MODELING CONTINUOUS SPATIO-TEMPORAL DISEASE EVOLUTION FOR LONGITUDINAL RADIOLOGY REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Radiology report generation aims to assist radiologists in drafting medical reports by automatically generating descriptive text from images, thereby reducing workload and improving diagnostic efficiency and accuracy. Existing report generation methods often treat disease changes under irregular temporal sampling as fixed-interval discrete jumps, resulting in inaccurate disease state characterization. To address this limitation, we propose a novel **Co**ntinuous spatio-temporal **D**isease **E**volution framework for longitudinal **R**adiology **R**eport **G**eneration (CoDE-RRG), which models continuous disease evolution from arbitrary numbers of prior visits and multi-view images. Specifically, we propose a view-stable multi-view fusion module, where a primary view serves as an anchor to aggregate information from all auxiliary views. Departing from traditional discrete modeling of longitudinal changes, we first employ disease-guided dimensionality reduction to project high-dimensional visual features into a low-dimensional disease-specific state space, and then innovatively introduce Neural Ordinary Differential Equations to compute the continuous temporal evolution trajectories of these disease states. It allows for the construction of individualized disease progression tracks that align with actual physiological changes. Extensive experiments demonstrate that our method consistently improves the modeling of longitudinal disease evolution and current disease states, with relative gains of up to 2.3% in BLEU-4 and 4.6% in CheXbert F1 over state-of-the-art methods. Our code is available at `https://anonymous.4open.science/r/CoDE-RRG-E007/`.

1 INTRODUCTION

Chest X-ray (CXR) is a widely used non-invasive modality for thoracic disease screening and diagnosis. However, interpreting CXRs and generating reports are time-consuming, while the growing imaging volume further burdens radiologists. Consequently, automated radiology report generation (RRG) has emerged as a key research direction in computer vision and natural language processing, aiming to translate visual features into coherent clinical reports.

Despite recent advances in RRG, most existing methods focus on interpreting a single current image (Shen et al., 2024; Liu et al., 2024a; Xiao et al., 2025; Yin et al., 2025), while radiologists routinely rely on longitudinal comparisons to assess disease progression. Some recent studies have begun to incorporate historical examinations (Zhu et al., 2023; Yun et al., 2025; Liu et al., 2025b; 2026; Gao et al., 2026b). However, as illustrated in Fig. 1, disease evolution is inherently a continuous physiological process governed by temporal regularities, where patients undergo multiple visits with irregular intervals and multi-view acquisitions, exhibiting progressive changes over time. In contrast, existing longitudinal methods typically compare the current study only with the most recent prior visit, while neglecting irregular temporal intervals. As a result, disease evolution is often modeled as a discrete transition under a fixed temporal-interval assumption rather than a continuous physiological progression, limiting the model's ability to capture disease states and temporal trends. These limitations raise a fundamental question: **how can we effectively model physiological dynamics from irregular longitudinal clinical data?**

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

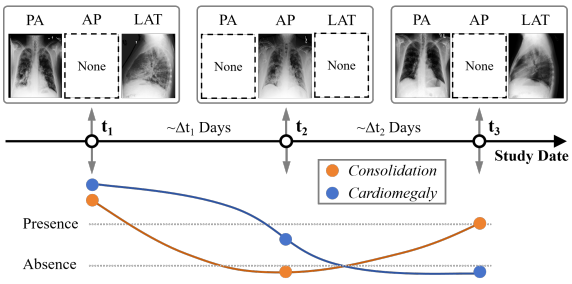

Figure 1: Illustration of longitudinal multi-view chest X-ray data with irregular visit intervals and incomplete view acquisitions. Each visit may contain different views, while disease states continuously evolve between irregularly spaced observations.

Traditional temporal models, such as Recurrent Neural Networks (RNNs) and Transformers, operate on discrete time steps and are therefore inherently limited in modeling continuous physiological dynamics. To address this limitation, Neural Ordinary Differential Equations (Neural ODEs) (Chen et al., 2018) provide a continuous-time formulation by parameterizing the evolution of hidden states via neural differential equations, enabling the modeling of continuous-time evolution and capturing long-term dependencies. While Neural ODEs have been widely used for modeling temporal dynamics (Qiu et al., 2023; 2025; Wang et al., 2025; Chan et al., 2026; Gao et al., 2026a), most existing methods focus on low-dimensional time series data, as optimizing continuous dynamical systems over high-dimensional visual feature spaces in RRG remains challenging.

In this paper, we propose a novel **Co**ntinuous spatio-temporal **D**isease **E**volution framework for longitudinal **R**adiology **R**eport **G**eneration (CoDE-RRG). Specifically, to extract stable and comprehensive features from an inconsistent number of multi-view images, we perform an adaptive fusion strategy that uses a primary view as an anchor to integrate information from all views. To model continuous disease evolution, we project high-dimensional visual features into a low-dimensional disease-specific space, enabling Neural ODEs to model continuous trajectories, while Graph Convolutional Networks (GCNs) further capture inter-disease interactions during evolution. Finally, the evolved disease states, together with current multi-view visual features and clinical context, are projected and fed as prompts into a pre-trained language model (PLM) to generate accurate and longitudinally informed reports.

Our contributions can be summarized as follows:

- We propose CoDE-RRG, a continuous spatio-temporal framework that leverages irregular longitudinal information to enhance radiology report generation by learning disease evolution from clinical timelines.
- For irregular time intervals, we proposed disease-guided dimensionality reduction to project high-dimensional visual features into a low-dimensional disease-specific state space and leverage Neural ODEs and GCNs to model continuous disease evolution trajectories. To our knowledge, this is the first work to explore Neural ODE-based continuous-time modeling for RRG.
- For irregular view availability, we introduce an adaptive multi-view fusion strategy that uses a primary view as an anchor to integrate complementary information from available views, yielding stable multi-view representations.
- Extensive experiments on MIMIC-CXR and Longitudinal-MIMIC datasets demonstrate the effectiveness of CoDE-RRG.

2 Related Work

Most longitudinal report generation methods primarily focus on modeling consistency and discrepancies between the current visit and recent historical visits. HC-LLM (Liu et al., 2025b) introduces a tri-consistency constraint mechanism to improve cross-visit multi-modal alignment. PriorRG (Liu et al., 2026) integrates clinical context guided contrastive learning with a coarse-to-fine decoding strategy. Diff-RRG (Yun et al., 2025) constructs disease difference maps to model temporal variations, while MARE (Gao et al., 2026b) introduces multi-modal analogical reasoning to model pathological evolution. However, these methods are built under a single-view longitudinal imaging setting, limiting the exploitation of complementary multi-view visual information. Recently, some methods start to explore multi-view information in longitudinal report generation. MLRG (Liu et al.,

