HIST-AID: Leveraging Historical Patient Reports for Enhanced Multi-Modal Automatic Diagnosis

Haoxu Huang[1,3] HH2740@NYU.EDU
Cem M. Deniz[2] CEM.DENIZ@NYULANGONE.ORG
Kyunghyun Cho[1,3,4,5] KYUNGHYUN.CHO@NYU.EDU
Sumit Chopra[1,2,3] SUMIT.CHOPRA@NYULANGONE.ORG
Divyam Madaan[1] DIVYAM.MADAAN@NYU.EDU
[1] *Courant Institute of Mathematical Sciences, New York University*
[2] *Department of Radiology, New York University Grossman School of Medicine*
[3] *Center of Data Science, New York University*
[4] *Prescient Design, Genentech*
[5] *CIFAR LMB*

Abstract

Chest X-ray imaging is a widely accessible and non-invasive diagnostic tool for detecting thoracic abnormalities. While numerous AI models assist radiologists in interpreting these images, most overlook patients' historical data. To bridge this gap, we introduce *Temporal MIMIC* dataset, which integrates five years of patient history, including radiographic scans and reports from MIMIC-CXR and MIMIC-IV, encompassing 12,221 patients and thirteen pathologies. Building on this, we present *HIST-AID*, a framework that enhances automatic diagnostic accuracy using historical reports. HIST-AID emulates the radiologist's comprehensive approach, leveraging historical data to improve diagnostic accuracy. Our experiments demonstrate significant improvements, with AUROC increasing by 6.56% and AUPRC by 9.51% compared to models that rely solely on radiographic scans. These gains were consistently observed across diverse demographic groups, including variations in gender, age, and racial categories. We show that while recent data boost performance, older data may reduce accuracy due to changes in patient conditions. Our work paves the potential of incorporating historical data for more reliable automatic diagnosis, providing critical support for clinical decision-making. The code for generating the data and model training is available at https://github.com/NoTody/HIST-AID.

Keywords: Temporal Dataset, Radiology Reports, Chest X-Rays (CXR), Time-Series, Multi-modal Learning

1. Introduction

Chest X-ray (CXR) is widely used for diagnosing thoracic abnormalities due to its affordability, accessibility, and non-invasive nature (Wang et al., 2017; Irvin et al., 2019; Johnson et al., 2019, 2023; Huang et al., 2023). AI-driven clinical decision support systems have shown potential to match or exceed human diagnostic accuracy (Rajpurkar et al., 2018; Killock, 2020; Gaube et al., 2023). However, most deep learning models only focus on the latest scan, neglecting patients' historical data (Wang et al., 2017; Rajpurkar et al., 2017; Irvin et al., 2019; Khan et al., 2020). This oversight is a critical limitation, as radiologists incorporate a patient's medical history and track changes over time to provide a more accurate diagnosis.

To address this, we introduce the *Temporal MIMIC* dataset (see Figure 1), which provides a comprehensive longitudinal view of patient data by combining five years of radiology images from MIMIC-CXR (Johnson et al., 2019) with corresponding clinical reports from MIMIC-IV (Johnson et al., 2023). This dataset includes 12,221 patients, each with an average of eleven reports and thirteen scans, providing a rich temporal multi-modal dataset that facilitates the development of multi-modal models capable of detecting subtle changes in a patient's condition over time.

To fully leverage our proposed dataset, we propose *HIST-AID*, an end-to-end framework that leverages historical chest X-rays and reports for abnormality detection (see Figure 2). In clinical deployment, when a patient undergoes a Chest X-Ray, our frame-

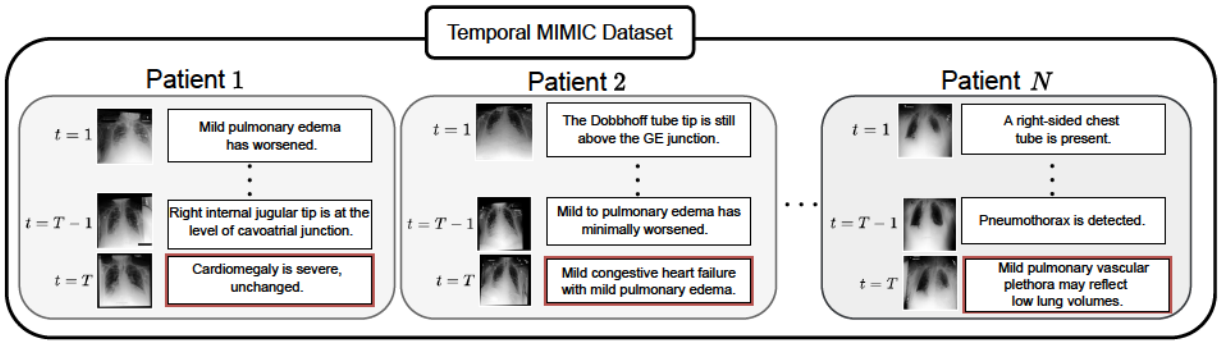


Figure 1: **Temporal MIMIC Dataset:** The dataset consists of radiographic scans and corresponding radiology reports collected over a span of five years, providing a comprehensive view of the progression of patient conditions over time. The final report, highlighted in red, is used to obtain the ground-truth labels for the patient's current condition.

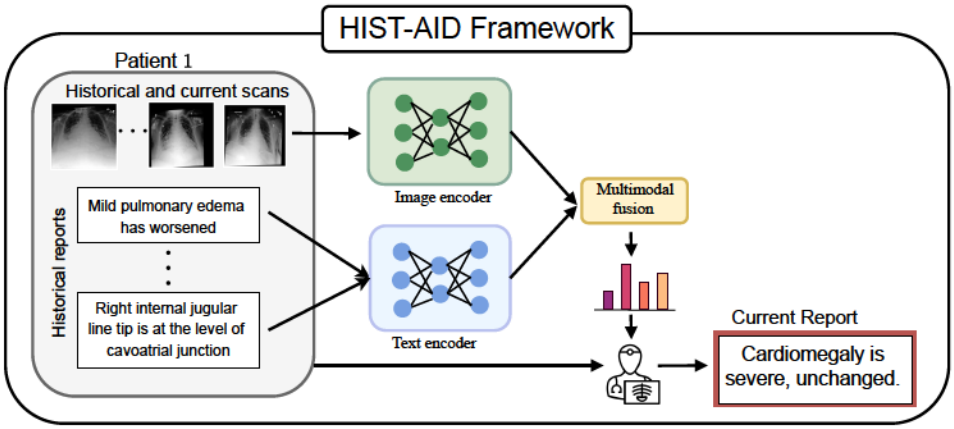


Figure 2: **Leveraging historical patient scans and reports for automatic diagnosis (HIST-AID) Framework:** We retrieve both current and historical image scans along with the radiology reports from the past. These inputs are first processed through image and text encoders. Subsequently, the resulting modality-specific representations are combined with time-series information (time offset from the current time stamp) to generate the final predictions with multi-modal fusion. This prediction, combined with the current scan and the patient's historical scans and reports, assists the physician in the final diagnosis.

work retrieves and integrates the historical clinical data from the healthcare system database. HIST-AID uses transformer-based image and text time-series encoders (Zerveas et al., 2021), effectively capturing temporal information from these past scans and reports. The temporal information from past scans and reports are then combined through multi-modal fusion (Kim et al., 2021) to make a preliminary diagnosis. The radiologists review and utilize this diagnosis to write the final report. By modeling these historical trends, HIST-AID can identify evolving patterns in a patient's condition that may not be apparent from a single scan.

Our evaluation shows that integrating past reports improves model performance across thirteen pathologies, with average AUROC and AUPRC increases of 6.56% and 9.51% compared to current scan only methods. This improvement is consistent across subgroups defined by gender, age, and race, ensuring a more equitable diagnostic approach. Incorporating past scans with reports did not yield additional gains, likely due to overlapping information between the two modalities for abnormality detection. Additionally, we found that reports from distant timestamps led to a decline in performance, highlighting that not all historical data is equally useful.

2. Related work

In this section, we provide an overview of the datasets for CXRs and radiology reports, along with the machine learning models developed using these datasets.

Datasets. Chest radiography is essential for early detection and diagnosis of critical health conditions, with recent deep learning advancements enhancing its effectiveness through large-scale datasets. Prominent datasets like ChestX-ray14 (Wang et al., 2017) and CheXpert (Irvin et al., 2019) provide extensive radiographic images with automated annotations, while the MIMIC dataset (Johnson et al., 2019, 2023) offers a large number of CXRs along with Electronic Health Record (EHR), including timestamps, patient identifiers, and hospital admission data.

Our work is the first to introduce a longitudinal dataset that combines historical CXRs from MIMIC-CXR (Johnson et al., 2019) and corresponding clinical reports from MIMIC-IV (Johnson et al., 2023), allowing models to leverage historical medical data to improve diagnostic accuracy.

Models. Numerous deep learning approaches have advanced pathology diagnosis prediction. Early works like CheXNet (Rajpurkar et al., 2017) demonstrate performance comparable to experienced radiologists in Chest X-ray pathology prediction. Recent studies in multi-modal pre-trainin and leveraging images, text, and tabular data further improves predictive models. ConVIRT (Zhang et al., 2022b) enhances visual representations with image-text pairs, while BiomedCLIP (Zhang et al., 2023) scales CLIP-style pre-training with biomedical data from PubMed. BioViL-T (Bannur et al., 2023) explores image-text contrastive pre-training using historical and current images.

Recently, time-series modeling has also gained attention in the community. Barbieri et al. (2020) evaluates RNNs on ICU readmission risk using electronic medical records, and Kaushik et al. (2020) shows the effectiveness of ensembles and LSTMs in predicting healthcare costs. HAIM (Soenksen et al., 2022) integrates CXRs and reports along with time-series information but lacks end-to-end training and proper timestamp segregation, resulting in biases and data leakage. Our work incorporates both historical radiology notes and images by addressing issues like improper data partitioning and refining the overall problem setup to enable accurate, end-to-end diagnosis using patient history.

