LEARNING HOW, NOT JUST WHAT: EXPLICIT EVOLUTION MODELING FOR LONGITUDINAL RADIOLOGY REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Longitudinal radiology report generation (LRRG) aims to describe current findings and their temporal evolution by analyzing prior and current examinations. Existing methods largely rely on static comparisons to identify what has changed, but explicit modeling of how the disease evolves remains limited. To address this gap, we propose EvoRG, a framework for explicitly modeling disease evolution. We first aggregate visual features from both examinations using shared disease prototypes and Sinkhorn-based optimal transport. This alignment reduces interference from non-pathological spatial shifts and establishes semantically comparable disease-wise endpoints. Latent transitions between these endpoints are then parameterized as continuous geodesic trajectories on a joint visual–semantic product-sphere manifold. Logarithmic-map tangents encode evolution direction, while geodesic midpoints provide intermediate representations. Geodesic distances quantify evolution magnitude and modulate a product-of-experts module that fuses the midpoint and directed-movement descriptors. The resulting disease-specific transition tokens are used by a large language model to generate evolution-aware reports. Extensive experiments on Longitudinal-MIMIC show that EvoRG achieves state-of-the-art performance. Relative improvements over the strongest baseline for each metric are 12.5% in CIDEr and 9.6% in CheXbert F1.

1 INTRODUCTION

Radiology report generation (RRG) aims to automatically produce diagnostic reports from medical images and alleviate the reporting burden on radiologists. Substantial progress has been made in single-timepoint report generation (Chen et al., 2020; 2021; Wang et al., 2023). However, clinical assessment often requires comparisons between prior and current examinations to identify interval changes. For example, similar radiographic findings on current chest X-rays may reflect improvement or worsening depending on the prior examination. Accurately characterizing these interval changes is important for assessing treatment response and informing clinical management. This need motivates longitudinal radiology report generation (LRRG), which incorporates information from prior examinations to describe current findings and their interval changes.

Early LRRG methods integrate contextual information from prior and current examinations (Zhu et al., 2023; Bannur et al., 2023). More recent approaches explicitly extract temporal differences to represent interval changes (Song et al., 2025; Yun et al., 2025; Liu et al., 2026b). As illustrated in Figure 1 (top), these methods can detect localized changes such as emerging or resolving abnormalities. For abnormalities that persist across examinations, reports must also describe changes in severity or extent. Their continued presence may be accompanied by either improvement or worsening, which calls for more detailed descriptions of disease evolution. However, existing approaches primarily encode temporal information through static comparisons of endpoint features, leaving the transition between examinations implicit. Consequently, the resulting representations emphasize what has changed but provide limited characterization of how the disease evolves.

From the perspective of disease dynamics, pathological changes unfold continuously over time, whereas prior and current examinations provide only sparse observations of this evolution (Young et al., 2024). Endpoint comparisons summarize the net change between visits but leave the transition

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.

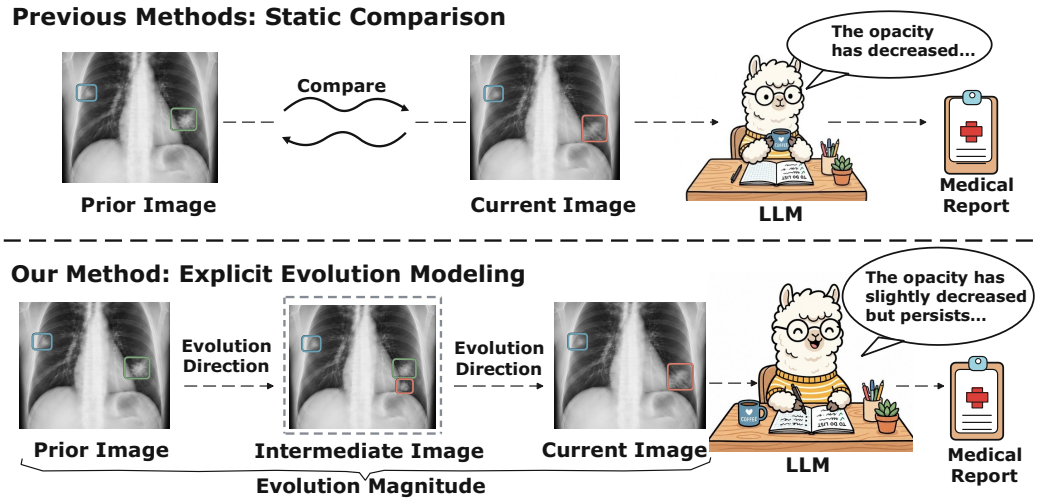


Figure 1: **From static comparison to explicit evolution modeling.** Existing methods rely on static endpoint comparisons to identify what has changed. Our method characterizes how the disease evolves through evolution direction, an intermediate representation, and evolution magnitude. The central image is a schematic visualization of the latent geodesic midpoint.

itself implicit. A continuous path in representation space can explicitly parameterize the transition between disease-specific endpoints. To characterize this path, its direction and length need to be defined and measured according to the geometry of that space. Riemannian geometry provides geodesic paths and associated operations for extracting complementary transition descriptors. As illustrated in Figure 1 (bottom), logarithmic-map tangents encode evolution direction, while geodesic midpoints provide intermediate representations. The corresponding geodesic distances quantify evolution magnitude (Miolane et al., 2020).

Based on this analysis, we propose EvoRG, a framework that moves longitudinal report generation from identifying what has changed to describing how the change occurs. We first aggregate visual features from prior and current examinations using shared disease prototypes and Sinkhorn-based optimal transport. This alignment reduces interference from non-pathological spatial shifts and yields semantically comparable disease-wise endpoints. Transitions between these endpoints are then parameterized as continuous geodesic trajectories on a joint visual-semantic product-sphere manifold. In this space, visual changes are modeled jointly with changes in predicted disease-label distributions. Midpoint and directed-movement descriptors are extracted from the trajectories. Finally, these descriptors are fused into disease-specific transition tokens within a distance-aware product-of-experts (PoE) module, with fusion weights modulated by geodesic distance. These tokens condition a large language model (LLM) for evolution-aware longitudinal report generation.

Our main contributions are summarized as follows:

- We propose EvoRG, an explicit disease evolution modeling framework for LRRG that moves from identifying what has changed to describing how the change occurs.
- We develop a cross-temporal disease alignment module using shared prototypes and Sinkhorn-based optimal transport, which produces semantically comparable disease-wise endpoints by mitigating the interference of non-pathological spatial shifts.
- We parameterize disease evolution as continuous geodesic trajectories on a joint visual-semantic product-sphere manifold. By extracting and fusing geometric descriptors into transition tokens, the model is effectively guided to capture how the disease evolves.
- Extensive experiments on the Longitudinal-MIMIC dataset show that EvoRG achieves state-of-the-art performance across both natural language generation and clinical efficacy metrics.

