, demonstrating that the models possess the requisite visual encoders but lack alignment for multi-event integration.

In Task 1 (Next-Event Forecasting), SFT drives a massive leap in clinical factuality. MedGemma1.5-4B-sft improves its RadGraph-F1 from 9.83 to 17.32, while Qwen2.5-VL-7B-sft more than doubles its Macro-F1 (25.24 to 59.78) for progression classification. These substantial gains validate our core hypothesis: interpreting non-monotonic trajectories (where conditions frequently worsen and improve) requires models to be explicitly trained on explicit state inversions and irregular temporal gaps, rather than pairwise deltas. The strong performance of the fine-tuned Qwen2.5-VL-7B backbone particularly underscores the efficiency of modern dense VLMs in capturing complex cross-modal interactions when properly supervised with structured longitudinal histories.

The Complexity of Interventional Treatment Inference. Task 3 (Interventional Treatment Inference) is the most rigorous test of clinical reasoning, requiring the model to map historical context and prior-to-current-state transitions to specific therapeutic actions. Zero-shot models struggle drastically with both clinical accuracy and basic instruction following in this domain. LLaVA-Med fails entirely (0.00 across all metrics), and most baselines exhibit low Parse-Rates, demonstrating an inability to structure complex treatment rationales.

Remarkably, Qwen2.5-VL-7B-sft establishes a definitive new state-of-the-art for this task. It achieves a flawless 100% Parse-Rate and drastically outperforms all other models in Entity-Micro-F1 (31.61) and Entity Recall (35.08). MedGemma1.5-4B-sft also improves, but remains clearly behind Qwen-SFT on both overlap and entity recall, suggesting that for decision-like text outputs, larger/generalist multimodal backbones may benefit more from task-specific instruction supervision. Interestingly, Hulu-Med-7B-zs exhibits unusually strong zero-shot performance here (Ent-Micro-F1 of 23.89), likely due to heavy exposure to MIMIC discharge summaries during its pre-training. However, the superior performance of the SFT models ultimately confirms that explicit multi-visit trajectory modeling is a prerequisite for reliable clinical decision support.

Generalization to Standard Pairwise Benchmarks. To verify that our longitudinal instruction tuning transfers effectively to conventional temporal settings, we evaluate our models on two established external benchmarks: MS-CXR-T (Temporal progression classification) and Med-Diff-VQA (pairwise difference reasoning). While our MIMIC-MTE dataset emphasizes complex, multi-visit trajectories (3–10 events), models fine-tuned on our corpus demonstrate superior generalization to these simpler two-time-point tasks. For instance, Qwen2.5-VL-7B-sft and MedGemma1.5-4B-sft substantial improvements over its zero-shot counterpart, achieving a Macro-F1 of 43.55 on MS-CXR-T and 40.17 on Med-Diff-VQA respectively. Notably, this fine-tuning paradigm enables general-purpose backbones to consistently outperform medical-specific zero-shot baselines, proving that robust multi-visit alignment inherently strengthens pairwise temporal grounding.

4 Conclusion

In this paper, we address a critical gap in medical vision-language modeling: the inability of current models to reason over non-monotonic, long-horizon clinical trajectories. We introduce MIMIC-MTE, a comprehensive multi-visit dataset that couples longitudinal chest radiographs with rich multimodal clinical context across irregular temporal gaps. Based on this dataset, we formulate a suite of tasks aligned with real-world longitudinal workflows, including next-event forecasting, longitudinal QA, and interventional treatment inference. Extensive experiments demonstrate that while off-the-shelf medical VLMs struggle significantly with multi-visit reasoning, direct instruction tuning on our dataset effectively bridges this gap and establishes strong state-of-the-art baselines. Furthermore, these fine-tuned models exhibit superior generalization capabilities when transferred to standard pairwise longitudinal benchmarks. In future work, we plan to expand MIMIC-MTE toward other institutions, care settings, or imaging modalities, and explore more robust temporal reasoning architectures to better handle dense follow-ups and subtle progression boundaries.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Addala, D.N., Kanellakis, N.I., Bedawi, E.O., Dong, T., Rahman, N.M.: Malignant pleural effusion: Updates in diagnosis, management and current challenges. Frontiers in Oncology **12**, 1053574 (2022)
2. Bannur, S., Hyland, S., Liu, Q., Pérez-García, F., Ilse, M., Coelho de Castro, D., Boecking, B., Sharma, H., Bouzid, K., Schwaighofer, A., Wetscherek, M.T., Richardson, H., Naumann, T., Alvarez Valle, J., Oktay, O.: MS-CXR-T: Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing. PhysioNet (Mar 2023). `https://doi.org/10.13026/pg10-j984`, `https://doi.org/10.13026/pg10-j984`, version 1.0.0
3. Chen, Z., Varma, M., Delbrouck, J.B., Paschali, M., Blankemeier, L., Van Veen, D., Valanarasu, J.M.J., Youssef, A., Cohen, J.P., Reis, E.P., et al.: Chexagent: Towards a foundation model for chest x-ray interpretation. In: AAAI 2024 Spring Symposium on Clinical Foundation Models (2024)
4. Gilbert, C.R., Porcel, J.M.: Management of recurrent transudative pleural effusions: can we reduce unnecessary interventions? (2022)
5. Gow, B., Pollard, T., Nathanson, L.A., Johnson, A., Moody, B., Fernandes, C., Greenbaum, N., Waks, J.W., Eslami, P., Carbonati, T., et al.: Mimic-iv-ecg: Diagnostic electrocardiogram matched subset. Type: dataset (2023)
6. Hu, X., Gu, L., An, Q., Zhang, M., Liu, L., Kobayashi, K., Harada, T., Summers, R.M., Zhu, Y.: Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 4156–4165 (2023)