3. Method

In this section, we outline the problem setup for abnormality prediction using historical data, followed by the process of generating the Temporal MIMIC dataset. We then introduce our HIST-AID framework, which leverages this dataset to enhance diagnostic accuracy.

3.1. Problem Setup

We consider a dataset of N patients, represented as $\{(X_i^{\text{image}}, X_i^{\text{text}}, y_i)\}_{i=1}^{N}$, where each patient i is associated with a temporal sequence of imaging scans and corresponding clinical reports over time. The temporal sequence is indexed by $T = (1, \ldots, t_n)$, where t_n denotes the latest timestamp for the imaging data.

The imaging data for each patient i is expressed as $X_i^{\text{image}} = \{x_{i,t}^{\text{image}}\}_{t=1}^{t_n}$, where each $x_{i,t}^{\text{image}}$ represents an image from timestamp t. The corresponding sequence of textual reports is denoted as $X_i^{\text{text}} = \{x_{i,t}^{\text{text}}\}_{t=1}^{t_{i'}}$, with $t_{i'}$ typically being one timestamp prior to t_n.

The objective is to predict the pathology label y_i for the scan at the current timestamp t_n, leveraging all available imaging data and associated historical reports from previous timestamps.

3.2. Temporal-MIMIC dataset generation

The objective of Temporal MIMIC dataset is to enable the use the historical images and patient reports to evaluate their utility for automatic diagnosis. To accomplish this, we integrate Chest-X rays from MIMIC-CXR (Johnson et al., 2019) and radiology reports from MIMIC-IV (Johnson et al., 2023), linked through patient subject identifiers.

Temporal MIMIC contains $12,221$ patients from $69,077$ radiographic studies, with an average of eleven reports and thirteen images per patient, paired with free-text radiology reports collected between 2011 and 2016 at Beth Israel Deaconess Medical Center in Boston, MA. Each data point is timestamped relative to the patient's most recent radiology image and includes corresponding ground truth pathology labels. The dataset spans thirteen distinct multi-label pathologies, including atelectasis, cardiomegaly, edema, lung opacity, pleural effusion, pneumonia, and pneumothorax, as well as fewer instances of fractures, lung lesions, and other pleural disorders.

The construction of Temporal-MIMIC dataset consists of the the following steps.

1. **Data merging.** We link the images from MIMIC-CXR (Johnson et al., 2019) with corresponding reports drawn from the time-series patient records of MIMIC-IV (Johnson et al., 2023). This connection is facilitated by the common identifiers present in both datasets, with labels derived directly from the radiology reports.

2. **Remove current time-stamp.** We exclude the current timestamp data from the reports to simulate a real-world scenario where a diagnosis must be made without immediate access to the latest diagnostic information.

3. **Augmentation with additional patient samples.** We treat each valid timestamp during a patient's admission period as a separate sample, significantly increasing the number of samples per patient. Each timestamp is linked to the corresponding labels in MIMIC. For each timestamp, we create a new datapoint by combining the label from that timestamp with all previous timestamps. For instance, if a patient has five timestamps, we take the label from the fifth timestamp and combine it with the prior four timestamps to form one datapoint. Similarly, we take the label from the fourth timestamp and combine it with the preceding three timestamps as another datapoint, and so on.

4. **Removing duplicates.** Post-merging, any duplicate records identified within the historical patient data are removed to ensure dataset integrity and prevent redundancy in the training process. Entries with empty impression section are also removed in this step.

5. **Dataset splitting.** Finally, the dataset is divided into 80% (n = 55,471) for training, 10% (n = 6,776) for validation, and 10% (n = 6,830) for testing. This split is performed using unique subject IDs from the MIMIC dataset to ensure that all data points related to a single patient are contained within one subset, thereby avoiding potential data leakage across the different phases of model evaluation.

Figure 1 shows various samples from our dataset. Figure A.9 shows the construction of Temporal MIMIC dataset. We provide the details on demographic distribution of the study population, label distribution and co-occurrences in the supplementary material.

3.3. HIST-AID framework

HIST-AID, shown in Figure 2, leverages historical images and reports in a temporal model, mimicking the radiologist's workflow and improving diagnostic performance over using only the current scan.

We use distinct modality-specific encoders: f_θ^{image} and f_ϕ^{text} for CXRs and corresponding reports from different time-stamps. These time-series representations are then processed by a separate time series encoder for each modality. The output of these encoders is followed by aggregation using multi-modal fusion encoder h_τ. The predicted pathology output $\widehat{y}$ leveraging both the historical imaging scans and associated reports as follow:

$$\widehat{y} = h_\tau \left(\bigoplus_i f_\theta^{\text{image}} \left(X_i^{\text{image}} \right), \bigoplus_i f_\phi^{\text{text}} \left(X_i^{\text{text}} \right) \right) \tag{1}$$

The aggregation operation $\bigoplus$ means the separate representations from different timestamps are concatenated to form a sequence of representations, where the sequence is padded by zero vectors if the sequence length is shorter than pre-defined maximum sequence length. We discuss the components in greater detail below.

3.3.1. Pre-trained encoders

In numerous studies on medical data (Sowrirajan et al., 2021; Mei et al., 2022; Wang et al., 2022; Eslami et al., 2023; Zhang et al., 2023), models pre-trained on relevant datasets consistently outperformed baseline models that were not pre-trained. In our work, CXRs and reports are encoded by pre-trained modality-specific encoders: a vision transformer (ViT) (Dosovitskiy et al., 2021) for images and a BERT encoder (Devlin et al., 2019) for text, both from Biomed-CLIP (Zhang et al., 2023).

These encoders are denoted as f_θ^{image} and f_ϕ^{text} and are used to process historical radiology images X_i^{image} and reports X_i^{text} respectively (see Equation (1)). We use the embedding of the $[CLS]$ token and append zero to the last dimension to accommodate time-series data that is shorter than the maximum length.

3.3.2. Time-series and multi-modal encoder

To effectively capture longitudinal information across different timestamps, our approach uses a transformer encoder for both time-series and multi-modal inputs, rather than averaging representations across

timestamps (Soenksen et al., 2022). The time-series and multi-modal inputs are encoded into $\{X_i^{\text{image}}, X_i^{\text{text}}\} \in \mathbb{R}^{B\times K\times D}$, where B is the batch size, K is the maximum time-series length, and D is the output dimension of the encoder using the $[CLS]$ token embedding for each image or text encoder. Learnable tokens, [IMG] for images and [TEXT] for text, are added to each representation at each timestamp. If the time-series length is less than K, zero padding is applied.

For simplicity, assume the maximum image timestamp is T and the maximum text timestamp is $T-1$. We define $f_\phi\left(x_{i,t_j}^{\text{modality}}\right) \in \mathbb{R}^{1\times d}$ as the output for each modality (image or text), and aggregate the outputs across timestamps as $\bigoplus f_\phi^{\text{modality}}(X_i) \in \mathbb{R}^{K\times d}$. The concatenated outputs, along with the $[CLS]$ token, form the transformer input with a shape of $\mathbb{R}^{2T}$.

We employ Rotary Positional Encoding (RoPE) (Su et al., 2021) to encode time-series information. We discuss the effectiveness of RoPE over other positional encoding methods in the supplementary material. A min-max normalized time offset serves as the positional indicator within the time-series. Given time offsets $\mathbf{t} = \{t_1, ..., t_n\}$, the normalized offset at timestamp t_i is:

$$t_i^{\text{norm}} = \frac{t_i - \min(t_1, ..., t_n)}{\max(t_1, ..., t_n) - \min(t_1, ..., t_n)} \quad (2)$$

$[CLS]$ token is added to the multi-modal representations as $\{[CLS], X_i^{\text{image}}, X_i^{\text{text}}\} \in \mathbb{R}^{B\times K\times(D+1)}$, capturing holistic information across modalities. This aggregated representation is encoded using a time-series transformer (Zerveas et al., 2021) for early fusion, and the encoded output is fed into a linear classifier for pathology classification.

4. Results

We compare the performance of HIST-AID on the Temporal MIMIC dataset against a uni-modal model that uses only the current scan (Rajpurkar et al., 2017; Wang et al., 2017; Irvin et al., 2019). In addition, we analyze the performance across different demographic subgroups, the impact of increasing the number of historical reports and the influence of reports from different time periods. All models are trained with multi-label classification with mean and standard deviation over five runs. Further details about the dataset and hyper-parameters for the models are deferred to the supplementary material.

4.1. Temporal multi-modal learning improves pathology prediction

The incorporation of historical radiology reports with our temporal multi-modal model during fine-turning shows a significant improvement in AUROC performance when compared to models trained with scans from the current time-stamp for all the examined pathologies in Figure 3. We observe over 5% improvement in AUROC in specific pathologies such as Consolidation ($5.05\%, p < 0.0001$), Pleural Other ($5.55\%, p < 0.0001$), Pneumonia ($6.20\%, p < 0.0001$). On average, HIST-AID shows an enhancement of 6.56% ($p < 0.0001$) and 3.41% ($p < 0.0001$) in AUROC for all pathologies when compared to models relying solely on chest radiographs and historical reports respectively. Additionally, we observe a 9.51% ($p < 0.0001$) and 2.63% ($p < 0.0001$) improvement in average AUPRC (see supplementary material) for all the pathologies. These results confirm the advantage of using CXRs with the historical reports for detecting thoraic abnormalities.

4.2. Subcohort analysis

We compare the performance of the model trained using only the current scan to HIST-AID across various demographic sub-cohorts, including gender, age, and race, as shown in Figure 4. This analysis was inspired by previous research highlighting significant disparities in model fairness and effectiveness when applied to diverse groups (Seyyed-Kalantari et al., 2021; Zhang et al., 2022a; Yang et al., 2024; Vaidya et al., 2024). Our results show a clear advantage for HIST-AID over the current scan only approach across all demographic subgroups.

HIST-AID consistently demonstrates accurate pathology diagnosis across both male and female patients, suggesting it helps reduce gender disparities. For age, the model achieved higher AUROC for younger and middle-aged adults, though performance slightly declined for individuals over 60, potentially due to the increased complexity of disease manifestations in older patients. While HIST-AID performed slightly less effectively for the black population compared to other racial groups, it provided a substantial improvement of $6-7\%$ AUROC compared to the current scan only model across all racial demographics.