2 Related Work

2.1 Radiology Report Generation

Radiology report generation (RRG) aims to automatically generate clinically meaningful text reports from medical images. Early studies primarily used encoder–decoder architectures, mapping visual features to textual descriptions while integrating memory mechanisms to facilitate cross-modal interactions (Chen et al., 2020; 2021; Zhu et al., 2023). To improve report quality, recent studies have incorporated large language models (LLMs), structured clinical data, and external knowledge (Wang et al., 2023; Kim et al., 2025; Delbrouck et al., 2025; Zhang et al., 2026). Despite these developments, the majority of existing models operate in single-visit settings, neglecting the longitudinal comparisons that are vital for clinical decision-making. Because disease assessment naturally relies on temporal changes, a single image remains insufficient for comprehensively understanding disease evolution over time.

2.2 Longitudinal Radiology Report Generation

Longitudinal radiology report generation (LRRG) incorporates prior examinations to describe current findings and their changes across visits. Recent studies have incorporated prior information to enrich current disease representations (Bannur et al., 2023; Liu et al., 2025; Cao et al., 2026; Liu et al., 2026a) or explicitly modeled cross-temporal differences to capture interval changes (Song et al., 2025; Yun et al., 2025; Liu et al., 2026b; Gao et al., 2026). Despite these advances, longitudinal transitions are typically encoded through cross-temporal interactions or endpoint-level representations. EvoRG instead explicitly parameterizes latent transitions between disease-specific endpoints using geometric path descriptors.

2.3 Change Process Modeling

Change process modeling aims to characterize transitions between discrete observations. Existing studies construct intermediate visual content through image interpolation to represent how one observation transforms into another (Zhang et al., 2024). Geometric approaches further formulate such transitions as transport or geodesic paths, demonstrating that continuous paths can provide intermediate states and characterize the magnitude of longitudinal changes in anatomical shapes and latent image representations (Gao et al., 2014; Hong et al., 2019; Yu et al., 2025). Recent work also incorporates temporal representations and procedural context into change understanding (Zhu et al., 2025; Sun et al., 2026). Motivated by these studies, we use geodesic paths to explicitly parameterize longitudinal transitions in a joint visual–semantic space.

3 Methodology

3.1 Overall Framework

Figure 2 presents the overall framework of EvoRG. Given a patient's prior image, prior report, and current image, EvoRG explicitly models disease evolution to generate the current report. Let $D_{\mathrm{cur}} = \{I_{\mathrm{cur}}, R_{\mathrm{cur}}\}$ denote the current examination and its reference report, and let $D_{\mathrm{prior}} = \{I_{\mathrm{prior}}, R_{\mathrm{prior}}\}$ denote the most recent prior examination and report. The report generation process is formulated as follows:

$$\hat{R}_{\mathrm{cur}} = F_\theta(I_{\mathrm{cur}}, D_{\mathrm{prior}}), \tag{1}$$

where F_θ denotes EvoRG with trainable parameters θ, and $\hat{R}_{\mathrm{cur}}$ is the generated current report.

3.2 Cross-Temporal Disease Alignment

We aggregate local visual features from both examinations through shared disease prototypes to obtain semantically comparable endpoints. For $t \in \{\mathrm{prior}, \mathrm{cur}\}$, a shared visual encoder E_{img} maps I_t to a set of local visual tokens:

$$V_t = E_{\mathrm{img}}(I_t) = \{v_{t,i}\}_{i=1}^{N} \in \mathbb{R}^{N \times D}, \tag{2}$$

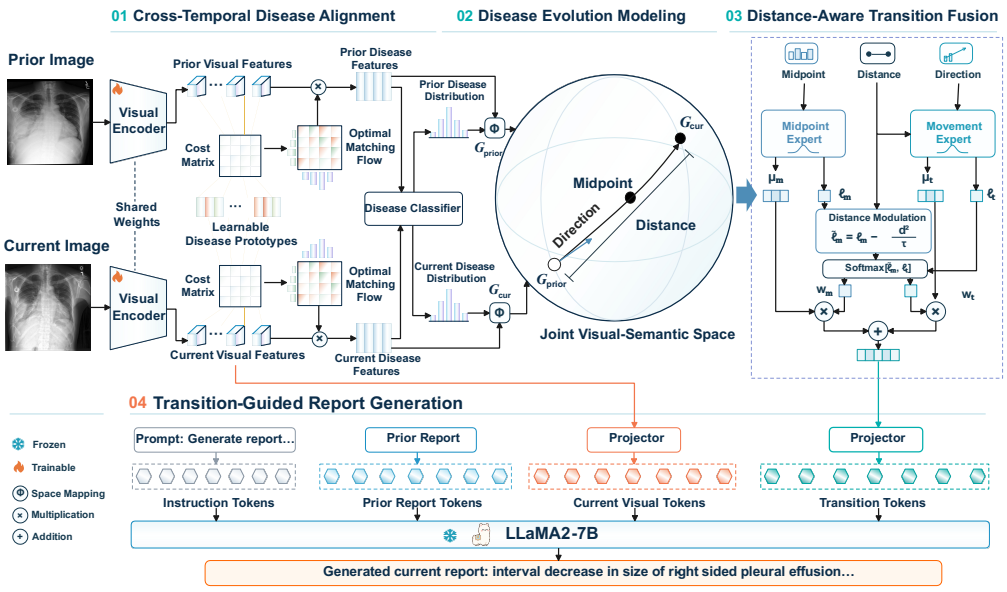


Figure 2: **Overview of EvoRG.** EvoRG constructs disease-wise endpoints via optimal transport and models their transitions as geodesic trajectories on a joint visual–semantic manifold. Distance-aware fusion integrates the geometric descriptors into transition tokens that guide longitudinal report generation.

where N and D denote the number and dimension of visual tokens. We introduce $K = 14$ learnable prototypes $P \in \mathbb{R}^{K\times D}$ shared across visits, with row $p_k^\top$ assigned to disease category k.

The dissimilarity between visual token $v_{t,i}$ and prototype p_k is measured by cosine distance:

$$C_{t,ik} = 1 - \frac{v_{t,i}^\top p_k}{\|v_{t,i}\|_2 \|p_k\|_2}, \tag{3}$$

where $C_t \in \mathbb{R}^{N\times K}$ is the cost matrix.

We compute a soft assignment plan by solving the entropy-regularized optimal transport problem:

$$\Pi_t^* = \arg\min_{\Pi_t \in \mathcal{U}(a,b)} \langle \Pi_t, C_t \rangle + \varepsilon \sum_{i=1}^{N}\sum_{k=1}^{K} \Pi_{t,ik}(\log \Pi_{t,ik} - 1). \tag{4}$$

The feasible set is $\mathcal{U}(a,b) = \{\Pi_t \in \mathbb{R}_+^{N\times K} : \Pi_t \mathbf{1}_K = a,\ \Pi_t^\top \mathbf{1}_N = b\}$, where $a = \frac{1}{N}\mathbf{1}_N$ and $b = \frac{1}{K}\mathbf{1}_K$ specify uniform marginals over visual tokens and disease prototypes, respectively. We solve Eq. (4) using Sinkhorn iterations (Cuturi, 2013) for each visit.