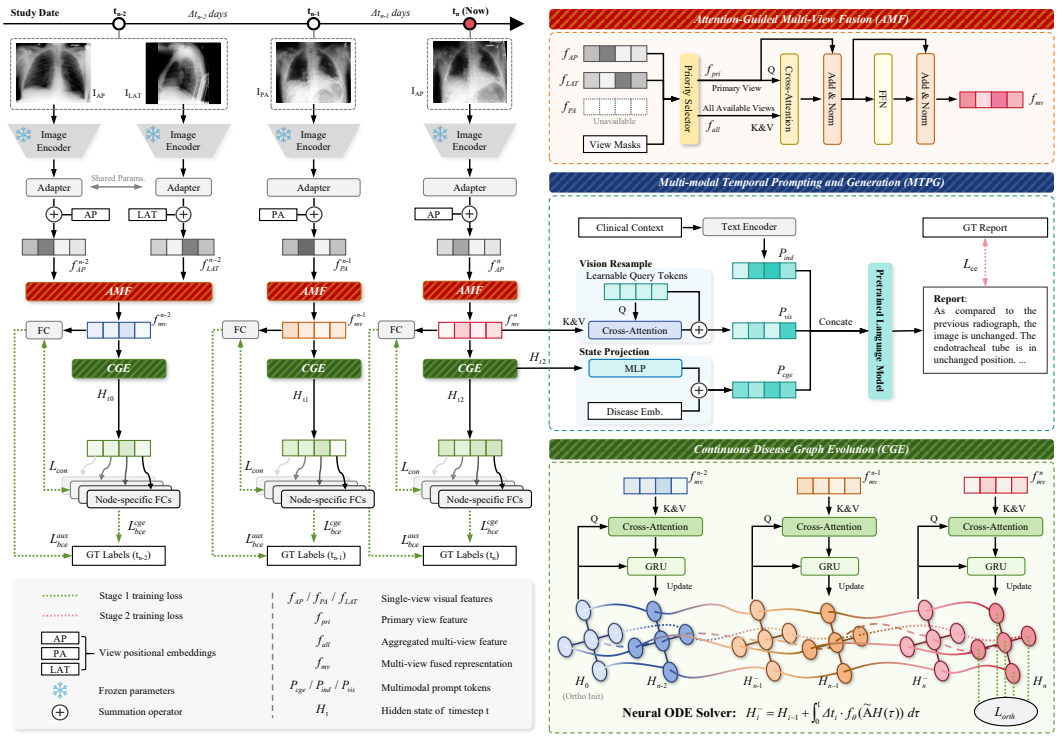

Figure 2: Overview of the proposed CoDE-RRG framework. It comprises three modules: attention-guided multi-view fusion (AMF), continuous disease graph evolution (CGE), and multi-modal temporal prompting and generation (MTPG), for multi-view integration, disease evolution modeling, and report generation, respectively.

2025a) explores multi-view contrastive pretraining to enhance longitudinal representation learning, and CXRMate (Nicolson et al., 2024) introduces reinforcement learning with clinical accuracy rewards to better align generation with multi-view longitudinal data. Despite these advances, these methods ignore long-range history and temporal irregularities between visits. Although recent methods (Wang et al., 2024; Yang et al., 2025) have started to exploit longer longitudinal contexts, their ability to model complex temporal dynamics remains limited, as they still model disease evolution as discrete transitions between visits.

3 Method

3.1 Overall

As shown in Figure 2, CoDE-RRG consists of three main modules: (1) attention-guided multi-view fusion (AMF), (2) continuous disease graph evolution (CGE), and (3) multi-modal temporal prompting and generation (MTPG). Given a patient, the input is a longitudinal sequence of visits $\mathcal{T} = \{t_0, t_1, \ldots, t_N\}$, where t_N is the current visit. Each visit t_i contains a variable number of multi-view images I^{t_i}. The model also takes irregular time intervals Δt between visits and the clinical context c_{t_N} at the current visit. CoDE-RRG aims to generate an context-aware and disease progression-oriented radiology report $\hat{y}$ for the current visit t_N by modeling the continuous longitudinal imaging history with the current multi-modal inputs.

3.2 Visual Features Extraction

We sequentially extract high-dimensional visual features from the multi-view images I^{t_i} for each visit in chronological order. Given the variable number of images per visit, we categorize all avail-

able images into three primary radiographic view types: Anteroposterior (AP), Posteroanterior (PA), and Lateral (LAT). For each view category, we retain a maximum of N_v images.

For each selected image in I^{t_i}, a frozen pre-trained image encoder extracts patch-level visual features $f_i \in \mathbb{R}^{N_p \times D}$, where N_p and D denote the number of patches and feature dimension, respectively. We then transform f_i into the task-specific feature space using a trainable adapter. Specifically, the features are first normalized and passed through a multi-layer perceptron (MLP), whose output is added residually to the original features, followed by another layer normalization (LN), yielding f_{adapt}. To encode view-specific information, we introduce a learnable embedding E_v for each view type $v \in \text{PA}, \text{AP}, \text{LAT}$ and add it to f_{adapt} to obtain view-aware features f_v. The resulting features are subsequently fed into AMF for multi-view fusion.

3.3 Attention-Guided Multi-View Fusion

To effectively aggregate multi-view images into a stable and unified representation, we propose the AMF module. Utilizing view masks to identify available images, the AMF module selects the highest-priority view as the primary view $f_{pri} \in \mathbb{R}^{N_p \times D}$ to aggregate complementary information from auxiliary views. In clinical practice, PA is the standard projection for chest radiography, followed by AP, while LAT typically serves as a supplementary view. Accordingly, we define the priority order as PA $>$ AP $>$ LAT. We then treat the primary view as the Query (Q), and the features of all available views as the Key (K) and Value (V):

$$Q = f_{pri}, \quad K = V = \text{Concat}(f_{PA}, f_{AP}, f_{LAT}), \tag{1}$$

where missing views are excluded via a binary key padding mask and $K, V \in \mathbb{R}^{3N_p \times D}$. The cross-attention output is refined using residual connections, layer normalization, and a feed-forward network (FFN). The resulting representation $f_{mv} \in \mathbb{R}^{N_p \times D}$ integrates both primary anatomical structures and complementary diagnostic cues.

3.4 Continuous Disease Graph Evolution

Graph Definition and Initialization. To model disease states at each time point while capturing relationships among clinical findings, we formulate a patient's condition as a disease graph, denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.The node set $\mathcal{V}$ consists of $N_c = 13$ semantic nodes, each corresponding to a clinical finding predefined in CheXbert (Smit et al., 2020). The edge set $\mathcal{E}$ represents the dependencies and co-occurrence patterns among these findings. We instantiate a learnable adjacency matrix $\mathbf{A} \in \mathbb{R}^{N_c \times N_c}$ to capture these relationships.

Before incorporating temporal progression, the graph requires an initial hidden state at the pre-visit baseline. We define this sequence-agnostic state as $H_0 \in \mathbb{R}^{N_c \times d_n}$, where $d_n = 32$ denotes the node embedding dimension. We use a relatively low-dimensional state space to avoid computational and optimization challenges of modeling high-dimensional visual dynamics with Neural ODEs. Furthermore, we adopt an orthogonal initialization strategy for H_0 to ensure decoupled initial features across disease nodes.

Continuous Trajectory Evolution. To trace these latent physiological trajectories, we formulate the temporal evolution of diseases as an Initial Value Problem (IVP). Specifically, the explicitly updated graph state $H_{t_{i-1}}$ from the previous visit serves as the exact initial value, and our objective is to predict the prior hidden state $H_{t_i}^-$, which represents the continuous physiological state computationally evolved over Δt_i. The continuous evolution is modeled by a neural ODE parameterized by f_θ. To enable information exchange among the N_c disease nodes during temporal evolution, we incorporate a graph convolutional mechanism into the dynamics. The learnable adjacency matrix $\mathbf{A}$ is normalized via Softmax to obtain $\tilde{\mathbf{A}}$, which defines message passing over disease nodes. We define the ODE dynamics over a normalized continuous time variable $\tau \in [0, 1]$ as:

$$\frac{dH(\tau)}{d\tau} = f_\theta\big(\tilde{\mathbf{A}} H(\tau)\big), \tag{2}$$

where f_θ is implemented as a two-layer MLP with $\tanh$ activation. For irregular time intervals Δt_i between visits, we map the physical time interval $[t_{i-1}, t_i]$ to the normalized domain via $t = t_{i-1} + \tau \Delta t_i$. Accordingly, the hidden state at time t_i is obtained by solving:

$$H_{t_i}^- = \text{ODESolve}\big(H_{t_{i-1}}, f_\theta, (0, 1)\big) = H_{t_{i-1}} + \int_0^1 \Delta t_i \cdot f_\theta(\tilde{\mathbf{A}} H(\tau)) d\tau. \tag{3}$$

Visual Observation and State Update. To incorporate newly available longitudinal observations at each clinical visit, we update the prior hidden state $H_{t_i}^-$ predicted by the Neural ODE with current visual evidence. Specifically, the prior state $H_{t_i}^-$ serves as disease-specific queries that interact with visual features $f_{mv}^{t_i}$ via cross-attention. This enables each disease node to selectively retrieve relevant pathological evidence from visual inputs, producing disease-aware visual embeddings $E_{t_i} \in \mathbb{R}^{N_c \times d_n}$.