These findings highlight the potential of HIST-AID to enhance diagnostic accuracy across diverse demographic groups, helping mitigate biases present in models trained with only the current scan.

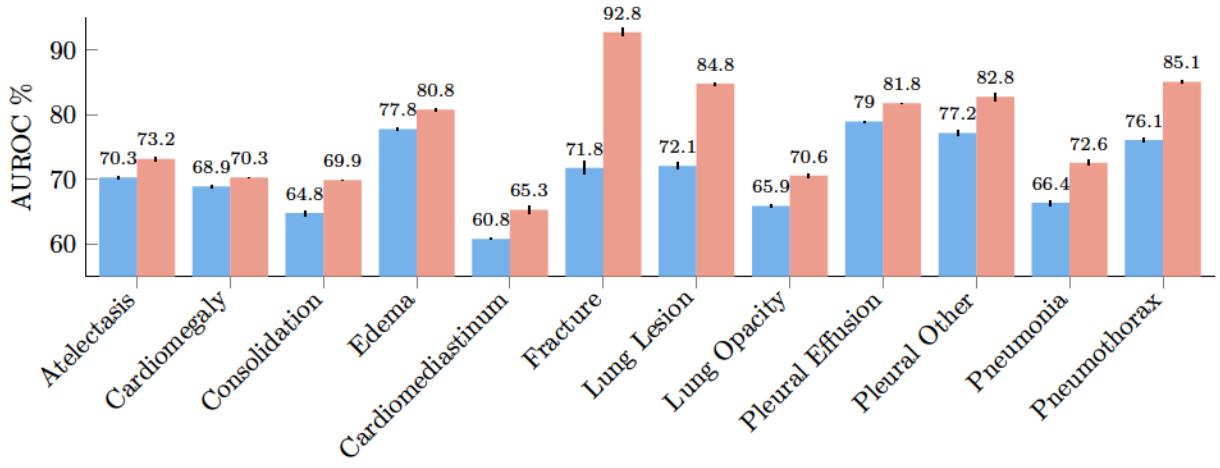


Figure 3: **AUROC comparison between different models for 13 pathologies.** We compare the image-only model in blue (left) bars that utilizes images from the current timestamp with HIST-AID in red (right), that integrates both current images and past textual data for diagnosis. We show that our model using both current scan and historical reports text enhances AUROC across all pathologies.

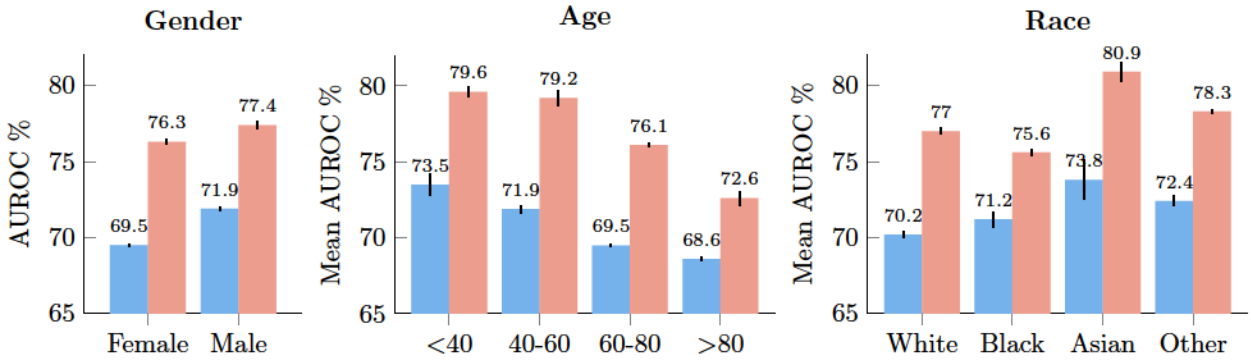


Figure 4: **AUROC comparison between model trained with current scan and HIST-AID across different demographic groups.** Our model in red (right) consistently outperforms the model in blue (left) trained with current scan images across gender, age groups, and racial categories. The error bars represent standard deviations calculated over five independent runs.

4.3. Enhancing pathology prediction with additional radiology reports

To assess the impact of the number of historical reports on automatic diagnosis, we measure the absolute AUROC improvement between models trained with both CXRs and historical reports and those trained with CXRs alone in Figure 5. We observe positive correlation between the number of reports and the AUROC improvement, indicating that incorporating more historical radiology reports improves diagnostic performance. Detailed breakdown across pathologies is in the supplementary material.

Figure 6 shows the temporal relevance of the reports by examining the impact of time intervals between the reports and the final diagnosis on AUROC. This analysis is crucial, as the number of reports does not necessarily correlate with their time distribution—multiple reports may originate from a single time period. We observe consistent performance improvement when reports are within 30 days of the diagnosis, while older reports tend to reduce AUROC. This highlights the importance of recent information in the diagnostic process and suggests caution when including older reports,

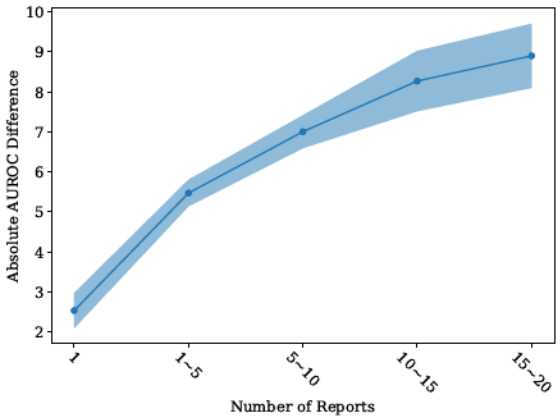


Figure 5: **Impact of increasing the number of reports on AUROC performance:** The performance of the temporal multi-modal model enhances as the number of reports increases, surpassing the model that relies solely on current timestamp images.

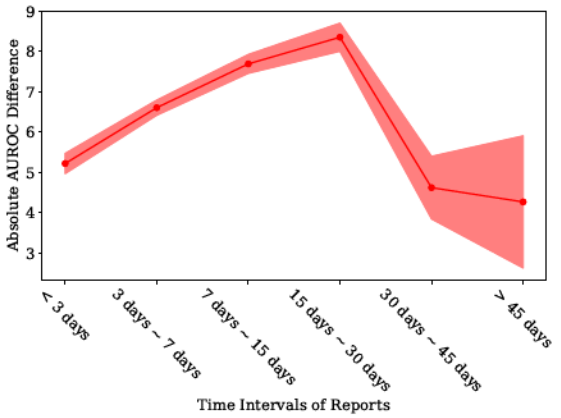


Figure 6: **Impact of report timing relative to final diagnosis on AUROC performance:** The performance of the temporal multi-modal model improves when utilizing reports from up to the last 30 days, but it declines with reports from more distant periods, cautioning against the use of older data.

Methods	FT	FLOPS
ConcatMLP	75.00±0.05	9.69G
Block (Ben-younes et al., 2019)	76.26±0.21	10.02G
MBT (Nagrani et al., 2021)	74.46±0.04	9.56G
ViLT (Kim et al., 2021)	**77.49**±0.14	9.38G
METER (Dou et al., 2022)	75.88±0.11	12.50G

Table 1: **Ablation on Different Fusion Methods.** Mean AUROC across all pathologies is reported in this study. We observe that ViLT represents the best trade-offs between model performance and compute.

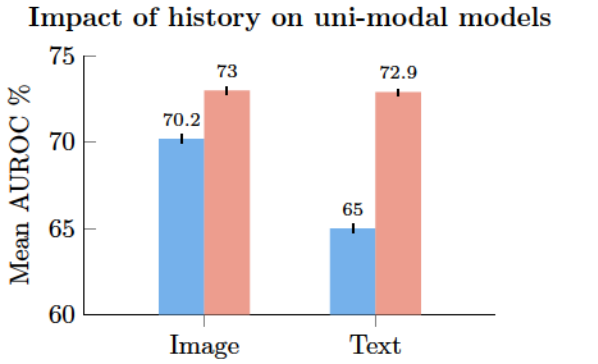
4.4. Effect of fusion methods

We show model performance on different multi-modal fusion methods in Table 1 with their detailed descriptions in the supplementary material. Our analysis highlights that ViLT (Kim et al., 2021) (early fusion) achieves an optimal balance between performance and computational efficiency within our framework. ViLT concatenates the representation tokens from image and historical reports and uses them as input to the time-series transformer encoder for information extraction. This demonstrates that early fusion is particularly effective for medical diagnosis tasks involving image and text modalities, where inter-modality dependencies (Madaan et al., 2024) are important. By integrating modality interactions from the outset, ViLT captures cross-modal relationships more effectively. Moreover, it avoids the need for separate time-series encoders for different modalities, making it more computationally efficient compared to alternative methods.

5. Limitations and future work

Challenges in leveraging historical radiographic scans for diagnostic performance. While our objective was to enhance diagnostic accuracy by incorporating both historical scans and reports, emulating the workflow of radiologists, our findings reveal limitations in the effectiveness of historical radiographic scans. Figure 7 compares models relying on the latest images and reports with those utilizing historical information in both uni-modal and multi-modal settings, with mean values and 95% confidence intervals across five independent runs.

In the uni-modal setting, incorporating historical CXRs and reports improves performance independently as expected. In the multi-modal setting, while historical reports boost diagnostic accuracy, adding historical images does not yield similar benefits. This limitation may stem from the fact that radiologists typically extract key pathology information from current images and document it in reports, making information from historical texts overlap with that of corresponding images.



Multi-modal models	Mean AUROC %
Current scan and last report	72.6 ± 0.19
Current scan and past reports	77.5 ± 0.17
All scans and past reports	77.7 ± 0.02

Figure 7: **Impact of historical images and reports on model performance.** When trained with images and text independently, the model's performance improves significantly with the inclusion of historical images and texts (red bars on the left) compared to using only the current scan and previous report (blue bars on the right). In multi-modal data, historical reports are beneficial, while historical images provide only marginal improvements.