Using the k-th column of Π_t^* as aggregation weights, we obtain the barycentric projection for disease k:

$$Z_t^{(k)} = \frac{\sum_{i=1}^{N} \Pi_{t,ik}^* v_{t,i}}{\sum_{i=1}^{N} \Pi_{t,ik}^*} \in \mathbb{R}^D. \tag{5}$$

Stacking these endpoints row-wise gives $Z_t = [(Z_t^{(1)})^\top; \ldots; (Z_t^{(K)})^\top] \in \mathbb{R}^{K\times D}$, which provides the disease-wise inputs for evolution modeling.

3.3 Disease Evolution Modeling

For each disease, we construct a joint visual–semantic endpoint on a product-sphere manifold. We first normalize each visual endpoint to unit length to remove variation in feature norms:

$$X_t^{(k)} = \frac{Z_t^{(k)}}{\|Z_t^{(k)}\|_2} \in \mathbb{S}^{D-1}. \tag{6}$$

A shared classifier predicts the disease-label distribution $\hat{Y}_t^{(k)} = \mathrm{Softmax}(WZ_t^{(k)} + c) \in \Delta^{S-1}$, where $W \in \mathbb{R}^{S\times D}$, $c \in \mathbb{R}^S$, and $S = 4$ is the number of label classes: positive, negative, uncertain, and unmentioned. Taking the element-wise square root of this distribution maps it to the positive orthant of a unit sphere, yielding the semantic component:

$$Q_t^{(k)} = \sqrt{\hat{Y}_t^{(k)}} \in \mathbb{S}_+^{S-1}. \tag{7}$$

Together, these components form the joint endpoint:

$$G_t^{(k)} = \left(X_t^{(k)}, Q_t^{(k)}\right) \in \mathcal{M}, \qquad \mathcal{M} = \mathbb{S}^{D-1} \times \mathbb{S}_+^{S-1}. \tag{8}$$

We equip $\mathcal{M}$ with the weighted product metric $g_{\mathcal{M}} = g_{\text{visual}} \oplus \beta_{\text{semantic}} g_{\text{semantic}}$, where $\beta_{\text{semantic}} > 0$ controls the semantic contribution. To preserve the intrinsic geometric structure of these representations, we model disease evolution in the latent space by the constant-speed shortest geodesic:

$$\gamma^{(k)}(s) = \left(\gamma_{\text{visual}}^{(k)}(s), \gamma_{\text{semantic}}^{(k)}(s)\right), \qquad s \in [0, 1], \tag{9}$$

where s is a normalized path parameter and each component follows $\gamma_{x\to y}(s) = \mathrm{Exp}_x(s\,\mathrm{Log}_x(y))$ from its prior endpoint x to its current endpoint y. Here, Exp and Log are the spherical exponential and logarithmic maps (Miolane et al., 2020); coincident endpoints yield a constant path. For non-antipodal endpoints in each component, this gives $\gamma^{(k)}(0) = G_{\text{prior}}^{(k)}$ and $\gamma^{(k)}(1) = G_{\text{cur}}^{(k)}$, providing a structured latent parameterization of the transition. We use the initial tangent, midpoint, and length to characterize the evolution direction, intermediate representation, and evolution magnitude, respectively.

For non-antipodal unit vectors x and y, the spherical angle is $\theta(x, y) = \mathrm{atan2}\left(\|y - \langle x, y\rangle x\|_2, \langle x, y\rangle\right) \in [0, \pi)$. The initial tangent $\dot{\gamma}_{x\to y}(0) = \mathrm{Log}_x(y)$ captures directed evolution:

$$\mathrm{Log}_x(y) = \begin{cases} \dfrac{\theta(x,y)}{\|y - \langle x, y\rangle x\|_2}\left(y - \langle x, y\rangle x\right), & 0 < \theta(x, y) < \pi, \\ \mathbf{0}, & x = y. \end{cases} \tag{10}$$

Applying this map to the two components gives $u_{\text{visual}}^{(k)} = \mathrm{Log}_{X_{\text{prior}}^{(k)}}(X_{\text{cur}}^{(k)})$ and $u_{\text{semantic}}^{(k)} = \mathrm{Log}_{Q_{\text{prior}}^{(k)}}(Q_{\text{cur}}^{(k)})$.

For non-antipodal endpoints, the geodesic midpoint $\gamma_{x\to y}(1/2)$ provides an intermediate representation of the modeled evolution and is computed as $m(x, y) = (x+y)/\|x+y\|_2$. This gives $m_{\text{visual}}^{(k)}$ and $m_{\text{semantic}}^{(k)}$.

The trajectory length under the weighted product metric gives the transition magnitude:

$$d^{(k)} = \sqrt{\left(\theta(X_{\text{prior}}^{(k)}, X_{\text{cur}}^{(k)})\right)^2 + \beta_{\text{semantic}}\left(\theta(Q_{\text{prior}}^{(k)}, Q_{\text{cur}}^{(k)})\right)^2}. \tag{11}$$

We concatenate the midpoint components to form $M^{(k)}$ and combine the visual tangent, the weighted semantic tangent, and the path length to form the directed-movement descriptor $T^{(k)}$:

$$M^{(k)} = [m_{\text{visual}}^{(k)}; m_{\text{semantic}}^{(k)}], \tag{12}$$

$$T^{(k)} = [u_{\text{visual}}^{(k)}; \sqrt{\beta_{\text{semantic}}} u_{\text{semantic}}^{(k)}; d^{(k)}]. \tag{13}$$

3.4 Distance-Aware Transition Fusion

To combine midpoint and directed-movement information, we encode $M^{(k)}$ and $T^{(k)}$ with two experts that output means $\mu_m^{(k)}, \mu_t^{(k)} \in \mathbb{R}^D$ and scalar log-precisions:

$$(\mu_m^{(k)}, \ell_m^{(k)}) = E_m(M^{(k)}), \qquad (\mu_t^{(k)}, \ell_t^{(k)}) = E_t(T^{(k)}). \tag{14}$$

Each expert defines an isotropic Gaussian $q_j^{(k)}(h) = \mathcal{N}(h; \mu_j^{(k)}, e^{-\ell_j^{(k)}} I_D)$ over the shared latent representation h, with $j \in \{m, t\}$.

To increase the relative contribution of movement information for larger transitions, we apply a distance-dependent bias to the midpoint log-precision:

$$\widetilde{\ell}_m^{(k)} = \ell_m^{(k)} - \frac{(d^{(k)})^2}{\tau}, \tag{15}$$

where $\tau > 0$ is the distance modulation temperature.