Subsequently, to fuse the historical evolution trends with the current observations, we employ a Gated Recurrent Unit (GRU) cell. The newly extracted visual evidence E_{t_i} serves as the current input, while the prior predicted state $H_{t_i}^-$ acts as the previous hidden state. This results in the posterior updated state H_{t_i}, which reflects the patient's true condition at visit t_i.

Algorithm 1 Overall procedure of the CGE module.

Input: Comprehensive multi-view visual features from AMF module with timestamps $\{(f_{mv}^{t_i}, t_i)\}_{i=1..N}$

$H_0 = \text{OrthogonalInit}()$
for i in $1, 2, \ldots, N$ **do**
  if $i = 1$ **then**
    $E_{t_i} = \text{CrossAttention}(H_0, f_{mv}^{t_i})$
    $H_{t_i} = \text{GRU}(E_{t_i}, H_0)$
  else
    $H_{t_i}^- = \text{ODESolve}\big(H_{t_{i-1}}, f_\theta, (0, 1)\big)$
    $E_{t_i} = \text{CrossAttention}(H_{t_i}^-, f_{mv}^{t_i})$
    $H_{t_i} = \text{GRU}(E_{t_i}, H_{t_i}^-)$
  end if
end for
Return H_{t_N}

By alternating continuous evolution and discrete observation updates, our framework constructs an individualized disease progression trajectory (see Algorithm 1). The process continues until t_N, where the final state H_{t_N} encodes longitudinal disease dynamics and is fed into the MTPG module.

3.5 Multi-modal Temporal Prompting and Generation

To integrate multi-view visual information, temporal disease dynamics, and clinical context, we propose the MTPG module, which constructs three complementary prompts to condition the PLM. Clinical context provides diagnostic focus by encoding patient indications and medical history, while visual and temporal prompts supply complementary imaging and progression information. Specifically, P_{ind} is obtained by encoding clinical text via the PLM text encoder; P_{vis} is generated by a vision resampler that cross-attends to the fused multi-view feature $f_{mv}^{t_N}$ using learnable queries to produce a compact representation of salient abnormalities; and P_{cge} is derived by aggregating the final hidden state H_{t_N} from the CGE module with disease embeddings and projecting them via an MLP into the PLM's embedding space to model disease progression. Finally, these three prompt sequences are concatenated to form a comprehensive multi-modal context P_{total}, and then fed as the prefix prompt into the PLM to generate the target radiology report $\hat{y}$.

3.6 Model Training

Stage 1. To ensure robust multi-view feature learning and clinically meaningful disease evolution modeling, we formulate Stage 1 training as a multi-label classification task with auxiliary supervision and regularization. Each CGE node is equipped with a node-specific classification head. The CGE and auxiliary classification losses are defined using multi-label Binary Cross-Entropy (BCE)

as follows:

$$\mathcal{L}_{bce}^{cge} = \frac{1}{\sum_{b,s} M_{b,s}} \sum_{b,s:M_{b,s}=1} \sum_{c=1}^{N_c} \mathrm{BCE}\Big(\hat{y}_{b,s,c}^{(cge)}, y_{b,s,c}\Big),$$
$$\mathcal{L}_{bce}^{aux} = \frac{1}{\sum_{b,s} M_{b,s}} \sum_{b,s:M_{b,s}=1} \sum_{c=1}^{N_c} \mathrm{BCE}\Big(\hat{y}_{b,s,c}^{(aux)}, y_{b,s,c}\Big). \tag{4}$$

Here, b and s denote the patient and visit indices, respectively, and $M_{b,s} \in \{0, 1\}$ is a validity mask indicating whether the corresponding visit is valid. To stabilize training and encourage agreement between the CGE and auxiliary branches, we introduce a consistency regularization term that aligns their predictive distributions:

$$\mathcal{L}_{con} = \frac{1}{\sum_{b,s} M_{b,s}} \sum_{b,s:M_{b,s}=1} \left\| p_{b,s}^{(cge)} - p_{b,s}^{(aux)} \right\|_2^2. \tag{5}$$

To mitigate over-smoothing in message passing and encourage disease-specific disentanglement, we impose an orthogonality constraint on normalized node representations:

$$\mathcal{L}_{orth} = \frac{1}{\sum_{b,s} M_{b,s}} \sum_{b,s:M_{b,s}=1} \frac{1}{N_c(N_c-1)} \sum_{i \neq j} \left| \tilde{H}_{b,s,i}^{\top} \tilde{H}_{b,s,j} \right|. \tag{6}$$

The overall objective in Stage 1 training is:

$$\mathcal{L}_{s1} = \mathcal{L}_{bce}^{cge} + \lambda_1 \mathcal{L}_{bce}^{aux} + \lambda_2 \mathcal{L}_{con} + \lambda_3 \mathcal{L}_{orth}. \tag{7}$$

Stage 2. In Stage 2, we fine-tune the PLM for report generation conditioned on multi-modal temporal prompts, while updating only the parameters of the MTPG module. Let $\mathbf{P}$ denote the integrated prefix prompt sequence, and let $y_{t_N} = \{w_1, w_2, \ldots, w_L\}$ be the ground-truth radiology report at the current visit t_N. The generative objective is defined as:

$$\mathcal{L}_{s2} = -\sum_{t=1}^{L} \log p(w_t \mid \mathbf{P}, w_{<t}). \tag{8}$$

4 Experiments

4.1 Dataset

To evaluate the effectiveness of our proposed method, we conduct experiments on the following two public datasets: 1) **MIMIC-CXR** (Johnson et al., 2019a;b; Goldberger et al., 2000): A public dataset comprising 377,110 chest X-ray images and 227,835 corresponding reports. 2) **Longitudinal-MIMIC** (Zhu et al., 2023): A subset of the MIMIC-CXR dataset including patients who have multiple visits. For these two datasets, we both adopt the official data split.

4.2 Implementation details

Following (Liu et al., 2025a; 2026), we use Rad-DINO (Pérez-García et al., 2025) as the image encoder and DistilGPT2 (Sanh et al., 2019) as the PLM. We restrict the number of images per view to 1 and the number of visits to 5. During training, the loss weights are set to $\lambda_1 = 0.8$, $\lambda_2 = 0.01$ and $\lambda_3 = 0.01$ based on preliminary experiments. The model is trained with batch size 4 and learning rate $1e-4$ for both stages, for 5 epochs in stage 1 and 6 epochs in stage 2. During inference, we use beam search with size 3. Details are provided in Appendix B.