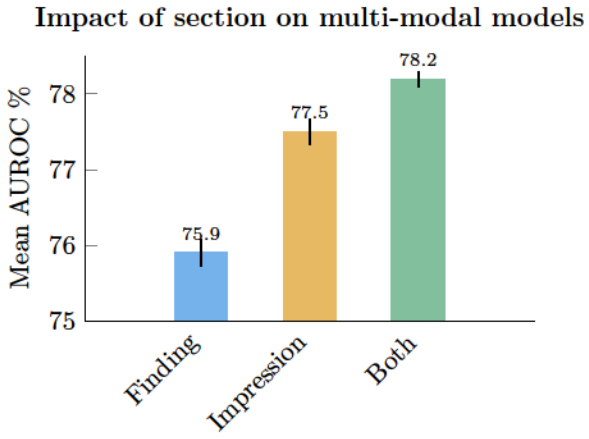


Figure 8: **Impact of report sections.** Mean AUROC for our multi-modal framework across all pathologies is reported in this study. We observe that impression is much more effective than finding and using both finding and impression can further improve the model performance when all historical reports are used.

Additionally, we hypothesize that optimization challenges in integrating high-dimensional historical scans into multi-modal models contribute to this issue. Addressing these challenges and developing more effective end-to-end multi-modal training techniques for incorporating historical scans is an important direction for future research.

Performance trade-offs in using different report sections. A limitation we identified is the varying contribution of different sections of radiology reports to model performance. As AI-based diagnostic tools evolve, understanding the impact of specific report sections becomes increasingly important, as these tools may influence how clinicians structure their reports. Radiology reports typically consist of five distinct sections, each serving a unique function:

- **History.** Provides a brief overview of the patient's medical background.
- **Indication.** Lists the reasons for conducting the radiological procedure.
- **Comparison.** References previous scans for comparison with the current one.
- **Findings.** Contains detailed observations made by the radiologist in each scanned area.
- **Impressions** Summarizes the key findings and their potential implications.

While all sections of the reports are important, many are frequently incomplete or missing. For instance, the history and comparison sections are often absent or lack sufficient detail. As a result, we focus on the findings and impression sections, as they contain the most critical diagnostic information and are more consistently present. Specifically, we analyze the impact of these sections on model performance to better understand their contribution. These sections are combined using the template: "Impression: ⟨*impression text*⟩ Finding: ⟨*finding text*⟩". As shown in Figure 8, including both sections improves performance by 0.69% ($p < 0.05$). However, this comes at a significant computational cost, as the findings section is two to three times longer than the impression section in terms of token length, as detailed in the supplementary material. Reducing training time and developing efficient models using all report sections is a promising direction for future research.

6. Conclusion

In this paper, we introduced the Temporal-MIMIC dataset, designed to assess model performance by integrating radiology images and reports across a patient's medical history. To leverage these historical images and reports for automated diagnosis, we proposed HIST-AID, a multi-modal framework that encodes modality-specific representations of both images and text, which are then combined through multi-modal fusion. Our results demonstrated that incorporating historical radiological reports alongside current scans significantly enhances the accuracy of automatic abnormality detection in chest X-rays, delivering consistent improvements across subgroups defined by gender, age, and race, thereby promoting a more equitable diagnostic approach. We showed that the impact of historical data varies across time, with the most recent reports – upto 30 days from diagnosis – being valuable, while older records tend to diminish predictive performance. This underscores the importance of carefully selecting relevant time periods when utilizing past medical information. By leveraging historical patient records, HIST-AID will enable specialists to comprehensively model patient histories, facilitating more effective identification of high-risk patients. This approach will help us to transform care delivery, improve treatment outcomes, and enhance overall healthcare efficiency.

Acknowledgments

We thank the anonymous reviewers for their insightful comments and suggestions. This research was supported by Samsung Advanced Institute of Technology (Next Generation Deep Learning: from pattern recognition to AI), Hyundai NGV (Uncertainty in Neural Sequence Modeling in the Context of Dialogue Modeling), NSF Award 1922658, the Center for Advanced Imaging Innovation and Research (CAI2R), a National Center for Biomedical Imaging and Bioengineering operated by NYU Langone Health, National Institute of Biomedical Imaging and Bioengineering through award number P41EB017183. The computational requirements for this work were supported in part by NYU IT High Performance Computing resources, services, and staff expertise and NYU Langone High Performance Computing Core's resources and personnel. This content is solely the responsibility of the authors and does not represent the views of the funding agencies.

References

Shruthi Bannur, Stephanie Hyland, and et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2023.

Sebastiano Barbieri, James Kemp, Oscar Perez-Concha, Sradha Kotwal, Martin Gallagher, Angus Ritchie, and Louisa Jorm. Benchmarking deep learning architectures for predicting readmission to the icu and describing patients-at-risk. *Scientific Reports*, 2020.

Hedi Ben-younes, Remi Cadene, Nicolas Thome, and Matthieu Cord. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *Proceedings of the AAAI National Conference on Artificial Intelligence*, 2019.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Minin*, 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *Proceedings of the International Conference on Learning Representations*, 2021.

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2022.

Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. Pubmedclip: How much does clip benefit visual question answering in the medical domain?

In *Findings of the Association for Computational Linguistics*, 2023.

Susanne Gaube, Harini Suresh, Martina Raue, Eva Lermer, Timo K Koch, Matthias FC Hudecek, Alun D Ackery, Samir C Grover, Joseph F Coughlin, Dieter Frey, et al. Non-task expert physicians benefit from correct explainable ai advice when reviewing x-rays. *Scientific reports*, 2023.

Haoxu Huang, Samyak Rawlekar, Sumit Chopra, and Cem M Deniz. Radiology reports improve visual representations learned from radiographs. In *Medical Imaging with Deep Learning*, 2023.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI National Conference on Artificial Intelligence*, 2019.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 2019.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 2023.

Shruti Kaushik, Abhinav Choudhury, Pankaj Kumar Sheron, Nataraj Dasgupta, Sayee Natarajan, Larry A Pickett, and Varun Dutt. AI in healthcare: Time-Series forecasting using statistical, neural, and ensemble architectures. *Front Big Data*, 2020.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In *Advances in Neural Information Processing Systems*, 2023.

Asif Iqbal Khan, Junaid Latief Shah, and Mohammad Mudasir Bhat. Coronet: A deep neural network for detection and diagnosis of covid-19 from chest x-ray images. *Computer Methods and Programs in Biomedicine*, 2020.

David Killock. Ai outperforms radiologists in mammographic screening. *Nature Reviews Clinical Oncology*, 2020.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the International Conference on Machine Learning*, 2021.

Divyam Madaan, Taro Makino, Sumit Chopra, and Kyunghyun Cho. Jointly modeling inter- & intra-modality dependencies for multi-modal learning. In *Advances in Neural Information Processing Systems*, 2024.

Xueyan Mei, Zelong Liu, Philip M. Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E. Link, Thomas Yang, Ying Wang, Hayit Greenspan, Timothy Deyer, Zahi A. Fayad, and Yang Yang. Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 2022.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems*, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 2020.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Yi Ding, Aarti Bagul, Curtis P. Langlotz, Katie S. Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint*, arXiv:1711.05225, 2017.

Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS medicine*, 2018.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 2021.

Luis R. Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M. Wiberg, Michael L. Li, Ignacio Fuentes, and Dimitris Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *npj Digital Medicine*, 2022.

Hari Sowrirajan, Jingbo Yang, Andrew Y. Ng, and Pranav Rajpurkar. Moco pretraining improves representation and transferability of chest x-ray models. In *Medical Imaging with Deep Learning*, 2021.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *CoRR*, 2021.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint*, arXiv:2302.13971, 2023.

Anurag Vaidya, Richard J. Chen, Drew F. K. Williamson, Andrew H. Song, Guillaume Jaume, Yuzhe Yang, Thomas Hartvigsen, Emma C. Dyer, Ming Y. Lu, Jana Lipkova, Muhammad Shaban, Tiffany Y. Chen, and Faisal Mahmood. Demographic bias in misdiagnosis by computational pathology models. *Nature Medicine*, 2024.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2017.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive learning from unpaired medical images and text. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2022.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

Yuzhe Yang, Haoran Zhang, Judy W. Gichoya, Dina Katabi, and Marzyeh Ghassemi. The limits of fair medical imaging ai in real-world generalization. *Nature Medicine*, 2024.

George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021.

Haoran Zhang, Natalie Dullerud, Karsten Roth, Lauren Oakden-Rayner, Stephen Pfohl, and Marzyeh Ghassemi. Improving the fairness of chest x-ray classifiers. In *Conference on health, inference, and learning*, 2022a.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Matthew Lungren, Tristan Naumann, and Hoifung Poon. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint*, arXiv:2303.00915, 2023.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Proceedings of the Machine Learning for Healthcare Conference*, 2022b.

Appendix A. Model Implementation and Training Details

A.1. Statistical Analysis

One-tailed Wilcoxon rank-sum test ($\alpha = 0.05$) was used to compute all p-values reported in the paper, where the samples are drawn from models trained on 5 different seeds to evaluate the stability of the models and robustness of the hypothesis.

A.2. Data Augmentations

For radiology images, we apply Random Resized Crop with scale $0.6 - 1.0$ and Bicubic Interpolation, Color Jittering with brightness $0.4 - 0.6$, contrast $0.4 - 0.6$, no saturation and hue change with probability 0.5, and Random Horizontal Flip with probability 0.5. We do not apply any data augmentations on radiology reports.

A.3. Hyperparameters

Table A.3 provides hyperparameters for training our framework. We keep the same hyperparameters in uni-modal cases. We performed a small-scale hyperparameters search to ensure our result does not change too much on different hyperparameters settings. For experiments of using finding section or both impression and finding sections, we use 400 Text Tokens Max Length due to computational constraint.

We show the distribution of time-series length for all samples in Figure A.11, where we select max sequence padding with truncation to 50 under evaluation of computational cost and length coverage.