Following the product-of-experts principle (Hinton, 2002), we use the modulated midpoint expert $\widetilde{q}_m^{(k)}$ and take the mean of the product $q^{(k)}(h) \propto \widetilde{q}_m^{(k)}(h) q_t^{(k)}(h)$ as the fused representation, with normalized precision weights $[w_m^{(k)}, w_t^{(k)}] = \mathrm{Softmax}([\widetilde{\ell}_m^{(k)}, \ell_t^{(k)}])$:

$$H^{(k)} = w_m^{(k)} \mu_m^{(k)} + w_t^{(k)} \mu_t^{(k)} \in \mathbb{R}^D. \tag{16}$$

We project the fused representation to the language-model hidden dimension D_{LLM}:

$$T_{\mathrm{LLM}}^{(k)} = \mathrm{Proj}(H^{(k)}) \in \mathbb{R}^{D_{\mathrm{LLM}}}. \tag{17}$$

Collecting the projected tokens for all K categories gives $T_{\mathrm{LLM}} = [T_{\mathrm{LLM}}^{(1)}, \ldots, T_{\mathrm{LLM}}^{(K)}]$.

3.5 Transition-Guided Report Generation

We form the conditioning sequence $\mathcal{C}$ from the projected current visual tokens $X_{\mathrm{cur}}^{\mathrm{LLM}} = P_{\mathrm{img}}(V_{\mathrm{cur}})$, transition tokens T_{LLM}, prior report embeddings $E(R_{\mathrm{prior}})$, and instruction tokens X_{inst}. The language model generates the current report conditioned on this sequence:

$$\hat{R}_{\mathrm{cur}} = F_{\mathrm{LLM}}(\mathcal{C}). \tag{18}$$

For a target report $R_{\mathrm{cur}} = \{r_t\}_{t=1}^T$, the autoregressive generation loss is:

$$\mathcal{L}_{\mathrm{LM}} = -\sum_{t=1}^{T} \log p_\theta \left(r_t \mid r_{<t}, \mathcal{C} \right). \tag{19}$$

3.6 Overall Objective

We supervise disease-label predictions for both examinations. Let $A_t = Z_t W^\top + \mathbf{1}_K c^\top \in \mathbb{R}^{K \times S}$ denote the stacked classifier logits at visit t, and let $Y_t \in \{1, \ldots, S\}^K$ denote the disease labels extracted from the corresponding reports using CheXbert for the $K = 14$ categories. These labels are used only for auxiliary endpoint supervision during training. The disease-label supervision loss is:

$$\mathcal{L}_{\mathrm{label}} = \mathrm{CE}(A_{\mathrm{cur}}, Y_{\mathrm{cur}}) + \mathrm{CE}(A_{\mathrm{prior}}, Y_{\mathrm{prior}}), \tag{20}$$

where each endpoint is supervised by the label of its corresponding disease category, and cross-entropy is averaged over categories.

The total training objective is:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{LM}} + \lambda_{\mathrm{label}} \mathcal{L}_{\mathrm{label}}, \tag{21}$$

where λ_{label} controls the strength of the auxiliary supervision. All trainable modules are jointly optimized with this objective.

4 Experiments

4.1 Experimental Setting

Dataset. We evaluate our method on the Longitudinal-MIMIC dataset (Zhu et al., 2023), a benchmark for longitudinal radiology report generation derived from the MIMIC-CXR (Johnson et al.,

Table 1: Performance comparison of state-of-the-art baselines across NLG and CE metrics on the Longitudinal-MIMIC dataset.

Model	Year	Inputs	NLG metrics							CE metrics		
			BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	PREC	REC	F1
Transformer	2017	Single	0.294	0.178	0.119	0.085	0.123	0.256	-	-	-	-
AoANet	2019	Single	0.272	0.168	0.112	0.080	0.115	0.249	-	-	-	-
CNN+Trans	2019	Single	0.299	0.186	0.124	0.088	0.120	0.263	-	-	-	-
R2Gen	2020	Single	0.302	0.183	0.122	0.087	0.124	0.259	-	-	-	-
R2CMN	2021	Single	0.305	0.184	0.122	0.085	0.126	0.265	-	-	-	-
CvT2DistilGPT2	2023	Single	0.365	0.226	0.151	0.107	0.143	0.275	-	0.367	0.258	0.261
R2GenGPT	2023	Single	0.365	0.230	0.156	0.111	0.139	0.271	0.180	0.297	0.240	0.247
PromptMRG	2024	Single	0.370	0.219	0.141	0.098	0.144	0.266	-	-	-	-
Prefilling	2023	Longitudinal	0.343	0.210	0.140	0.099	0.137	0.271	-	-	-	-
HERGen	2024	Longitudinal	0.389	0.242	0.163	0.117	0.155	0.282	-	0.421	0.289	0.295
Diff-RRG	2025	Longitudinal	0.401	0.257	0.176	0.127	0.156	0.284	0.237	0.417	0.381	0.370
HC-LLM	2025	Longitudinal	0.404	0.260	0.178	0.128	0.160	0.287	-	0.417	0.357	0.357
TSMRG	2026	Longitudinal	0.399	0.244	0.159	0.110	0.158	0.271	-	-	-	-
MARE	2026	Longitudinal	0.409	0.265	0.184	0.133	0.161	0.291	0.271	0.433	0.378	0.375
BiOTPrompt	2026	Longitudinal	0.397	0.253	0.174	0.126	0.155	0.285	0.236	0.471	0.424	0.417
EvoRG (Ours)	–	Longitudinal	**0.420**	**0.273**	**0.190**	**0.139**	**0.167**	**0.300**	**0.305**	**0.512**	**0.465**	**0.457**

2019). MIMIC-CXR contains 377,110 chest X-ray images and 227,835 radiographic studies from 65,379 patients. Following prior work, we select 26,625 patients with at least two visits and pair consecutive visits chronologically. Clinical observations are represented using the 14 CheXbert categories. Each sample contains a prior and a current examination, together with their corresponding chest X-ray images and radiology reports. The dataset is divided into 92,374 training samples from 26,156 patients, 737 validation samples from 203 patients, and 2,058 test samples from 266 patients. Within each patient, examinations are ordered by the StudyDate attribute so that each prior–current pair preserves the temporal order.

Evaluation Metrics. We evaluate the quality of generated radiology reports from both natural language generation (NLG) and clinical efficacy (CE) perspectives. For NLG evaluation, we report BLEU (Papineni et al., 2002), METEOR (Denkowski & Lavie, 2011), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015). These metrics evaluate the similarity between generated reports and their corresponding reference reports. For clinical evaluation, we use the CheXbert labeler (Smit et al., 2020) to extract clinical observation labels from generated and reference reports. We compute precision, recall, and F1 for each report pair and average them across the dataset to obtain example-based CE scores. Additional evaluations, including macro- and micro-averaged CheXbert scores, BERTScore (Zhang et al., 2019), F1RadGraph (Delbrouck et al., 2022), RaTEScore (Zhao et al., 2024), Temporal F1 (Zhang et al., 2025), and RadCliQ-v1 (Yu et al., 2023), are provided in Appendix E.