4.3 Baselines

We compare our method with several state-of-the-art (SOTA) baselines, which are categorized into three groups. 1) **SI**: Methods utilize only one single image from the current visit as input, including MAN (Shen et al., 2024), SEI (Liu et al., 2024b), MPO (Xiao et al., 2025), KIA (Yin et al., 2025), and D2Gen (Wu et al., 2026); 2) **LI**: Methods that leverage single images from longitudinal multiple

Table 1: Performance comparison on MIMIC-CXR (M-CXR) and Longitudinal-MIMIC (L-M) datasets. * denotes the reproduced results. (**Bold**: Best, Underline: Second Best)

Dataset	Input	Method	Venue	NLG Metrics ↑						CE Metrics ↑			
				B-1	B-2	B-3	B-4	MTR	R-L	RG	P	R	F1
M-CXR	SI	MAN	AAAI24	0.396	0.244	0.162	0.115	0.151	0.274	-	0.411	0.398	0.389
	SI	SEI*	MICCAI24	0.382	0.248	0.177	0.135	0.158	0.299	0.244	0.527	0.407	0.460
	SI	MPO	AAAI25	0.416	0.269	0.191	0.139	0.162	0.309	-	0.436	0.376	0.353
	SI	KIA	COLING25	0.415	0.274	0.187	0.138	0.167	0.307	-	0.504	0.425	0.461
	SI	D2Gen	AAAI26	**0.428**	0.272	0.187	0.131	-	0.232	-	-	-	-
	LI	HERGen*	ECCV24	0.367	0.226	0.150	0.105	0.146	0.277	0.209	0.557	0.433	0.487
	LI	PriorRG*	AAAI26	0.420	0.295	0.223	0.176	**0.186**	0.338	0.278	0.576	0.502	0.536
	MVL	CXRMate	IMU24	0.361	0.223	0.150	0.108	0.159	0.263	0.238	0.495	0.367	0.422
	MVL	STREAM	TMI25	0.422	0.271	0.188	0.137	0.165	0.290	0.219	0.527	0.412	0.462
	MVL	MLRG*	CVPR25	0.406	0.273	0.201	0.156	0.174	0.316	0.260	0.582	0.471	0.520
	MVL	**Ours**	/	0.427	**0.301**	**0.228**	**0.180**	0.185	**0.341**	**0.282**	**0.599**	**0.526**	**0.560**
L-M	LI	Prefilling	MICCAI23	0.343	0.210	0.140	0.099	0.137	0.271	-	0.538	0.434	0.480
	LI	HERGen*	ECCV24	0.369	0.227	0.150	0.106	0.146	0.276	0.207	0.563	0.438	0.493
	LI	Diff-RRG	MICCAI25	0.405	0.251	0.169	0.120	0.164	0.276	-	0.528	0.430	0.474
	LI	HC-LLM	AAAI25	0.406	0.260	0.178	0.127	0.162	0.285	-	0.415	0.358	0.360
	LI	MARE	AAAI26	0.409	0.265	0.184	0.133	0.161	0.291	-	-	-	-
	LI	PriorRG*	AAAI26	0.423	0.298	0.225	0.178	**0.187**	0.338	0.276	0.583	0.509	0.544
	MVL	MLRG*	CVPR25	0.405	0.272	0.200	0.155	0.174	0.315	0.257	0.587	0.477	0.526
	MVL	**Ours**	/	**0.431**	**0.304**	**0.231**	**0.182**	0.186	**0.342**	**0.282**	**0.607**	**0.535**	**0.569**

visits, including Prefilling (Zhu et al., 2023), HERGen (Wang et al., 2024), Diff-RRG (Yun et al., 2025), HC-LLM (Liu et al., 2025b), MARE (Gao et al., 2026b), and PriorRG (Liu et al., 2026); 3) **MVL**: Methods that incorporate multi-view images across longitudinal multiple visits, including CXRMate (Nicolson et al., 2024), STREAM (Yang et al., 2025), and MLRG (Liu et al., 2025a).

4.4 METRICS

We employ three categories of metrics: 1) **NLG metrics**, including BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), and ROUGE-L (Chin-Yew, 2004), to measure the lexical and syntactic alignment between the generated and ground-truth reports; 2) **CE metrics**, including CheXbert (Smit et al., 2020) and RadGraph (Jain et al., 2021), where the former evaluates diagnostic accuracy and the latter assesses the matching of clinical entities and their relations; and 3) an **LLM-based metric**, GREEN (Ostmeier et al., 2024), to detect factual errors and evaluate the alignment with expert preferences.

4.5 COMPARISON WITH SOTA METHODS

As shown in Table 1, our method achieves the best performance across almost all metrics on both datasets. On the MIMIC-CXR dataset, compared with SOTA SI method D2Gen, our method improves BLEU-4 from 0.131 to 0.180, showing that historical images provide complementary information. Compared with SOTA LI method PriorRG and MVL method MLRG, our method achieves overall superior performance. CoDE-RRG improves CheXbert-F1 by approximately 4.5% over PriorRG and 7.7% over MLRG. For BLEU-4, we achieve a substantial 15.4% relative improvement over MLRG. The advantages become more pronounced on the Longitudinal-MIMIC dataset. Compared to STREAM, which also exploits long-term history, our method achieves a 31.4% improvement in BLEU-4 and a 21.2% gain in CheXbert-F1 on the MIMIC-CXR dataset, demonstrating the effectiveness of modeling continuous disease evolution.

As shown in Table 2, we report four GREEN metrics: the overall GREEN score, Matches (correctly identified findings), MC (mentioned comparisons that are not in the reference report), and OC (omitted comparisons detailing a change from prior study). CoDE-RRG achieves the second-highest GREEN and Matches scores, slightly below PriorRG. This difference may arise from its tendency to retain historically positive findings, which can occasionally introduce findings not supported by the current reference. In addition, PriorRG's image-report contrastive pre-training may provide stronger capabilities in lesion localization and describing less common diseases. Nevertheless, CoDE-RRG achieves the lowest MC and OC scores, indicating fewer erroneous comparisons and better sen-

Table 2: LLM-based evaluation results on the MIMIC-CXR dataset. † denotes the results are sourced from Liu et al. (2025a). ♦ denotes non-longitudinal methods.

Method	GREEN ↑	Matches ↑	MC ↓	OC ↓
♦R2Gen†	0.283	1.803	-	-
♦CMN†	0.297	1.935	-	-
♦SEI*	0.276	1.859	-	-
HERGen*	0.308	2.014	0.104	0.142
MLRG*	0.341	2.032	0.102	0.102
PriorRG*	**0.368**	**2.292**	0.101	0.101
Ours	0.357	2.228	**0.094**	**0.093**

Table 3: Ablation study on the MIMIC-CXR dataset. $F1_{14}$ and $F1_5$ denote the CheXbert F1 scores evaluated on 14 observations and 5 key observations, respectively.

w/o	B-2	B-4	MTR	R-L	RG	$F1_{14}$	$F1_5$
PI	0.242	0.149	0.165	0.336	0.270	0.513	0.559
CC	0.219	0.097	0.146	0.263	0.209	0.551	0.603
$\mathcal{L}_{orth}$	0.286	0.172	0.178	0.339	0.280	0.521	0.569
$\mathcal{L}_{aux}$	0.201	0.112	0.133	0.286	0.212	0.172	0.134
GCN	0.296	0.174	0.181	0.335	0.277	0.538	0.587
AMF	0.258	0.154	0.163	0.324	0.252	0.415	0.442
Full	**0.301**	**0.180**	**0.185**	**0.341**	**0.282**	**0.560**	**0.610**

Table 4: Performance comparison of different backbone architectures on the MIMIC-CXR dataset.

Model	B-2	B-4	MTR	R-L	RG	$F1_{14}$	$F1_5$
RNN	0.292	0.174	0.181	0.338	0.277	0.547	0.589
RNN+T	0.296	0.176	0.182	0.339	0.281	0.548	0.596
Ours	**0.301**	**0.180**	**0.185**	**0.341**	**0.282**	**0.560**	**0.610**

Table 5: Performance comparison of different view configurations on the MIMIC-CXR multi-view subset.

Model	B-2	B-4	MTR	R-L	RG	$F1_{14}$	$F1_5$
Single	0.283	0.170	0.178	0.340	0.298	0.489	0.579
Dual	0.303	0.187	0.186	0.353	0.305	0.499	**0.600**
Full	**0.305**	**0.188**	**0.187**	**0.354**	**0.306**	**0.503**	0.596

sitivity to longitudinal changes, which further demonstrates its ability to capture temporal disease changes.