A.4. Training Details

We use 12 layers Vision Transformer (ViT-Base) with path size 16×16 (Dosovitskiy et al., 2021) as Image Encoder and 12 layers BERT-Base (Devlin et al., 2019) as Text Encoder in all our experiments. The pre-trained weights of both encoders are loaded from BiomedCLIP (Zhang et al., 2023). We use the model checkpoint with best validation AUROC for testset performance evaluation. The validation performance is calculated after each epoch. We do not use mixed-precision training as we find the training to be unstable with mixed-precision. All experiments are performed on two NVIDIA A100 80GB GPUs with total training time range from approximately 15 to 20 hrs varied based on different settings.

A.5. Dataset

We show concept plot on how our dataset is constructed in Figure A.9. The demographic distribution of our generated dataset is shown in Table A.2, where it shows a wide range of race, sex and age and demographics. Additionally, label distribution and their co-occurrences are shown in Figure A.10.

Appendix B. Additional results

B.1. Different multi-modal fusion methods

- ViLT (Kim et al., 2021): We adapt ViLT as general early fusion case for transformer. Given input $\mathbf{x_1} = \mathbb{R}^{L_1 \times D}$ and $\mathbf{x_2} \in \mathbb{R}^{L_2 \times D}$, where L_1, L_2 are sequence length and D is dimension length of representation. $\mathbf{x_1}$ and $\mathbf{x_2}$ are first concatenated with [CLS] token $\mathbf{t}_{\text{cls}}$ of shape $\mathbb{R}^{1 \times D}$ to get input with size $\mathbf{x} = [\mathbf{t}_{\text{cls}} || \mathbf{x_1} || \mathbf{x_2}] \in \mathbb{R}^{(1+L_1+L_2) \times D}$. Then, the concatenated input is directly feed into a standard transformer model as $\text{Transformer}(\mathbf{x}; \theta)$ with learnable tokens $\mathbf{t_1} \in \mathbb{R}^{L_1 \times D}$ and $\mathbf{t_2} \in \mathbb{R}^{L_2 \times D}$ added to the input along with positional embedding to indicate different modalities. We used 1 layers transformer for early fusion.

- MBT (Nagrani et al., 2021): We adapt MBT as general intermediate fusion case for transformer. Given input $\mathbf{x_1}, \mathbf{x_2}$ following the notation of ViLT. An extra fusion token $\mathbf{x}_{\text{fsn}} \in \mathbb{R}^{L_3 \times D}$ is added in the intermediate layers of transformer such that $\mathbf{x} = [\mathbf{x_1} || \mathbf{x}_{\text{fsn}} || \mathbf{x_2}]$. Output for each layer of transformer output is then calculated as $[\mathbf{x}_i^{l+1} || \hat{\mathbf{z}}_{\text{fsn}_i}^{l+1}] = \text{Transformer}([\mathbf{x}_i^l || \mathbf{z}_{\text{fsn}}^l; \theta_i])$ given i as index for different modalities. The fusion token is then updated with $\mathbf{z}_{\text{fsn}}^{l+1} = \text{Avg}_i(\hat{\mathbf{z}}_{\text{fsn}_i}^{l+1})$. We used 6 layers transformer with last 3 layers as intermediate fusion layers in our experiment following the best experiment setting in original paper.

- ConcatMLP: For given input $\mathbf{x_1} = \mathbb{R}^I$ and $\mathbf{x_2} \in \mathbb{R}^J$, where concatenation $\oplus[\mathbf{x_1}, \mathbf{x_2}] \in \mathbb{R}^{I+J}$. Given two weight matrices $\mathbf{W}_1 \in \mathbb{R}^{(I+J) \times K}$ and $\mathbf{W}_2 \in \mathbb{R}^{K \times M}$ and sigmoid function σ, fusion output is represented as $\mathbf{y} = \mathbf{W_2}^T \sigma(\mathbf{W_1}^T \oplus [\mathbf{x_1}, \mathbf{x_2}])$

- Block Ben-younes et al. (2019): For given bilinear model $\mathbf{y} = \mathcal{T} \times_1 \mathbf{x_1} \times_2 \mathbf{x_2}$ with $\mathbf{x_1} \in \mathbb{R}^I, \mathbf{x_2} \in \mathbb{R}^J$ and $\mathcal{T} \in \mathbb{R}^{I \times J \times K}$, Block decomposes the bilinear model on $\mathcal{T}$ to get form of $\mathbf{y} = \mathbf{C}(\mathcal{D} \times_1$

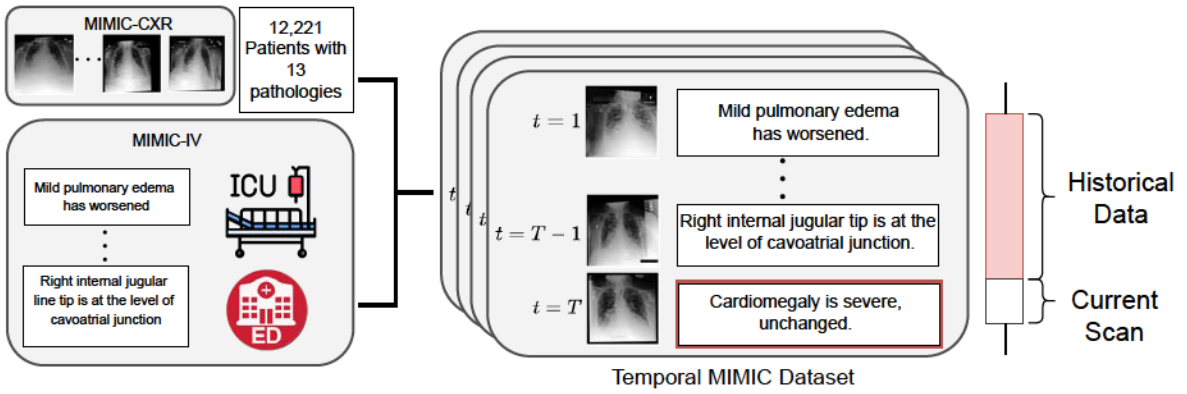


Figure A.9: **Overview of Temporal MIMIC creation:** We integrate historical and current Chest X-rays from MIMIC-CXR with radiology reports from MIMIC-IV using unique patient identifiers.

Characteristics		Count (Proportion %)
Race	Asian	2,269 (3.3%)
	Black	8,147 (11.8%)
	White	45,637 (66.1%)
	Other	13,024 (18.8%)
Sex	Female	29,021 (42.0%)
	Male	40,056 (58.0%)
Age	0-40	6,597 (9.6%)
	40-60	20,461 (29.6%)
	60-80	32,039 (46.4%)
	80-100	9,980 (14.4%)

Table A.2: **Demographic distribution of the study population.** The demographic breakdown reveals a predominantly white cohort, a slightly male-dominated gender ratio. Additionally, there is a significant segment of participants aged 40-80.

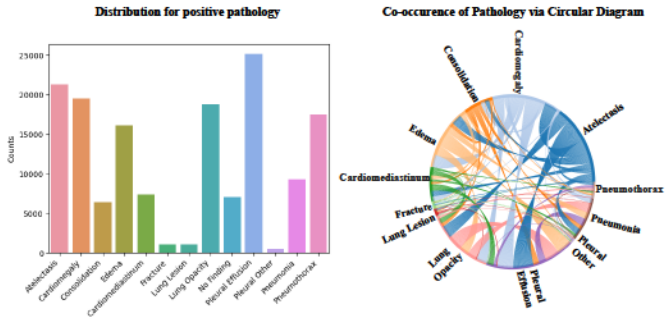


Figure A.10: **Distribution and co-occurrence of various pathologies.** The left bar graph representing the frequency of various pathologies such as cardiomegaly, consolidation, edema, lung opacity, nodule/mass, pneumonia, and pneumothorax within the Temporal-MIMIC dataset. The right circular diagram maps the interrelationships among these pathologies where the pathologies are interconnected by colored lines. The thickness of the lines reflect the prevalence of these co-occurrences.

$(\mathbf{x_1}^T\mathbf{A}) \times_2 (\mathbf{x_2}^T\mathbf{B}))$ with $\mathcal{D} \in \mathbb{R}^{L\times M\times N}, \mathbf{A} \in \mathbb{R}^{I\times L}, \mathbf{B} \in \mathbb{R}^{J\times M}$ and $\mathbf{C} \in \mathbb{R}^{K\times N}$.

- METER (Dou et al., 2022): Two transformer encoders composed of self-attention and cross-attention modules are used with two different modalities as input. The final output $\mathbf{y}$ is computed by a MLP layer on top of concatenation of $[CLS]$ token output of these two transformers. Taking Query, Key, Value of the transformers $Q_1 \in \mathbb{R}^{d_k}, K_1 \in \mathbb{R}^{d_k}, V_1 \in \mathbb{R}^{d_v}$ from one modality and $Q_2 \in \mathbb{R}^{d_k}, K_2 \in \mathbb{R}^{d_k}, V_2 \in \mathbb{R}^{d_v}$ from another modality, the cross-attention mechanism of these two transformers is calculated as $CrossAtt(Q_1, K_2, V_2) = \text{softmax}(\frac{Q_1 K_2^T}{\sqrt{d_k}})V_2$ and $CrossAtt(Q_2, K_1, V_1) = \text{softmax}(\frac{Q_2 K_1^T}{\sqrt{d_k}})V_1$.

B.2. Per Pathology Model Performance on Different Report Numbers and Time Intervals

We show all 12 classes performance for number of reports and time intervals ablation in Figure B.13 and Figure B.14 with mean and 95% confidence interval,

Hyper-Parameter	Value
Batch Size	16
Learning Rate (Image Encoder)	1e-5 × (Batch Size / 64)
Learning Rate (Text Encoder)	1e-5 × (Batch Size / 64)
Learning Rate (Time Series Encoder)	1e-4 × (Batch Size / 64)
Epochs	15
Weight Decay	1e-2
Optimizer	AdamW
AdamW Betas	(0.9, 0.999)
Scheduler	Cosine with Linear Warmup
Linear Warmup Steps	0.10 × Total Training Steps
Minimum Learning Rate	1e-3 × Learning Rate
Image Time-Series Max Sequence Length	1
Text Tokens Max Length	200
Text Time-Series Max Sequence Length	50

Table A.3: Hyper-parameters for Model Training

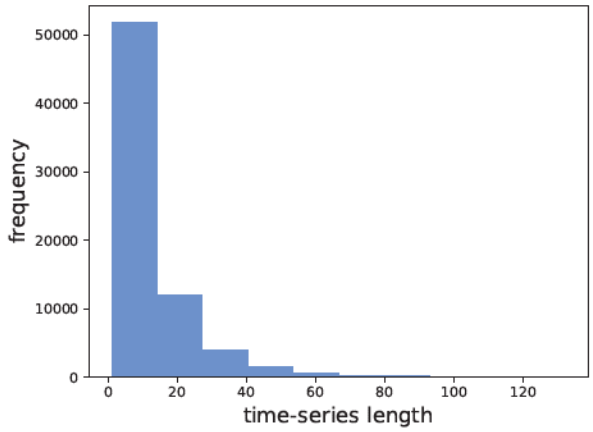


Figure A.11: Distribution for Text Time-Series Length

where we show that the general trend for all pathology in main paper hold for majority of pathologies with few exceptions.