Implementation Details. We use the base Swin Transformer (Liu et al., 2021) as the visual encoder and LLaMA2-7B-Chat (Touvron et al., 2023) as the frozen language model. The number of Sinkhorn iterations is set to 50. The entropic regularization coefficient ε, semantic geometry weight β_{semantic}, distance modulation temperature τ, and disease-label supervision weight λ_{label} are set to 0.1, 3.0, 0.25, and 1.0, respectively. We implement the model in PyTorch and train it on Longitudinal-MIMIC for 5 epochs using mixed-precision training on a single NVIDIA A40 GPU with 48 GB memory. We use AdamW with an initial learning rate of 1×10^{-4}, a cosine annealing learning-rate scheduler, and a batch size of 8. During inference, we use beam search with a beam size of 3.

4.2 Comparison with State-of-the-Art Methods

We compare EvoRG with single-visit radiology report generation methods, including Transformer (Vaswani et al., 2017), AoANet (Huang et al., 2019), CNN+Trans, R2Gen (Chen et al., 2020), R2CMN (Chen et al., 2021), CvT2DistilGPT2 (Nicolson et al., 2023), R2GenGPT (Wang et al., 2023), and PromptMRG (Jin et al., 2024), as well as longitudinal methods, including Prefilling (Zhu et al., 2023), HERGen (Wang et al., 2024), Diff-RRG (Yun et al., 2025), HC-LLM (Liu et al., 2025), TSMRG (Cao et al., 2026), MARE (Gao et al., 2026), and BiOTPrompt (Liu et al., 2026b). All comparisons and ablations follow the same experimental protocol. Table 1 reports

Table 2: Ablation study of EvoRG components on Longitudinal-MIMIC. "DA", "DEM", and "DTF" denote Disease Alignment, Disease Evolution Modeling, and Distance-Aware Transition Fusion, respectively.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	PREC	REC	F1
w/o DA	0.410	0.264	0.182	0.131	0.162	0.292	0.285	0.489	0.441	0.436
w/o DEM	0.414	0.268	0.185	0.133	0.164	0.295	0.287	0.499	0.440	0.439
w/o DTF	0.415	0.270	0.188	0.136	0.165	0.298	**0.309**	0.493	0.435	0.434
EvoRG	**0.420**	**0.273**	**0.190**	**0.139**	**0.167**	**0.300**	0.305	**0.512**	**0.465**	**0.457**

Table 3: Ablation study of disease evolution descriptors in DEM on Longitudinal-MIMIC. "M", "U", and "D" denote the midpoint components, tangent components, and explicit transition distance, respectively.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	PREC	REC	F1
w/o M	0.414	0.269	0.187	0.135	0.164	0.297	0.300	0.497	0.438	0.436
w/o U	0.408	0.264	0.183	0.132	0.162	0.294	0.273	0.495	0.451	0.442
w/o D	0.412	0.266	0.184	0.134	0.162	0.293	0.276	0.491	0.444	0.438
EvoRG	**0.420**	**0.273**	**0.190**	**0.139**	**0.167**	**0.300**	**0.305**	**0.512**	**0.465**	**0.457**

published baseline results, together with Diff-RRG reproduced using its official implementation. EvoRG achieves the highest scores among the compared methods across all reported NLG and CE metrics. Compared with MARE, the strongest baseline on NLG metrics, EvoRG improves CIDEr from 0.271 to 0.305. Compared with BiOTPrompt, the strongest baseline on CE metrics, it improves CheXbert F1 from 0.417 to 0.457, with corresponding gains in precision and recall. These results indicate improvements in both textual similarity and clinical agreement with reference reports.

Supplementary evaluation in Appendix E further supports these findings across additional semantic and clinical metrics. EvoRG achieves a Temporal F1 of 0.238, compared with 0.233 for BiOTPrompt and 0.231 for Diff-RRG, indicating improved agreement in temporal expressions. Together, these results support the effectiveness of the proposed disease-wise evolution representation for longitudinal report generation.

4.3 Model Analysis

Ablation Study. We first evaluate the three EvoRG components in Table 2. Removing DA consistently degrades both NLG and clinical metrics, reducing CheXbert F1 from 0.457 to 0.436. This result supports the role of disease alignment in establishing comparable disease-wise endpoints across examinations. Replacing DEM with a static endpoint difference decreases CIDEr from 0.305 to 0.287 and CheXbert F1 from 0.457 to 0.439, supporting the effectiveness of the proposed transition representation. Appendix B further compares static difference, Euclidean transition, and product-sphere transition representations to examine the contributions of structured descriptors and their geometry. The w/o DTF variant achieves a CIDEr score of 0.309, compared with 0.305 for the full model. Incorporating DTF improves clinical precision, recall, and F1 from 0.493, 0.435, and 0.434 to 0.512, 0.465, and 0.457, respectively, supporting its contribution to clinical agreement with reference reports.

We further analyze the disease evolution descriptors within DEM in Table 3. Removing the midpoint, directed tangent, or explicit transition distance decreases performance across the reported metrics. Removing the tangent components produces the largest decline in CIDEr, from 0.305 to 0.273. Removing the explicit distance descriptor reduces CIDEr to 0.276 and CheXbert F1 to 0.438, supporting the benefit of explicitly incorporating transition magnitude. Removing the midpoint decreases CheXbert F1 to 0.436, suggesting that the intermediate representation contributes complementary transition information. Together, these results support the use of midpoint, tangent, and distance descriptors for longitudinal report generation.

Qualitative Results. As shown in Figure 3, EvoRG correctly describes the interval decrease in right pleural effusion while retaining the finding that the pleural catheter remains in place, consistent with the reference report. In contrast, BiOTPrompt describes the effusion as unchanged, missing

Current Image	Prior Image	
		**BiOTPrompt:** as compared to the previous radiograph there is no relevant change . unchanged moderate-to-large right pleural effusion with adjacent atelectasis . unchanged appearance of the cardiac silhouette and of the left hemithorax . unchanged position of the left picc line . unchanged appearance of the surgical clips at the level of the sternum . no new parenchymal opacities .
		EvoRG (Ours): impression : interval decrease in size of right-sided pleural effusion . findings : pa and lateral views of the chest are compared to previous exam from . right chest wall port is again seen with catheter tip in the lower svc . right-sided pleural catheter is seen which appears to course in the fissure . interval decrease in size of right-sided pleural effusion . no pneumothorax . cardiomediastinal and hilar contours are unchanged .
		w/o DA: impression : no significant interval change . findings : compared to the prior radiograph there is no significant interval change . right-sided pleural catheter remains in place with tip in the lower svc . left chest wall port is again seen with catheter tip in the fissure . cardiomediastinal silhouette and hilar contours are unchanged . bibasilar atelectasis is unchanged . no pneumothorax .
Reference change: mild decrease in right loculated effusion; pleural catheter retained.		**w/o DEM:** as compared to the previous radiograph the right pleural catheter has been removed . there is no evidence of pneumothorax . the pre-existing moderate right pleural effusion has slightly decreased in size . the left lung is unremarkable . no new parenchymal opacities . unchanged appearance of the cardiac silhouette and the hilar contours .
Ground Truth: in comparison with the study of there is mild decrease in the still substantial opacification along the mid and lower lateral chest wall on the right most likely reflecting a loculated effusion . apparent pleural catheter remains in place as does the port-a-cath . little overall change in the appearance of the heart and lungs .		**w/o DTF:** impression : persistent right-sided pleural effusion with associated atelectasis . findings : pa and lateral views of the chest are compared to previous exam from . right chest wall port is again seen with catheter tip in the lower svc . right-sided pleural catheter is seen which appears to course in the fissure . there is persistent right-sided pleural effusion with associated atelectasis . there is no pneumothorax . left lung is clear . cardiomediastinal silhouette is stable .