4.6 Ablation Study

Effect of prior images (PI). Table 3 shows that using only current-time images degrades all metrics, with BLEU-4 dropping from 0.180 to 0.149 and $F1_{14}$ from 0.560 to 0.513, indicating that PI provides crucial evidence for the current diagnosis.

Effect of clinical context (CC). As shown in Table 3, removing CC leads to a decrease in ROUGE-L from 0.341 to 0.263 and BLEU-4 from 0.180 to 0.097, indicating degraded semantic overlap and n-gram matching with reference reports. Moreover, the RadGraph score drops from 0.282 to 0.209, reflecting reduced ability to capture and model the key clinical entities and their relationships.

Effect of auxiliary signals. As shown in Table 3, removing the orthogonality constraint ($\mathcal{L}_{orth}$) leads to a drop in F1 scores, indicating that explicitly disentangling feature spaces for different diseases helps improve the model's ability to accurately identify individual disease states. Removing the auxiliary supervision terms ($\mathcal{L}_{aux}$), which consists of the auxiliary classification loss $\mathcal{L}_{bce}^{aux}$ and consistency regularization $\mathcal{L}_{con}$, results in a significant performance degradation across all metrics. This is mainly because the CGE module introduces a long temporal reasoning path with Neural ODE, leading to attenuated gradients during backpropagation and making it difficult to optimize both CGE and upstream modules.

Effect of backbone on the CGE module. As shown in Table 3, removing GCN in the CGE module degrades CE performance, as $F1_{14}$ dropping from 0.560 to 0.538 and $F1_5$ from 0.610 to 0.587, along with consistent declines in NLG metrics. This demonstrates that temporal graph propagation enhances the modeling of inter-disease interactions, leading to more accurate inference of disease states. To further investigate the differences between continuous and discrete temporal modeling, we replace the Neural ODE with RNN (i.e., GRU) and a RNN variant with time embeddings (RNN+T). As shown in Table 4, performance consistently degrades after switching to discrete-time modeling, with the vanilla RNN achieving the worst results in both NLG and CE metrics. This suggests that discrete-time modeling is less effective than continuous-time modeling in capturing disease state.

Effect of views. As shown in Table 3, removing AMF and restricting each visit to a single view consistently degrades performance, particularly clinical correctness ($F1_{14}$: 0.560 to 0.415). On a multi-view MIMIC-CXR subset, Table 5 shows that combining frontal and lateral views improves all metrics over single-view input, while using all available views performs best overall. These results confirm the benefit of multi-view information for clinically consistent report generation.

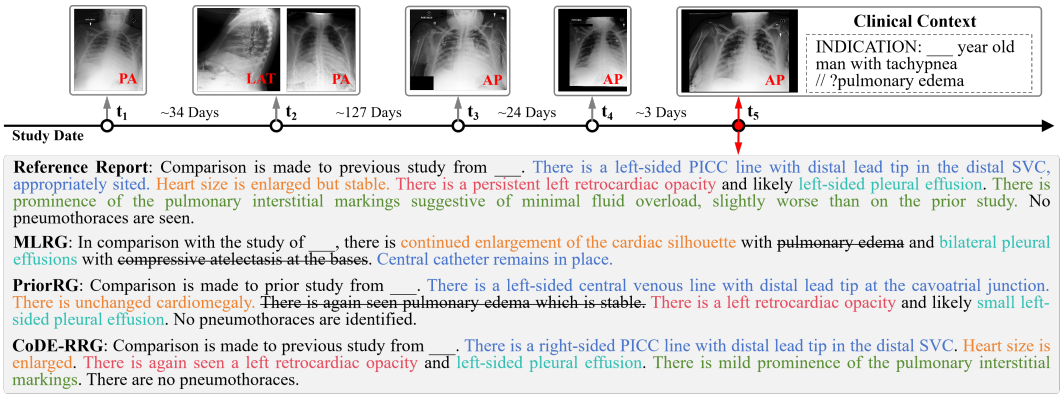

Figure 3: Illustration of the case study. Different colors are used to highlight distinct diseases.

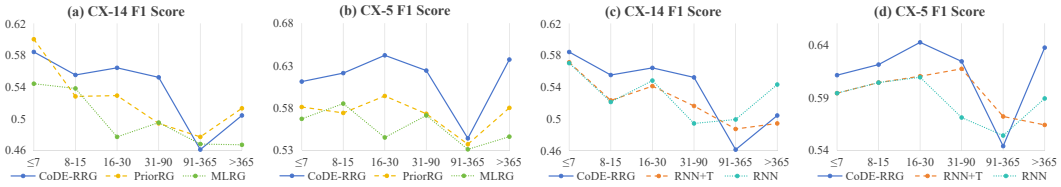

Figure 4: CheXbert F1 performance comparison across different temporal intervals between visits.

4.7 Case Study

Case in Figure 3 demonstrates the benefit of continuous-time modeling for longitudinal disease evolution. Our model preserves key persistent findings, including the left retrocardiac opacity and left-sided pleural effusion, and captures the prominence of the pulmonary interstitial markings. In contrast, baselines omit some of these findings and introduce unsupported interpretations such as pulmonary edema and atelectasis. These results highlight the advantage of continuous-time modeling in retaining longitudinal findings and reducing erroneous interpretations.

4.8 Temporal Interval Analysis

We further evaluate the model's performance under different time intervals between visits. As shown in Figure 4, all methods achieve comparable performance under short intervals ($\leq$ 7 days). As the temporal gap increases, the performance of baseline methods degrades significantly, while CoDE-RRG maintains more stable results. For the 91-365 days interval, CoDE-RRG exhibits a certain degree of performance fluctuation, which may be attributed to the more complex disease progression patterns in this period. However, under long intervals ($>$ 365 days), CoDE-RRG remains highly competitive, achieving strong performance on CX-5, which evaluates five key diseases. This performance remains close to its peak across all subsets. These results further validate that modeling disease progression as a continuous process is beneficial for handling irregular temporal intervals.

5 Conclusion

We propose CoDE-RRG, a continuous-time disease evolution framework for longitudinal RRG that models dynamic disease progression under irregular sampling radiology images. The method first adaptively fuses multi-view images at each time point to obtain unified visual representations. It then exploits the actual temporal trajectory of each patient to learn physiological priors for long-term disease evolution, where Neural ODEs and GCNs jointly characterize continuous disease dynamics along the real timeline. These representations are further integrated with clinical context to generate more temporally consistent radiology reports. Experiments demonstrate that our method more effectively models disease evolution under irregular temporal intervals and view sampling, thereby improving RRG accuracy.

AI USE STATEMENT

In this work, we used generative AI tools for design or provide feedback on research methodology or experiments, and assist with translation. We have not used generative AI tools to generate synthetic datasets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, implement methods, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for sourcing/searching for information, identifying relevant literature, drafting parts of the research paper, and editing the research paper to improve readability. We have reviewed all AI-assisted work. AI-assisted suggestions for research methodology and experiments were critically evaluated by the authors and verified through the corresponding literature, implementations, and experiments where applicable. AI-assisted manuscript text was reviewed and revised by the authors to ensure its accuracy and consistency with the actual work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study utilizes the publicly available MIMIC-CXR database (Johnson et al., 2019a). The dataset was collected with approval from the institutional review boards and has been thoroughly de-identified in compliance with the US Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor requirements, ensuring that all protected health information (PHI) is removed. The authors accessed and analyzed the data in accordance with the PhysioNet Credentialed Health Data Use Agreement.

REPRODUCIBILITY STATEMENT

We provide sufficient details to facilitate the reproducibility of our work. The proposed method and its major components are described in Section 3, together with the corresponding formulations and training objectives. Implementation and training details, including model configurations, optimization settings, and hyperparameter settings, are provided in Section 4.2 and Appendix B. Detailed descriptions of the datasets, data preprocessing, and experimental protocols are provided in Appendix A. The source code is available at `https://anonymous.4open.science/r/CoDE-RRG-E007/`.