We show additional ablation studies in Figure B.12 to justify our architecture selection. The interpretation for additional three ablation studies are present as follows.

B.3. Different Position Encoding

Since previous works have shown that position encoding can have important impact on model performance for transformer-based language model (Raffel et al., 2020; Workshop et al., 2022; Touvron et al., 2023; Kazemnejad et al., 2023), we ablate on some popular position encoding in our framework with position to be indicated with time-stamps. The purpose of this ablation is to identify the optimal ways of injecting temporal information to input tokens. In Figure B.12 (a), we compare model performance on different positional encoding, where the definition of sine-cosine, learnable and RoPE (Su et al., 2021) positional encoding are defined in the Supplementary Material. While using no, sine-cosine or learnable positional encoding show similar performance, RoPE shows clear improvement over other methods. This indicates that adding time-series information with relative positional encoding (RoPE) to our architecture can clearly benefit model performance.

B.4. Different Pooling Methods

In order to explore the effectiveness of pooling time-series representations with transformer encoder, we compare time-series pooling with transformer (TST) vs. mean pooling (Mean) used in HAIM (Soenksen et al., 2022) for evaluating the benefits of pooling with time-series transformer in Figure B.12 (b). The result shows that pooling time-series reports with time-series transformer is more effective than mean pooling when either training with time-series reports along or multi-modal training with both image and reports. When mean pooling is performed, the model is not capable of retaining information interactions across various time-stamps, where time information in each time-stamp is encoded equally. Conversely, the integration of self-attention, complemented by time-stamp positional encoding, enhanced the model's ability to discern and leverage nuanced interactions. This facilitates a more refined and detailed interaction of the representations extracted from different time-stamps.

We further compare with HAIM [1] using feature extraction and fusion with XGBoost (Chen and Guestrin, 2016). The results indicate suboptimal performance, achieving $\sim$ 61.14 average AUROC for the image uni-modal, $\sim$ 59.61 for the text uni-modal, and $\sim$ 63.92 for the image-text multi-modal model. We attribute this under-performance to the lack of weight updates for the modality-specific models. Further investigation and incorporation of more modalities would be an interesting direction for future work.

1. https://github.com/lrsoenksen/HAIM

B.5. Different Modality Combinations

In order to examine the effectiveness of multi-modal fusion in comparison to simplest way of combining different modality with logits averaging, we compare model performance on different modality combination methods in Figure B.12 (c). Ensemble means the logits output from two modality-specific encoders are directly averaged and fusion means early fusion is performed on the representation of two modality-specific encoders. For fair comparison, two encoders trained on different seeds are used for ensembling for image and text result. The result shows that fusion always outperform other ensembling options, showing the effectiveness of our proposed framework.

B.6. Per Pathology Model Performance on AUROC and AUPRC Plots

We show all 12 classes performance with AUROC and AUPRC plots in Figure B.16 and Figure B.17. In total, there are 4.48% labels multi-modal gets right while image based uni-model gets wrong.

B.7. Dataset Sample

We show an example of a current radiology image with associated past reports and ground truth in Figure Figure B.18.

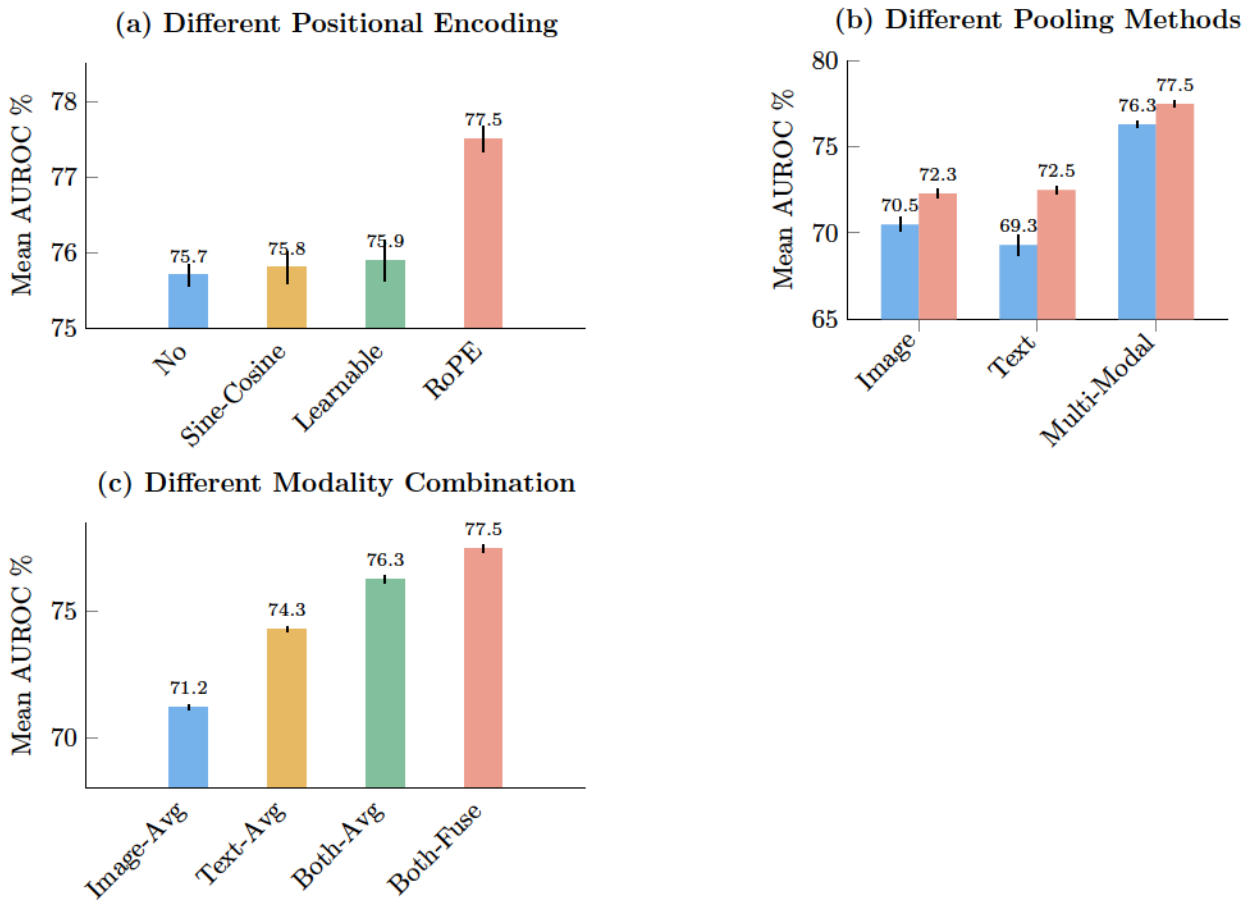


Figure B.12: **Additional Ablation Studies. a) For positional encoding ablation**, we evaluate four most commonly used positional encoding, where we find using relative encoding with RoPE consistently perform better. **b) For pooling method ablation**, we compared mean pooling (blue bars) and time-series transformer pooling (red bars), where time-series transformer pooling consistent outperform mean pooling. **c) For different modality combination comparison**, we compared ensembling with logits averaging (blue, yellow and green bars) vs. early fusion (pink bar), where early fusion consistently performs better than logits averaging from two models of experimented modalities.