7. Jain, S., Agrawal, A., Saporta, A., Truong, S.Q., Duong, D.N., Bui, T., Chambon, P., Zhang, Y., Lungren, M.P., Ng, A.Y., et al.: Radgraph: Extracting clinical entities and relations from radiology reports. arXiv preprint arXiv:2106.14463 (2021)
8. Jiang, S., Wang, Y., Song, S., Hu, T., Zhou, C., Pu, B., Zhang, Y., Yang, Z., Feng, Y., Zhou, J.T., et al.: Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. arXiv preprint arXiv:2510.08668 (2025)
9. Johnson, A.E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Hao, S., Moody, B., Gow, B., et al.: Mimic-iv, a freely accessible electronic health record dataset. Scientific data **10**(1), 1 (2023)
10. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
11. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
12. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
13. Lindow, T., Quadrelli, S., Ugander, M.: Noninvasive imaging methods for quantification of pulmonary edema and congestion: a systematic review. Cardiovascular Imaging **16**(11), 1469–1484 (2023)
14. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
15. Pellegrini, C., Özsoy, E., Bani-Harouni, D., Keicher, M., Navab, N.: Ehr2path: Scalable modeling of longitudinal health trajectories with llms
16. Pellegrini, C., Özsoy, E., Busam, B., Wiestler, B., Navab, N., Keicher, M.: Radialog: Large vision-language models for x-ray reporting and dialog-driven assistance. In: Medical Imaging with Deep Learning (2025)
17. Saporta, A., Puli, A.M., Goldstein, M., Ranganath, R.: Symile-MIMIC: a multimodal clinical dataset of chest X-rays, electrocardiograms, and blood labs from MIMIC-IV. PhysioNet (Jan 2025). `https://doi.org/10.13026/3vvj-s428`, `https://doi.org/10.13026/3vvj-s428`, version 1.0.0
18. Schulz, D., Rasch, S., Heilmaier, M., Abbassi, R., Poszler, A., Ulrich, J., Steinhardt, M., Kaissis, G.A., Schmid, R.M., Braren, R., et al.: A deep learning model enables accurate prediction and quantification of pulmonary edema from chest x-rays. Critical Care **27**(1), 201 (2023)
19. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
20. Takada, K., Ogawa, K., Miyamoto, A., Nakahama, H., Moriguchi, S., Murase, K., Hanada, S., Takaya, H., Tamaoka, M., Takai, D.: Risk factors and interventions for developing recurrent pneumonia in older adults. ERJ Open Research **9**(3) (2023)
21. Thawakar, O.C., Shaker, A.M., Mullappilly, S.S., Cholakkal, H., Anwer, R.M., Khan, S., Laaksonen, J., Khan, F.: Xraygpt: Chest radiographs summarization using large medical vision-language models. In: Proceedings of the 23rd workshop on biomedical natural language processing. pp. 440–448 (2024)
22. Wang, F., Du, S., Yu, L.: Hergen: Elevating radiology report generation with longitudinal data. In: European Conference on Computer Vision. pp. 183–200. Springer (2024)

23. Wu, C., Zhang, X., Zhang, Y., Hui, H., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. Nature Communications **16**(1), 7866 (2025)
24. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 21372–21383 (2023)
25. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
26. Wu, J.T., Agu, N.N., Lourentzou, I., Sharma, A., Paguio, J.A., Yao, J.S., Dee, E.C., Mitchell, W., Kashyap, S., Giovannini, A., et al.: Chest imagenome dataset for clinical reasoning. arXiv preprint arXiv:2108.00316 (2021)
27. Yang, Z., Shen, L.: Tempa-vlp: Temporal-aware vision-language pretraining for longitudinal exploration in chest x-ray image. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4625–4634. IEEE (2025)
28. Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E.P., Fonseca, E.K.U.N., Lee, H.M.H., Abad, Z.S.H., Ng, A.Y., et al.: Evaluating progress in automatic chest x-ray radiology report generation. Patterns **4**(9) (2023)
29. Zeiser, F.A., Zeiser, M.H., de Oliveira Ramos, G., da Costa, C.A.: Multitimesurv: Temporal multimodal networks for dynamic survival analysis with longitudinal data
30. Zhu, Q., Mathai, T.S., Mukherjee, P., Peng, Y., Summers, R.M., Lu, Z.: Utilizing longitudinal chest x-rays and reports to pre-fill radiology reports. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 189–198. Springer (2023)