REFERENCES

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, pp. 65–72, 2005.

Julian Chan, Alessia Gualandris, and Payel Das. Continuous-time modelling of black hole binary evolution with neural odes. *Monthly Notices of the Royal Astronomical Society*, pp. stag135, 2026.

Ricky T. Q. Chen. torchdiffeq, 2018. URL `https://github.com/rtqichen/torchdiffeq`.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 1439–1449, 2020.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pp. 5904–5914, 2021.

Lin Chin-Yew. ROUGE: A package for automatic evaluation of summaries. In *Proc. Text Summarization Branches Out*, pp. 74–81, 2004.

Mengzhou Gao, Huangqian Yu, and Pengfei Jiao. Mlgo: Multi-layer graph neural odes for traffic forecasting. *Neural Networks*, pp. 108540, 2026a.

Qingqing Gao, Tengfei Liu, Xiaoyan Li, Xiaodan Zhang, Zhongfan Sun, Boyue Wang, Baocai Yin, and Zhaohui Liu. Mare: Multimodal analogical reasoning for disease evolution-aware radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21180–21188, 2026b.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, Physiotoolkit, and Physionet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220., 2000. RRID:SCR_007345.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664, 18–24 Jul 2021.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis Langlotz, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports. In *Adv. neural inf. proces. syst.*, 2021.

Alistair Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. MIMIC-CXR database, 2019a. (version 2.0.0) *PhysioNet*. RRID:SCR_007345.

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*, 6(1):317, 2019b.

Hongzhao Li, Hongyu Wang, Xia Sun, Hua He, and Jun Feng. Context-enhanced framework for medical image report generation using multimodal contexts. *Knowledge-Based Systems*, 310: 112913, 2025. ISSN 0950-7051.

Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. Bootstrapping large language models for radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18635–18643, 2024a.

Kang Liu, Zhuoqi Ma, Xiaolu Kang, Zhusi Zhong, Zhicheng Jiao, Grayson Baird, Harrison Bai, and Qiguang Miao. Structural entities extraction and patient indications incorporation for chest x-ray report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 433–443, 2024b.

Kang Liu, Zhuoqi Ma, Xiaolu Kang, Yunan Li, Kun Xie, Zhicheng Jiao, and Qiguang Miao. Enhanced contrastive learning with multi-view longitudinal data for chest x-ray report generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10348–10359, 2025a.

Kang Liu, Zhuoqi Ma, Zikang Fang, Yunan Li, Kun Xie, and Qiguang Miao. Priorrg: Prior-guided contrastive pre-training and coarse-to-fine decoding for chest x-ray report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7206–7214, 2026.

Tengfei Liu, Jiapu Wang, Yongli Hu, Mingjie Li, Junfei Yi, Xiaojun Chang, Junbin Gao, and Baocai Yin. Hc-llm: Historical-constrained large language models for radiology report generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 5595–5603, 2025b.

Aaron Nicolson, Jason Dowling, Douglas Anderson, and Bevan Koopman. Longitudinal data and a semantic similarity reward for chest x-ray report generation. *Informatics in Medicine Unlocked*, 50:101585, 2024.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, et al. Green: Generative radiology report evaluation and error notation. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 374–390, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proc. Annu. Meet. Assoc. Comput Linguist.*, pp. 311–318, 2002.

Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Matthew P Lungren, et al. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7(1):119–130, 2025.

Xihe Qiu, Shaojie Shi, Xiaoyu Tan, Chao Qu, Zhijun Fang, Hailing Wang, Yongbin Gao, Peixia Wu, and Huawei Li. Gram-based attentive neural ordinary differential equations network for video nystagmography classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 21339–21348, 2023.

Xihe Qiu, Shaojie Shi, Bin Li, Xiaoyu Tan, Yongbin Gao, and Shuo Li. Llm-gaode: Large-language-model augmented neural ordinary differential equation network for video nystagmography classification. *Knowledge-Based Systems*, pp. 114050, 2025.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS EMC*$^2 Workshop$, 2019.

Hongyu Shen, Mingtao Pei, Juncai Liu, and Zhaoxing Tian. Automatic radiology reports generation via memory alignment network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4776–4783, 2024.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, pp. 1500–1519, 2020.

Fuying Wang, Shenghui Du, and Lequan Yu. Hergen: Elevating radiology report generation with longitudinal data. In *European Conference on Computer Vision*, pp. 183–200. Springer, 2024.

Xiaoda Wang, Yuji Zhao, Kaiqiao Han, Xiao Luo, Sanne van Rooij, Jennifer Stevens, Lifang He, Liang Zhan, Yizhou Sun, Wei Wang, and Carl Yang. Conditional neural ode for longitudinal parkinson's disease progression forecasting. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 4131–4134. IEEE, 2025.

Puzhen Wu, Hexin Dong, Yi Lin, Yihao Ding, and Yifan Peng. A disease-aware dual-stage framework for chest x-ray report generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 40, pp. 33953–33961, 2026.

Ting Xiao, Lei Shi, Peng Liu, Zhe Wang, and Chenjia Bai. Radiology report generation via multi-objective preference optimization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 8664–8672, 2025.

Yan Yang, Xiaoxing You, Ke Zhang, Zhenqi Fu, Xianyun Wang, Jiajun Ding, Jiamei Sun, Zhou Yu, Qingming Huang, Weidong Han, et al. Spatio-temporal and retrieval-augmented modelling for chest x-ray report generation. *IEEE Transactions on Medical Imaging*, 2025.

Heng Yin, Shanlin Zhou, Pandong Wang, Zirui Wu, and Yongtao Hao. Kia: Knowledge-guided implicit vision-language alignment for chest x-ray report generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 4096–4108, 2025.

Hannah Yun, Junyeong Maeng, Eunsong Kang, and Heung-Il Suk. Diff-rrg: Longitudinal disease-wise patch difference as guidance for llm-based radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 152–161. Springer, 2025.

Qingqing Zhu, Tejas Sudharshan Mathai, Pritam Mukherjee, Yifan Peng, Ronald M Summers, and Zhiyong Lu. 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.

APPENDIX

A DATA PROCESSING

A.1 LONGITUDINAL SEQUENCE CONSTRUCTION

For patients with closely spaced visits, we first sort all studies according to their acquisition timestamps. We then examine the temporal intervals between consecutive studies. If the interval between two consecutive studies is zero, we introduce a small positive temporal offset to distinguish these visits and repeat this procedure until no zero temporal intervals remain.

For each target report generation task, we construct a longitudinal sequence consisting of the current visit and historical records. Specifically, we retrieve up to four previous visits before the current time point. If fewer than four historical visits are available, zero-padding is applied to the missing temporal steps and corresponding longitudinal representations to maintain a fixed sequence length. For patients with more than four historical visits, only the four most recent records are retained. Therefore, each input sequence contains five temporal steps, including four historical visits and the current visit.

A.2 CLINICAL CONTEXT CONSTRUCTION

For patient clinical context modeling, we follow the preprocessing strategy in Li et al. (2025). Specifically, we extract three clinically relevant fields from the textual information in MIMIC-CXR (Johnson et al., 2019a), including "indication", "history", and "reason for the exam". These fields are combined as the clinical context associated with each visit and are used as additional patient-level information for model learning.

A.3 LABEL CONSTRUCTION

In the first-stage training, we utilize disease classification supervision derived from the radiology reports in the MIMIC-CXR dataset. Specifically, we adopt the labels generated by CheXbert (Smit et al., 2020), which extracts disease labels from free-text reports. The original CheXbert outputs represent each finding with four possible states: positive (1), negative (0), uncertain (-1), and blank (not mentioned). To formulate binary classification targets, we merge the positive and uncertain states as positive findings, while treating negative and blank states as negative findings. Therefore, the original labels are converted into binary ground-truth labels.