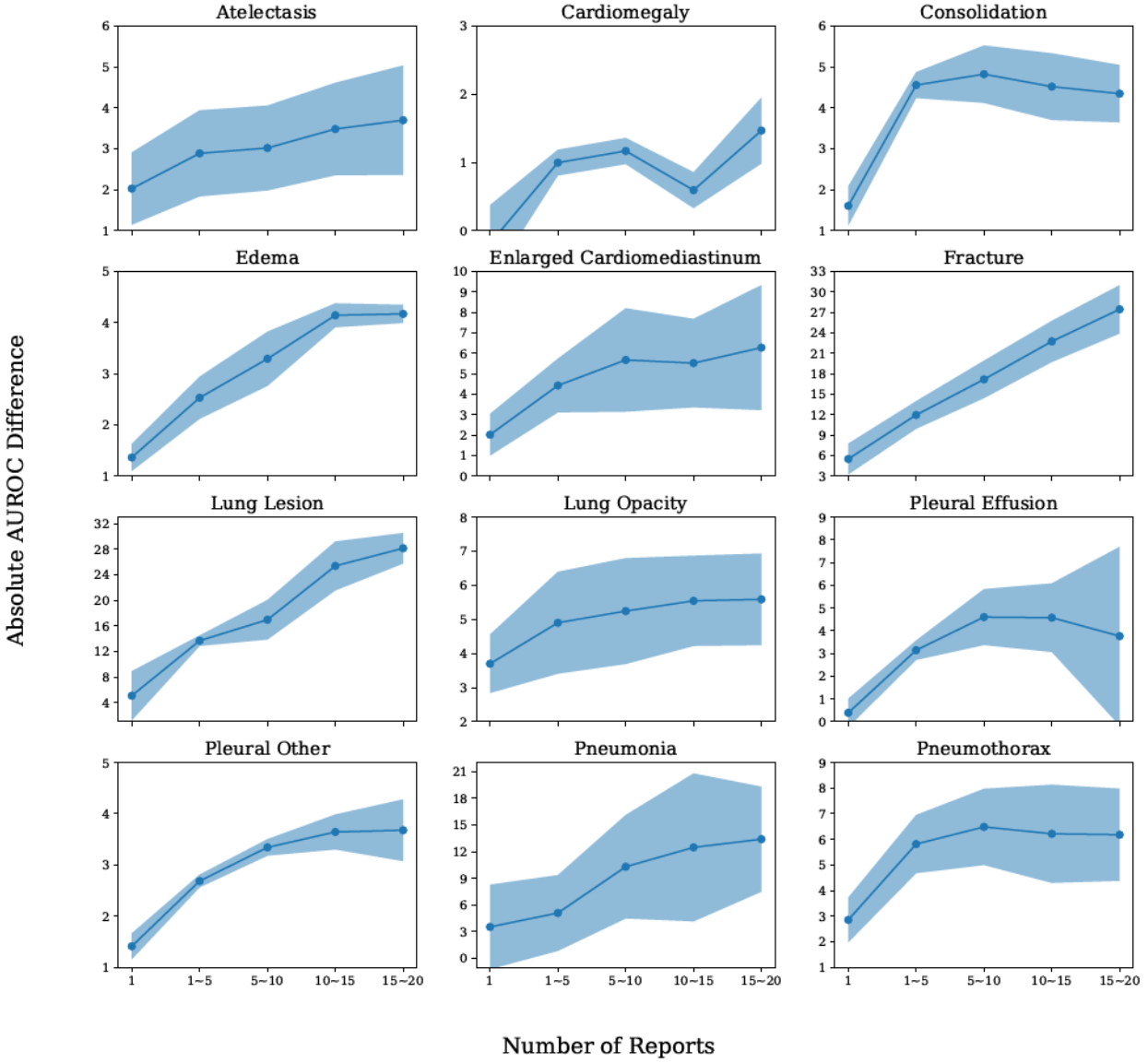


Figure B.13: Performance Difference of model trained only on current scans in comparison to current scan Images and historical reports on different number of reports for different pathologies.

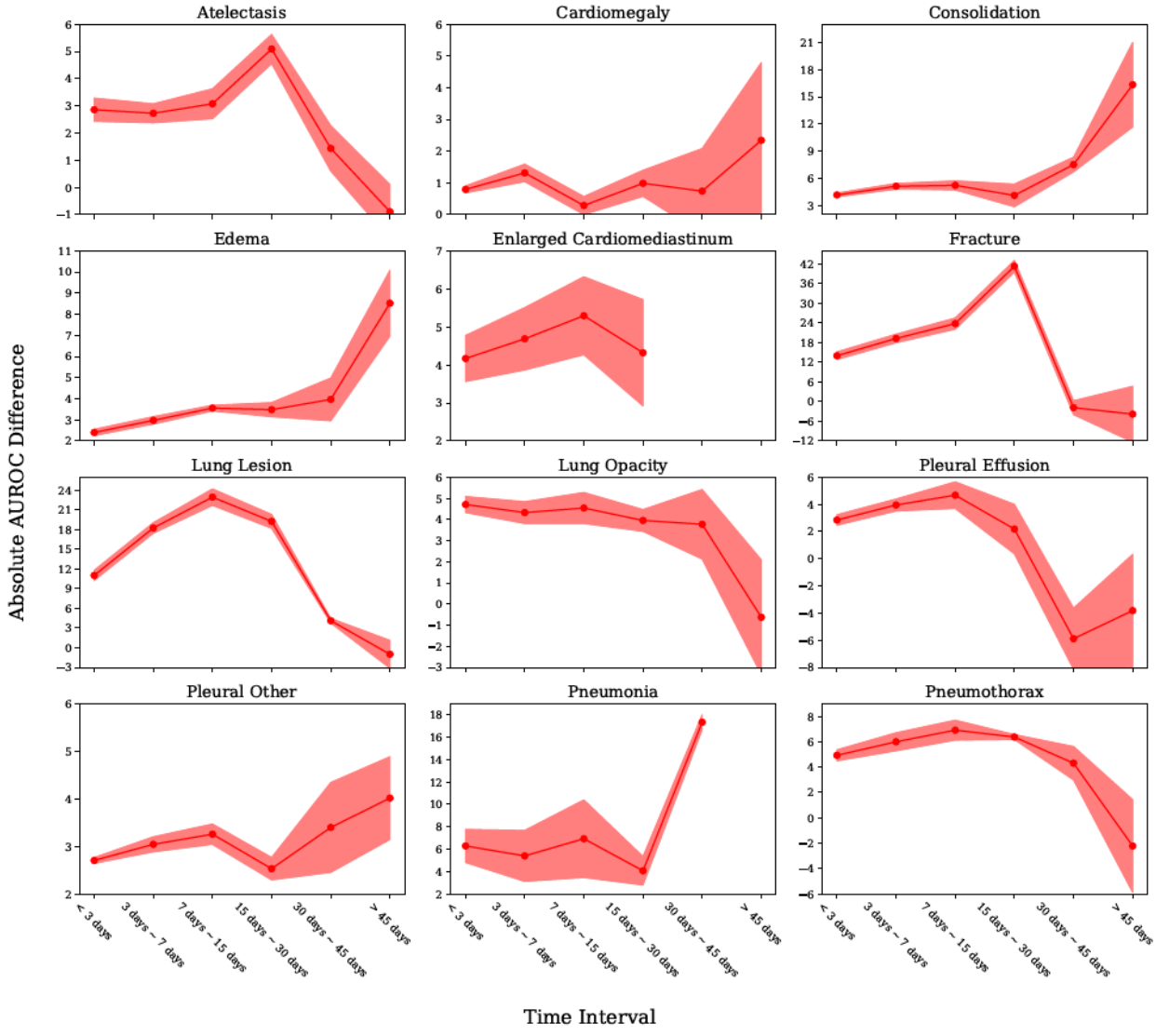


Figure B.14: Performance Difference of model trained only on current scans in comparison to current scan Images and historical reports on different time intervals for different pathologies.

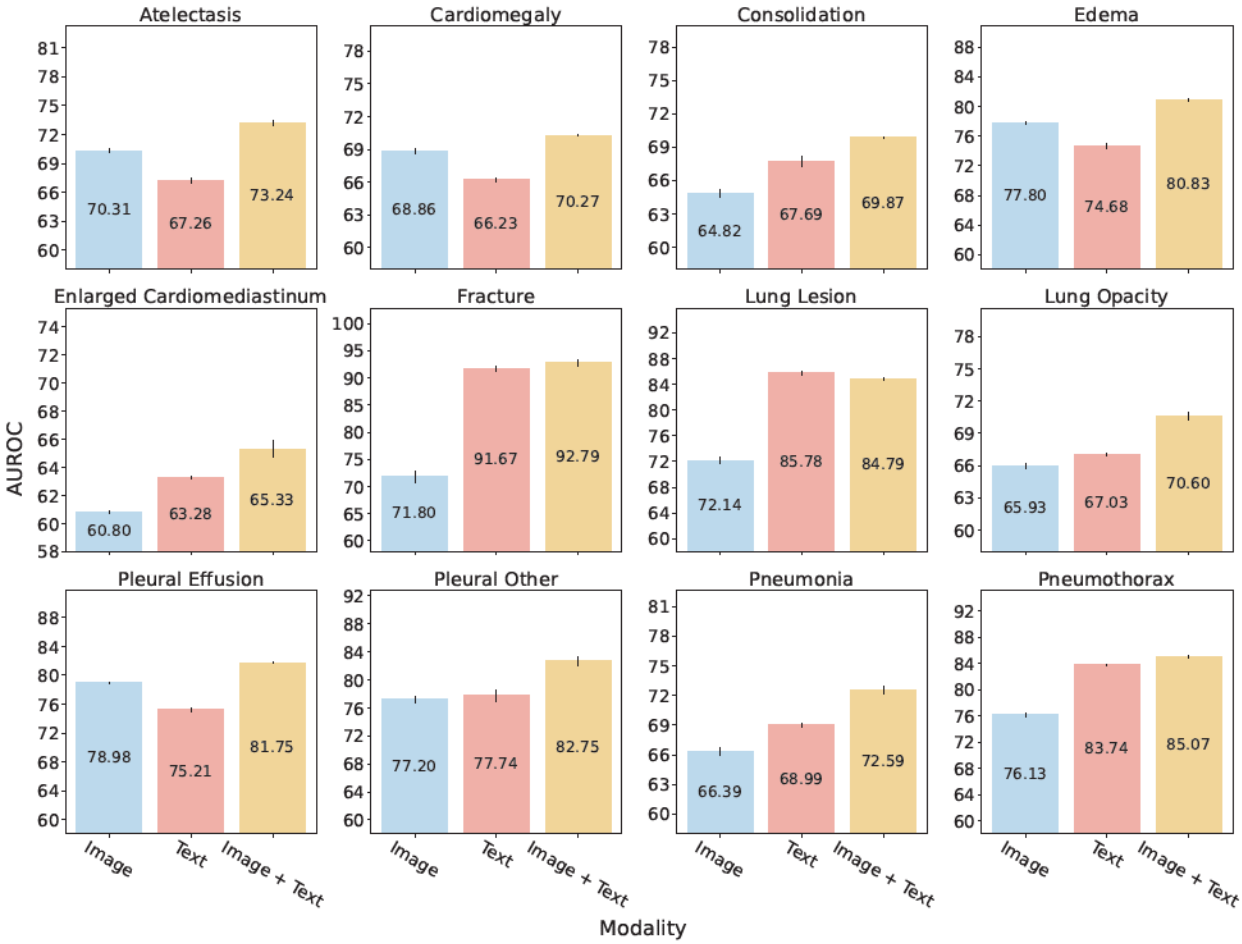


Figure B.15: **AUROC comparison between different models for 13 pathologies.** We evaluate the performance of three models (bars in the following order): an image-only model in that utilizes images from the current timestamp, text-only model that uses reports from previous timestamps, and our comprehensive model that integrates both current images and past textual data for diagnosis. Our model markedly enhances diagnostic accuracy across all examined pathologies.