Figure 3: Comparison of generated reports across models on a longitudinal case from the test set of Longitudinal-MIMIC.

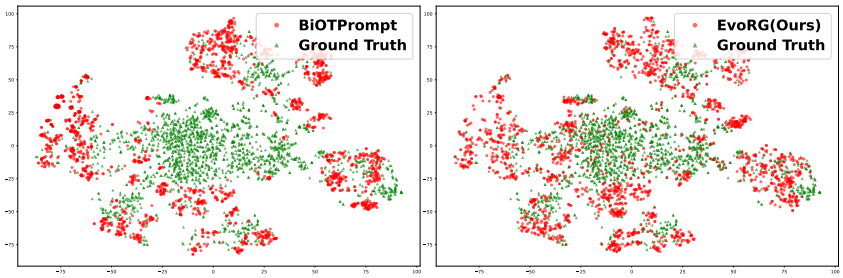


Figure 4: t-SNE visualization of report representations for BiOTPrompt and EvoRG.

the improvement between examinations. The ablated variants also exhibit distinct errors: w/o DA reports no significant interval change, while w/o DTF describes persistent effusion without capturing its reduction. Although w/o DEM identifies the slight decrease in effusion, it incorrectly states that the pleural catheter has been removed. This example illustrates how the proposed representation can support interval-change descriptions while retaining relevant clinical findings. We further visualize the representations of generated and reference reports using t-SNE (Van der Maaten & Hinton, 2008) in Figure 4, where red circles and green triangles denote generated and reference reports, respectively. Compared with BiOTPrompt, EvoRG exhibits greater overlap with the reference-report distribution, suggesting improved semantic consistency between generated and reference reports. Together, these qualitative results support the effectiveness of the proposed disease evolution modeling for longitudinal report generation.

5 Conclusion

In this paper, we propose EvoRG, a longitudinal radiology report generation framework that moves from identifying what has changed to describing how the change occurs. Comparable disease representations are constructed across visits via optimal transport and shared prototypes. This alignment reduces the impact of non-pathological spatial shifts. To model the disease evolution, latent transitions are parameterized as continuous geodesic trajectories on a joint visual-semantic manifold. These trajectories yield geometric descriptors capturing direction, magnitude, and intermediate states. The extracted components are then fused within a distance-aware module to condition the language model for evolution-aware report generation. Experimental results on the Longitudinal-MIMIC dataset show that by modeling continuous transitions rather than static differences, the proposed approach achieves state-of-the-art performance across natural language generation and clinical efficacy metrics. Ablation studies confirm the effectiveness of the proposed modules. Currently, the evaluation of disease evolution relies on paired consecutive chest X-rays. Future work will extend this approach to multi-visit sequences and other medical imaging modalities.

AI USE STATEMENT

Large language models (LLMs) were used to polish the writing and to support literature retrieval and discovery, including the search for related work. They were not used to develop the research idea or methodology, design the experiments, analyze the data, or generate the reported results. The authors take full responsibility for the final manuscript and all of its scientific claims.

ETHICS STATEMENT

This work uses de-identified longitudinal chest X-ray images and reports from the Longitudinal-MIMIC dataset. No new participants were recruited, and no animal experiments or clinical interventions were conducted. We followed the dataset's access and usage requirements and did not use or release personally identifiable information. The proposed system is intended for research purposes and is not a substitute for professional clinical judgment.

REPRODUCIBILITY STATEMENT

The paper and appendices provide the model formulation, training objectives, and implementation details, including the configuration in Appendix D. Code, configuration files, and evaluation scripts will be released to support reproduction, subject to the access requirements of the Longitudinal-MIMIC dataset.

REFERENCES

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15016–15027. IEEE, 2023.

Lili Cao, Haoran Wu, and Bo Xu. Tsmrg: Temporal semantic enhancement for medical report generation with longitudinal data. *IEEE Transactions on Multimedia*, 2026.

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.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, 2013.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4348–4360, 2022.

Jean-Benoit Delbrouck, Justin Xu, Johannes Moll, Alois Thomas, Zhihong Chen, Sophie Ostmeier, Asfandyar Azhar, Kelvin Zhenghao Li, Andrew Johnston, Christian Bluethgen, et al. Automated structured radiology report generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26813–26829, 2025.

Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the sixth workshop on statistical machine translation*, pp. 85–91, 2011.

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*, pp. 21180–21188, 2026.

Yi Gao, Liang-Jia Zhu, Sylvain Bouix, and Allen Tannenbaum. Interpolation of longitudinal shape and image data via optimal mass transport. In *Proceedings of SPIE–the International Society for Optical Engineering*, volume 9034, pp. 90342X, 2014.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

Sungmin Hong, James Fishbaugh, Jason J Wolff, Martin A Styner, Guido Gerig, and IBIS Network. Hierarchical multi-geodesic model for longitudinal analysis of temporal trajectories of anatomical shape and covariates. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 57–65. Springer, 2019.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4633–4642. IEEE, 2019.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 2607–2615, 2024.

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, 2019.

Yunsoo Kim, Jinge Wu, Su Hwan Kim, Pardeep Vasudev, Jiashu Shen, and Honghan Wu. Look & mark: Leveraging radiologist eye fixations and bounding boxes in multimodal large language models for chest x-ray report generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 17680–17694, 2025.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

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*, pp. 7206–7214, 2026a.

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*, pp. 5595–5603, 2025.

Tengfei Liu, Yijian Fan, Boyue Wang, Yongli Hu, Mingjie Li, Jinghua Li, Junbin Gao, Xiaojun Chang, Zhihui Li, and Baocai Yin. Biotprompt: Bidirectional optimal transport guided prompting for disease evolution-aware radiology report generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13755–13765, 2026b.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9992–10002. Ieee, 2021.

Nina Miolane, Nicolas Guigui, Alice Le Brigant, Johan Mathe, Benjamin Hou, Yann Thanwerdas, Stefan Heyder, Olivier Peltre, Niklas Koep, Hadi Zaatiti, et al. Geomstats: A python package for riemannian geometry in machine learning. *Journal of Machine Learning Research*, 21(223):1–9, 2020.