Finally, we exclude the "No Finding" label, resulting in a final target set of 13 categories defined as:

1. Atelectasis
2. Cardiomegaly
3. Consolidation
4. Edema
5. Enlarged Cardiomediastinum
6. Fracture
7. Lung Lesion
8. Lung Opacity
9. Pleural Effusion
10. Pleural Other
11. Pneumonia

Table 6: Statistics of images used for view classifier training.

Split	AP	PA	LAT	Total
# Train	143,556	94,416	115,647	353,619
# Validation	1,212	747	908	2,867
# Test	2,405	998	1,431	4,834
Total	147,173	96,161	117,986	361,320

Table 7: Distribution of projection views in the MIMIC-CXR dataset used for our experiments.

Split	AP	PA	LAT	Total
# Train	86,705	82,744	101,324	270,773
# Validation	687	649	794	2,130
# Test	1,761	851	1,246	3,858
Total	89,153	84,244	103,364	276,761

12. Pneumothorax
13. Support Devices

A.4 Multi-view Data Preprocessing

Since our method requires multi-view chest X-ray inputs, complete view information is essential for subsequent multi-view feature aggregation. However, a portion of images in the MIMIC-CXR dataset lack explicit view annotations. To recover the missing view information, we first utilize images with available view labels to train a view classifier and then apply the trained classifier to infer the views of images with missing annotations.

Specifically, we follow the official data split of MIMIC-CXR to divide the dataset into training, validation, and test sets. Images with reliable view annotations are selected from each split to construct the training data for the view classifier. The distribution of images used for view classifier training is summarized in Table 6. We train a view classifier based on ResNet-18. The classifier consists of a ResNet-18 backbone followed by a fully connected classification layer, which predicts three commonly used chest X-ray views: AP, PA, and LAT. During training, we apply standard data augmentation techniques, including resizing images to 256×256 pixels, randomly cropping regions of size 224×224. The classifier is trained for 3 epochs with a learning rate of 1×10^{-4}. The final classification accuracy reaches 99.39%.

After training, the classifier is used to predict the missing view labels of images without explicit annotations, enabling the construction of complete multi-view inputs.

A.5 Dataset Construction

We first remove examinations without corresponding textual reports. During training and validation, we treat each patient as one sample, where the sample contains the complete longitudinal sequence of the patient's previous visits. This patient-level sampling strategy prevents repeatedly using multiple visits from the same patient as independent training samples, thereby reducing the risk of overfitting. During testing, we treat each visit as one sample to ensure that the sampling protocol is consistent with previous methods. The resulting dataset contains 59,799, 459, and 2,347 samples for training, validation, and testing, respectively. The distribution of different projection views in the constructed dataset is summarized in Table 7.

B Additional Implementation Details

B.1 Vision Resampler Configuration

In the MTPG module, we employ a vision resampler (Jaegle et al., 2021) to reduce the number of visual tokens while preserving informative visual representations. Specifically, the vision resampler introduces 128 learnable query tokens, which interact with image features through a multi-head cross-attention mechanism. Given the input visual features, the query tokens are used as queries, while the image features serve as keys and values. The resulting attended representations are obtained by adding the cross-attention outputs to the original learnable query tokens, producing a compact set of 128 visual tokens for subsequent processing.

Table 8: Summary of baseline methods and our approach in terms of auxiliary input data and maximum longitudinal visits.

Method	Auxiliary Inputs				Max Visits
	Prior Image	Prior Report	Retrieval Data	Clinical Context	
MAN					1
SEI				✓	1
MPO					1
KIA					1
D2Gen			✓		1
HERGen	✓	✓			5
PriorRG	✓			✓	2
CXRMate	✓	✓			2
STREAM	✓		✓		3
MLRG	✓	✓		✓	2
Prefilling	✓	✓			2
Diff-RRG	✓	✓			2
HC-LLM	✓				2
MARE	✓	✓			2
Ours	✓			✓	5

B.2 Baseline Configurations

Table 8 summarizes the key characteristics of all evaluated baseline methods in our main text, detailing their auxiliary input data and maximum longitudinal visit capacities:

1. **Auxiliary Inputs:** Whether a method relies on additional information during generation, such as ground-truth prior text report, clinical context, or external retrieval data.
2. **Max Visits:** The maximum number of longitudinal patient visits the architecture can accommodate (where 1 indicates single-visit methods and > 1 denotes longitudinal multi-visit context).

B.3 Evaluation Settings

During testing, we do not impose any explicit constraints on the generated report length, except for a maximum decoding length of 512 tokens to prevent excessively long outputs. The same inference setting is consistently applied to all reproduced baseline methods for fair comparison. For methods that generate reports based on a single image, we randomly select one image from each visit as the representative input for report generation, ensuring that each visit is treated as one report generation sample.

B.4 ODE Solver Configuration

To model continuous-time dynamics, we employ the adaptive-step Dormand-Prince solver (dopri5) implemented in the `torchdiffeq` package (Chen, 2018). We adhere to the library's default tolerance configurations across all training and evaluation stages, specifically setting the relative tolerance (rtol) to 1×10^{-7} and the absolute tolerance (atol) to 1×10^{-9}. During training, gradients are computed via direct backpropagation through the solver's computational graph.

C Additional Experimental Results

C.1 Detailed GREEN Results

GREEN (Generative Radiology Report Evaluation and Error Notation) (Ostmeier et al., 2024) is a factual consistency metric designed to evaluate generated radiology reports by identifying clinically meaningful discrepancies between candidate reports and reference reports. Different from

Table 9: Detailed GREEN evaluation results on the test set. The error categories (a)–(f) correspond to different types of factual inconsistencies. † denotes the results are sourced from MLRG (Liu et al., 2025a). * denotes the reproduced results.

Method	GREEN Score ↑	(a) ↓	(b) ↓	(c) ↓	(d) ↓	(e) ↓	(f) ↓	# Errors ↓	# Matched Findings ↑
R2Gen†	0.283	1.310	3.089	0.103	0.201	0.082	0.142	4.926	1.803
CMN†	0.297	1.383	2.963	0.127	0.228	0.081	0.161	4.942	1.935
SEI*	0.276	1.507	2.784	0.184	0.269	0.118	0.118	4.981	1.859
HERGen*	0.308	1.264	2.802	0.160	0.306	0.104	0.142	4.778	2.014
MLRG*	0.341	1.364	2.311	0.205	0.279	0.102	0.102	4.364	2.032
PriorRG*	0.368	1.481	2.300	0.241	0.329	0.101	0.101	4.552	2.292
Ours	0.357	1.582	2.345	0.283	0.315	0.094	0.093	4.713	2.228

traditional lexical similarity metrics, GREEN focuses on clinical correctness by explicitly modeling factual errors and matched findings between the generated and reference reports. Specifically, GREEN categorizes factual errors into six types:

(a) False report of a finding in the candidate.

(b) Missing a finding present in the reference.

(c) Misidentification of a finding's anatomic location/position.

(d) Misassessment of the severity of a finding.

(e) Mentioning a comparison that is not present in the reference.

(f) Omitting a comparison detailing a change from a prior study.

For each generated report, GREEN counts the number of errors in each category and identifies the matched findings, which represent clinically consistent findings shared between the candidate and reference reports. The final GREEN score jointly considers factual errors and matched findings, where fewer errors and more matched findings indicate better factual consistency.

As shown in Table 9, CoDE-RRG achieves competitive performance compared with existing approaches (Chen et al., 2020; 2021; Liu et al., 2024b; Wang et al., 2024; Liu et al., 2025a; 2026). It is worth noting that R2Gen, CMN, and SEI are designed for single-visit report generation rather than longitudinal report generation, making comparison-related categories (e) and (f) less applicable to these methods. Although PriorRG obtains the highest overall GREEN score, CoDE-RRG demonstrates advantages in comparison-related factual errors, achieving the lowest error rates in categories (e) and (f) among longitudinal report generation approaches. This indicates that the proposed longitudinal reasoning mechanism helps preserve temporal consistency when generating reports involving comparisons with prior studies.