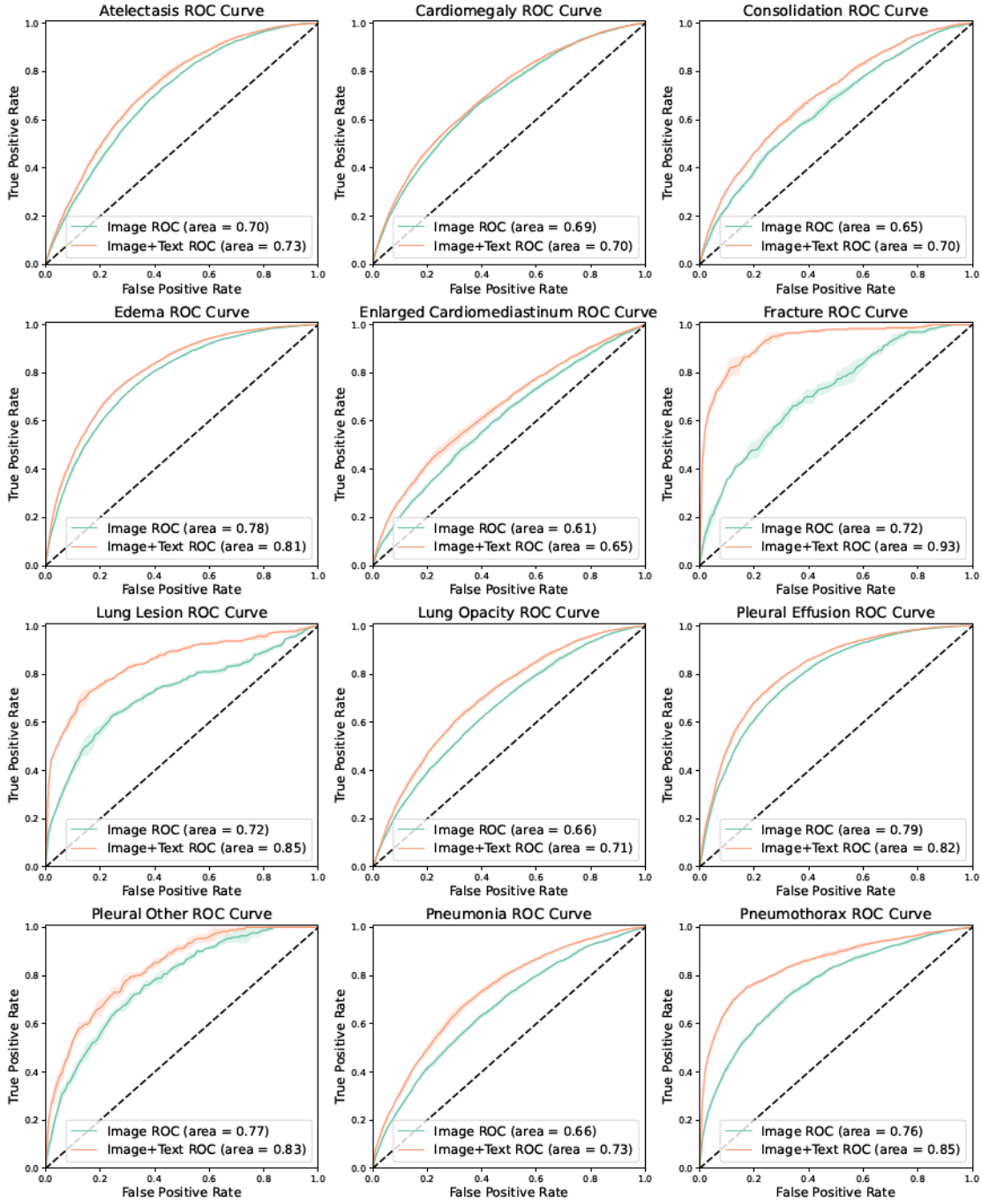


Figure B.16: Per Pathology ROC Curve

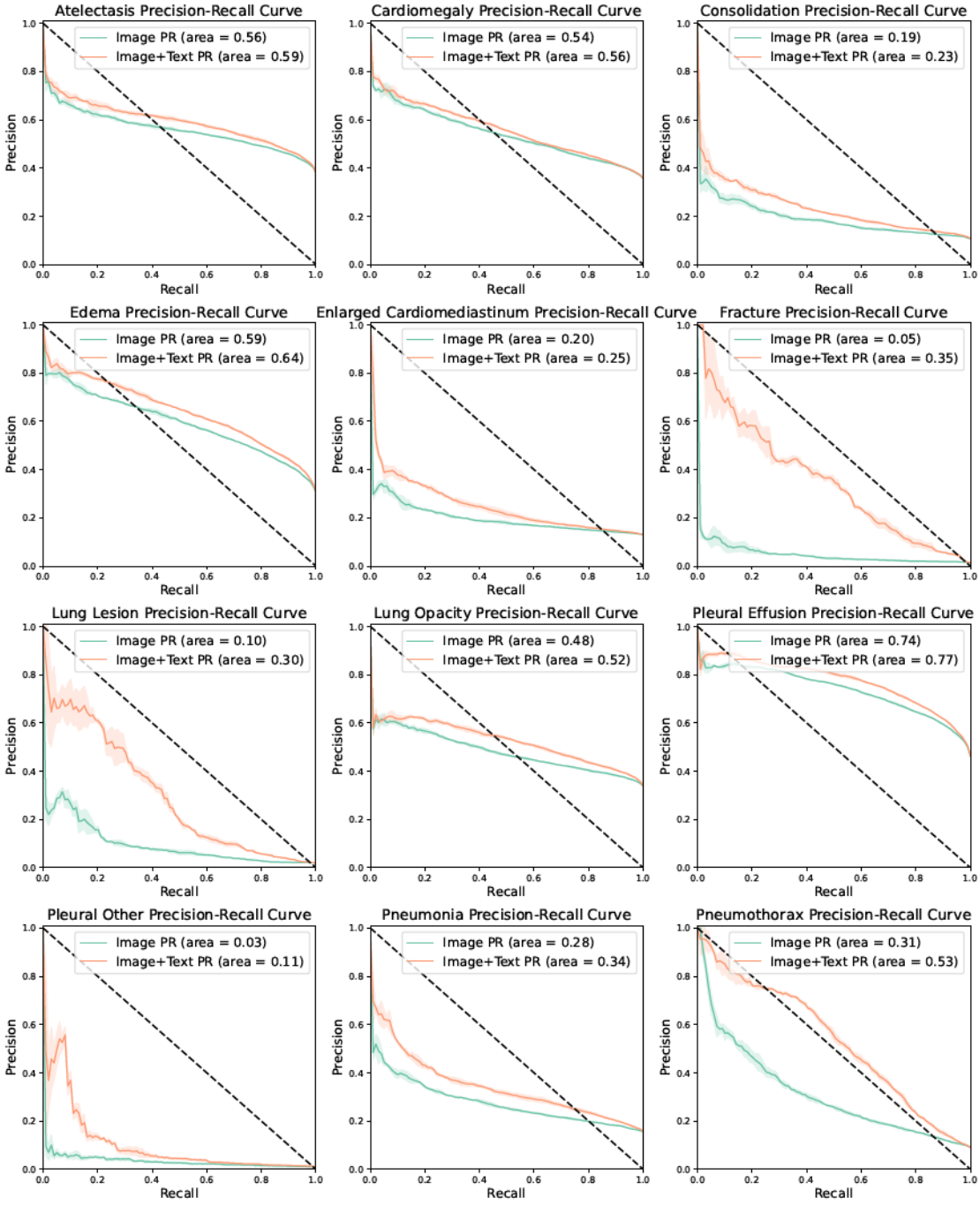


Figure B.17: Per Pathology Precision-Recall Curve

Current Image:

Labels for Current Image:

Atelectasis: 0.0
Cardiomegaly: 1.0
Consolidation: NaN
Edema: 0.0
Enlarged Cardiomediastinum: NaN
Fracture: NaN
Lung Lesion: NaN
Lung Opacity: 1.0
No Finding: NaN
Pleural Effusion: 0.0
Pleural Other: NaN
Pneumonia: NaN
Pneumothorax: NaN
Support Devices: NaN

Radiology Reports in the Past

Time

1. ET tube terminates approximately 2.9 cm above the carina. 2. Retrocar diac opacity with air bronchograms. Small left pleural effusion. FINDING S could be seen with an infectious process.

1. A 4 (AP) x 5 (TV) x 4 (SI) mm postero-lateral pointing A-comm aneurys m. 2. Diffuse subarachnoid hemorrhage, better visualized on the prior non contrast CT. 3. Small, nonspecific left temporal and left nasal soft tiss ue emphysema. Recommend clinical correlation.

In COMPARISON with the earlier study of this date, the tip of the orogas tric tube is within the stomach, though the side port is at the level of the esophagogastric junction. The tube should be pushed forward several cm. Little overall change in the appearance of the heart and lungs.

1.Right-sided intraventricular catheter traversing frontal on the right lateral ventricle and terminating in the third ventricle. 2. Diffuse sub arachnoid hemorrhage, increased compared to prior imaging, with extensio n into the subdural horn of the lateral ventricles, particularly promine nt in the right sylvian fissure. 3. Possible small left and right subdur al hematomas along the parietal convexities. 4. Mild ventriculomegaly, u nchanged from prior imaging. 5. No infarction or new hemorrhage.

In COMPARISON with the earlier study of this date, the nasogastric tube is been pushed forward so that the side port is distal to the esophagoga stric junction. Endotracheal tube tip remains in good position. There i s increasing opacification at the left base consistent with volume loss left lower lobe and pleural fluid. Little change in the mild atelectati c FINDINGS at the right base.

1. There is no evidence of splenomegaly. 2. Vascular calcifications and/ or granulomas in splenic parenchyma.

Left IJ catheter ends in the upper SVC. Stable consolidation in the right lower lung may reflect pneumonia.

Figure B.18: **Dataset Sample:** We show a pre-processed sample from our Temporal MIMIC dataset containing current image and corresponding labels and all previous reports in chronological order.