Aaron Nicolson, Jason Dowling, and Bevan Koopman. Improving chest x-ray report generation by leveraging warm starting. *Artificial intelligence in medicine*, 144:102633, 2023.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

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 *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 1500–1519, 2020.

Shanshan Song, Hui Tang, Honglong Yang, and Xiaomeng Li. Ddatr: Dynamic difference-aware temporal residual network for longitudinal radiology report generation. *IEEE Transactions on Medical Imaging*, 44(12):5345–5357, 2025.

Jiayang Sun, Zixin Guo, Min Cao, Guibo Zhu, and Jorma Laaksonen. Imagine how to change: Explicit procedure modeling for change captioning. *arXiv preprint arXiv:2603.05969*, 2026.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

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.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3):100033, 2023.

Alexandra L Young, Neil P Oxtoby, Sara Garbarino, Nick C Fox, Frederik Barkhof, Jonathan M Schott, and Daniel C Alexander. Data-driven modelling of neurodegenerative disease progression: thinking outside the black box. *Nature Reviews Neuroscience*, 25(2):111–130, 2024.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, et al. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9), 2023.

Qingtao Yu, Jaskirat Singh, Zhaoyuan Yang, Peter Henry Tu, Jing Zhang, Hongdong Li, Richard Hartley, and Dylan Campbell. Probability density geodesics in image diffusion latent space. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27989–27998. IEEE, 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.

Erjian Zhang, Yatong Hao, Liejun Wang, and Zhiqing Guo. The double dilemma in multi-task radiology report generation: A gradient dynamics analysis and solution. *arXiv preprint arXiv:2605.22635*, 2026.

Kaiwen Zhang, Yifan Zhou, Xudong Xu, Bo Dai, and Xingang Pan. Diffmorpher: Unleashing the capability of diffusion models for image morphing. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7912–7921. IEEE, 2024.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Xi Zhang, Zaiqiao Meng, Jake Lever, and Edmond SL Ho. Libra: Leveraging temporal images for biomedical radiology analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 17275–17303, 2025.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Ratescore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15004–15019, 2024.

Duowang Zhu, Xiaohu Huang, Haiyan Huang, Hao Zhou, and Zhenfeng Shao. Change3d: Revisiting change detection and captioning from a video modeling perspective. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24011–24022. IEEE, 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.

A Appendix Overview

The appendix provides the following details:

B. Effect of Transition Representation: A comparison of static difference, Euclidean transition, and product-sphere transition representations.
C. Hyperparameter Sensitivity Analysis: An analysis of the effects of the four key hyperparameters in EvoRG.
D. Implementation and Inference Details: The model configuration and settings used for training and inference.
E. Supplementary Evaluation: Additional clinical and report evaluation metrics on Longitudinal-MIMIC.
F. Additional Qualitative Analysis: Representative cases of improvement, worsening, stable findings, and newly emerging findings.

B Effect of Transition Representation

We compare three transition representations between aligned disease-wise endpoints. Static Difference corresponds to the w/o DEM baseline in Table 2. Euclidean Transition and Product-Sphere Transition use the same normalized visual endpoints $X_t^{(k)}$, semantic endpoints $Q_t^{(k)}$, expert networks, and fusion configuration. Euclidean Transition uses arithmetic midpoints, directed endpoint differences, and a weighted Euclidean distance. Product-Sphere Transition uses geodesic midpoints, logarithmic-map tangents, and a weighted geodesic distance. The training settings and hyperparameters are shared between these two structured variants.

Table 4: Comparison of transition representations on Longitudinal-MIMIC.

Representation	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	PREC	REC	F1
Static Difference	0.414	0.268	0.185	0.133	0.164	0.295	0.287	0.499	0.440	0.439
Euclidean Transition	0.415	0.269	0.187	0.136	0.164	0.296	0.295	0.503	0.449	0.445
Product-Sphere Transition	**0.420**	**0.273**	**0.190**	**0.139**	**0.167**	**0.300**	**0.305**	**0.512**	**0.465**	**0.457**

As shown in Table 4, Euclidean Transition improves CIDEr from 0.287 to 0.295 and CheXbert F1 from 0.439 to 0.445 over Static Difference. This suggests that representing the transition with a midpoint, directed movement, and distance is more effective than using the endpoint difference alone. Product-Sphere Transition further increases CIDEr to 0.305 and CheXbert F1 to 0.457. These results support the effectiveness of structured transition representations and demonstrate further gains from computing their components on the product-sphere manifold.

C Hyperparameter Sensitivity Analysis

We vary each hyperparameter independently and evaluate its effect on CIDEr and CheXbert F1-score in Figure 5. Both metrics improve and then decline as the entropic regularization strength increases, suggesting that the quality of disease alignment depends on the degree of smoothing in the transport plan. The setting $\beta_{\text{semantic}} = 0$ serves as an ablation that removes the semantic contribution to the transition distance and weighted tangent while retaining the semantic midpoint. Increasing the semantic geometry weight initially improves performance, whereas larger weights reduce it, indicating that semantic displacement should contribute to evolution modeling without dominating the visual component. Disease-label supervision improves both metrics compared with its absence, but assigning it greater weight eventually reduces these gains, particularly in CheXbert F1-score. Performance peaks at an intermediate distance modulation temperature and declines at the tested extremes, indicating the importance of balancing midpoint and directed-movement information.

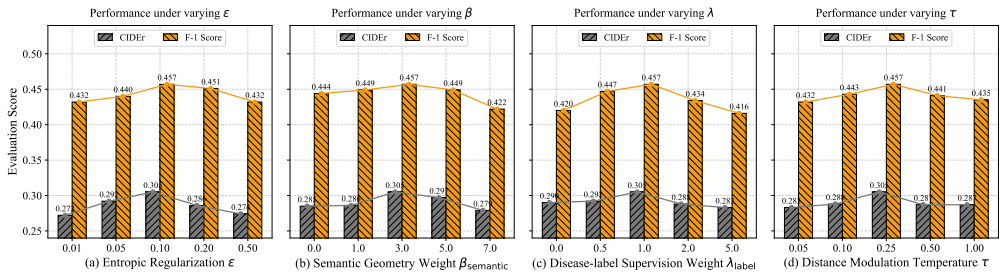


Figure 5: Essential hyperparameter analysis of EvoRG.

D Implementation and Inference Details

Table 5 summarizes the configuration used to obtain the full EvoRG results in Table 1. The visual encoder, transition modules, image projector, disease prototypes, and disease classifier are jointly optimized using the total objective in Eq. (21), while the LLaMA2-7B-Chat parameters are frozen. We use mixed-precision training with a batch size of 8 for five epochs on one NVIDIA A40 GPU. At inference, the model takes the prior image, prior report, and current image as input and generates the current report using beam search. The semantic endpoints are computed from the disease classifier's predicted softmax distributions.

Table 5: Implementation configuration used for the full EvoRG model.