C.2 Role of Real Physical Time

To evaluate the impact of temporal intervals on modeling continuous disease progression via Neural ODEs, we conduct an ablation study comparing three time interval configurations in CoDE-RRG. In our default model, Real Time, the exact physical time intervals recorded between consecutive longitudinal visits are utilized. For Standard Time, we replace all temporal gaps with a fixed constant value, which is computed as the median time interval across all adjacent visit pairs in the dataset. In contrast, for Random Time, we randomly permute the ground-truth time intervals across all samples in the dataset.

As shown in Table 10, CoDE-RRG (Real Time) consistently outperforms both the Random Time and Standard Time variants across all NLG metrics and CE metrics. This confirms that exact physical temporal intervals provide critical context for Neural ODEs to model continuous disease progression accurately. Interestingly, assigning random time intervals yields slightly better results than fixing intervals to a constant value (e.g., $F1_5$ of 0.592 vs. 0.579). A plausible explanation is that random intervals may retain certain temporal variance during trajectory integration via Neural ODEs, which could help avoid the rigid bias introduced by static steps.

Table 10: Ablation study on the impact of time interval in CoDE-RRG for longitudinal report generation. **Standard** replaces all intervals with the fixed dataset median. **Random** randomly shuffles ground-truth intervals across samples. **Real** uses the exact physical intervals between consecutive visits.

Model	B-2	B-4	MTR	R-L	RG	$F1_{14}$	$F1_5$
Standard	0.294	0.172	0.179	0.332	0.274	0.533	0.579
Random	0.295	0.174	0.181	0.337	0.279	0.540	0.592
Real	**0.301**	**0.180**	**0.185**	**0.341**	**0.282**	**0.560**	**0.610**

Table 11: Distribution of test samples across different time intervals.

Time Interval	Sample Size	Proportion
$\leq$ 7 days	1,020	48.41%
7–15 days	166	7.88%
15–30 days	216	10.25%
30–90 days	283	13.43%
90–365 days	320	15.19%
$>$ 365 days	102	4.84%
Total	**2,107**	**100.00%**

Table 12: Performance comparison across different time intervals.

Time Interval	Model	B-1	B-2	B-3	B-4	MTR	R-L	$F1_{14}$	RG	$F1_5$
$\leq$ 7 days	MLRG	0.396	0.268	0.198	0.154	0.172	0.314	0.544	0.247	0.567
	PriorRG	0.414	0.292	0.220	0.174	0.185	0.337	0.600	0.266	0.581
	RNN	0.418	0.293	0.220	0.173	0.181	0.332	0.570	0.258	0.594
	RNN+T	0.430	0.299	0.223	0.174	0.183	0.331	0.571	0.263	0.594
	CoDE-RRG	0.431	0.301	0.226	0.177	0.185	0.333	0.584	0.266	0.611
7–15 days	MLRG	0.409	0.269	0.193	0.147	0.173	0.313	0.538	0.269	0.585
	PriorRG	0.419	0.294	0.221	0.175	0.187	0.331	0.528	0.275	0.574
	RNN	0.405	0.283	0.213	0.167	0.179	0.331	0.521	0.280	0.604
	RNN+T	0.398	0.276	0.207	0.162	0.176	0.334	0.523	0.271	0.604
	CoDE-RRG	0.413	0.282	0.210	0.164	0.177	0.326	0.555	0.272	0.621
15–30 days	MLRG	0.406	0.267	0.190	0.143	0.172	0.311	0.477	0.264	0.545
	PriorRG	0.417	0.287	0.213	0.166	0.183	0.333	0.529	0.283	0.594
	RNN	0.399	0.281	0.211	0.165	0.176	0.339	0.548	0.287	0.609
	RNN+T	0.409	0.287	0.217	0.171	0.180	0.344	0.541	0.295	0.610
	CoDE-RRG	0.412	0.292	0.222	0.176	0.182	0.346	0.564	0.293	0.642
30–90 days	MLRG	0.410	0.273	0.199	0.153	0.174	0.308	0.495	0.262	0.571
	PriorRG	0.422	0.294	0.222	0.175	0.184	0.336	0.494	0.281	0.573
	RNN	0.410	0.290	0.220	0.174	0.179	0.337	0.494	0.298	0.571
	RNN+T	0.414	0.295	0.226	0.181	0.181	0.345	0.516	0.305	0.617
	CoDE-RRG	0.427	0.304	0.233	0.187	0.185	0.349	0.552	0.301	0.624
90–365 days	MLRG	0.420	0.285	0.210	0.164	0.178	0.325	0.468	0.287	0.531
	PriorRG	0.428	0.300	0.227	0.180	0.188	0.338	0.477	0.297	0.537
	RNN	0.410	0.291	0.223	0.178	0.181	0.350	0.499	0.307	0.554
	RNN+T	0.403	0.286	0.218	0.174	0.179	0.352	0.487	0.308	0.572
	CoDE-RRG	0.415	0.295	0.227	0.182	0.181	0.354	0.461	0.310	0.544
$>$ 365 days	MLRG	0.461	0.327	0.249	0.201	0.201	0.357	0.467	0.315	0.546
	PriorRG	0.478	0.353	0.277	0.224	0.212	0.387	0.513	0.340	0.580
	RNN	0.444	0.332	0.264	0.217	0.201	0.389	0.543	0.331	0.589
	RNN+T	0.442	0.325	0.255	0.210	0.197	0.383	0.494	0.323	0.564
	CoDE-RRG	0.459	0.344	0.273	0.224	0.208	0.393	0.504	0.342	0.637

C.3 Detailed Results Across Different Time Intervals

Table 11 details the distribution of test samples across different time intervals between longitudinal visits. Table 12 provides the complete numerical results across different time intervals. We compare our proposed CoDE-RRG against recent SOTA baselines (MLRG and PriorRG) and two ablation variants (RNN and RNN+T).

Table 13: Performance comparison with different numbers of visits. P, R, and F1 represent precision, recall, and F1-score, respectively.

Visits	B-1	B-2	B-3	B-4	MTR	R-L	$F1_{14}$	R_{14}	P_{14}	RG	$F1_5$	R_5	P_5
5	0.427	0.301	0.228	0.180	0.185	0.341	**0.560**	**0.526**	0.599	0.282	**0.610**	**0.588**	0.634
4	0.420	0.298	0.228	0.181	0.187	0.349	0.553	0.515	0.598	0.290	0.588	0.544	0.640
3	0.424	0.302	0.230	0.183	0.188	**0.350**	0.558	0.521	**0.601**	0.291	0.592	0.549	**0.642**
2	**0.431**	**0.306**	**0.233**	**0.184**	**0.190**	0.349	0.558	0.522	0.599	**0.292**	0.594	0.553	0.641
1	0.407	0.287	0.217	0.172	0.182	0.336	0.534	0.497	0.576	0.274	0.576	0.525	0.638

C.4 Impact of the Number of Historical Visits

To investigate the impact of longitudinal context on report generation quality, Table 13 presents a detailed performance comparison across different sequence lengths of historical visits (ranging from 1 to 5). "Visits = 1" serves as the single-visit baseline without historical context.

Based on the experimental results, we select Visits $= 5$ as the default configuration for our model. Through a comprehensive consideration of both natural language fluency and diagnostic accuracy, we place greater emphasis on CE metrics to ensure clinical utility. As shown in Table 13, Visits $= 5$ achieves the highest performance on key CE metrics ($F1_{14} = 0.560$, $F1_5 = 0.610$), alongside the highest recall rates ($R_{14} = 0.526$, $R_5 = 0.588$). This indicates that a longer historical horizon enables the model to better capture longitudinal disease progression and effectively reduce false negatives.