Component	Setting	Component	Setting
Visual encoder	Swin-Base, patch4-window7-224	Language model	LLaMA2-7B-Chat
Trainable visual encoder	Yes	Trainable LLaMA weights	No
OT regularization (ε)	0.1	Sinkhorn iterations	50
Disease prototypes (K)	14	Label classes per category (S)	4
Semantic geometry weight	3.0	Distance modulation temperature (τ)	0.25
Geometry tolerance	10^{-6}	Transition dropout	0.05
Disease-label supervision weight (λ_{label})	1.0	Optimizer	AdamW
Initial learning rate	10^{-4}	Scheduler	CosineAnnealing
Training epochs	5	Training batch size	8
Precision	bf16-mixed	Inference beam size	3
Minimum new tokens	80	Maximum new tokens	120
Repetition penalty	2.0	Length penalty	2.0

E Supplementary Evaluation

We compare EvoRG with Diff-RRG and BiOTPrompt using additional clinical and report evaluation metrics on Longitudinal-MIMIC. The supplementary metrics are computed from the generated and reference reports using a shared evaluation pipeline.

E.1 Macro- and Micro-Averaged Clinical Evaluation

The CheXbert CE scores in Table 1 are example-based: precision, recall, and F1 are computed separately for each report and then averaged across reports. Here, we additionally report macro-averaged and micro-averaged CE scores over the 14 observation categories. Macro averaging computes each metric for each category across reports and then takes the unweighted mean over categories. Micro averaging pools true positives, false positives, and false negatives across all reports and categories before computing each metric. These aggregation schemes complement the example-based evaluation by assessing clinical agreement with equal weight per category and through pooled prediction counts, respectively.

Table 6: Macro- and micro-averaged CheXbert CE scores on Longitudinal-MIMIC. Higher is better for all metrics.

Method	**Macro average**			**Micro average**		
	Precision	Recall	F1	Precision	Recall	F1
Diff-RRG	0.397	0.290	0.311	0.488	0.420	0.451
BiOTPrompt	0.315	0.256	0.255	0.533	0.443	0.484
EvoRG	**0.413**	**0.321**	**0.346**	**0.566**	**0.497**	**0.529**

As shown in Table 6, EvoRG achieves the highest precision, recall, and F1 under both aggregation schemes. Compared with the strongest baseline under each scheme, macro F1 increases from 0.311 to 0.346 over Diff-RRG, while micro F1 increases from 0.484 to 0.529 over BiOTPrompt. Together with the example-based CE results in Table 1, these results show that EvoRG's advantage in clinical agreement holds across report-level, category-level, and pooled evaluation, supporting the effectiveness of the disease-wise evolution representation in improving clinical agreement with reference reports.

E.2 Additional Report Evaluation Metrics

We further use RadEval to compute BERTScore, all three F1RadGraph scores (Simple, Partial, and Complete), RaTEScore, Temporal F1, and RadCliQ-v1. Temporal F1 measures agreement between temporal expressions extracted from the generated and reference reports. We report the reciprocal of the mean RadCliQ-v1 score, denoted by 1/RadCliQ-v1, so that higher values indicate better performance for all reported metrics.

Table 7: Additional report evaluation results on Longitudinal-MIMIC. Higher is better for all metrics.

Method	**BERTScore**	**F1RadGraph**			**RaTEScore**	**Temporal F1**	**1/RadCliQ-v1**
		Simple	Partial	Complete			
Diff-RRG	0.517	0.241	0.214	0.148	0.521	0.231	0.677
BiOTPrompt	0.515	0.239	0.214	0.151	0.513	0.233	0.671
EvoRG	**0.528**	**0.263**	**0.234**	**0.168**	**0.536**	**0.238**	**0.726**

Table 7 shows that EvoRG outperforms both baselines across all reported metrics, including all three F1RadGraph scores. The improvements in BERTScore, F1RadGraph, and RaTEScore provide complementary evidence of better semantic and clinical agreement with reference reports. EvoRG also increases 1/RadCliQ-v1 from 0.677 to 0.726 compared with Diff-RRG. Temporal F1 reaches 0.238, compared with 0.231 for Diff-RRG and 0.233 for BiOTPrompt, indicating improved agreement in temporal expressions. These results complement the clinical evaluation and support the use of explicit disease evolution modeling for longitudinal report generation.

F Additional Qualitative Analysis

Table 8 presents representative test-set cases involving improvement, worsening, stable findings, and newly emerging findings. The examples compare how EvoRG and its ablated variants describe interval changes and preserve current clinical findings.

Table 8: Representative longitudinal cases from the test-set outputs. Each row corresponds to a distinct test example. Entries summarize clinically relevant findings and interval-change descriptions in the reference and generated reports.

Reference evolution	**Reference findings**	**EvoRG**	**w/o DA**	**w/o DEM**	**w/o DTF**
Improvement	Stable post-surgical and post-radiation changes with interval improvement of the previous right lower-lobe opacity; no new opacity, effusion, or pneumothorax.	Reports stable post-treatment changes with slightly decreased right-lung opacification and no new focal consolidation, effusion, or pneumothorax.	Reports the stable post-treatment appearance and slightly decreased right-lung opacification.	Reports the stable post-treatment appearance and slightly decreased right-lung opacification.	Preserves the stable post-treatment changes but omits the interval decrease.
Worsening	Diffuse interstitial opacities have increased, consistent with worsening pulmonary edema on a background of pulmonary fibrosis.	Explicitly identifies worsening pulmonary edema and preserves the absence of pneumothorax or large pleural effusion.	Describes mild pulmonary edema without explicitly stating worsening.	Describes mild pulmonary edema without explicitly stating worsening.	Describes mild pulmonary edema and reduced lung volumes without explicitly stating worsening.
Stable	No acute intrathoracic process; no focal consolidation, effusion, or pneumothorax; unchanged elevation of the right hemidiaphragm.	Reports no acute cardiopulmonary process and preserves the absence of consolidation, effusion, and pneumothorax.	Reports no acute cardiopulmonary abnormality with clear lungs and no effusion or pneumothorax.	Reports no acute intrathoracic process with no focal consolidation, effusion, or pneumothorax.	Reports no acute intrathoracic process with no focal consolidation, effusion, or pneumothorax.
New finding	New diffuse perihilar airspace opacities, concerning for widespread or atypical pneumonia, with possible superimposed pulmonary edema.	Detects new airspace opacities in the right upper and lower lobes and left upper lobe and describes multifocal pneumonia.	Describes diffuse bilateral airspace opacities concerning for multifocal pneumonia.	Describes worsening bilateral opacities concerning for multifocal pneumonia.	Describes multifocal airspace opacification throughout both lungs concerning for pneumonia.

The examples show that EvoRG captures improvement, worsening, and newly emerging findings while preserving relevant current findings. In the stable case, EvoRG retains the absence of acute cardiopulmonary abnormalities. These observations illustrate its ability to describe clinically meaningful interval changes and current findings.