VISIT-ORDER EXCHANGE EQUIVARIANCE MODELING FOR REGION-LEVEL LONGITUDINAL REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated radiology report generation (RRG) has proven its potential to reduce the workload of radiologists. However, most existing RRG methods rely on single-visit data, overlooking the longitudinal changes in the disease. While existing longitudinal methods track disease progression, they typically rely on global features that limit fine-grained regional correspondence, and they are prone to temporal template shortcuts due to visit-order sensitivity. To address this, we propose VEE-RG, a longitudinal radiology report generation framework that generates visit-order stable reports. Specifically, to overcome spatial misalignments and missed detections caused by patient posture variations, we design a paired-region stabilization module to construct reliable visual correspondence. Furthermore, we propose an exchange-equivariant change factorization module that decouples longitudinal differences into symmetric and antisymmetric components, effectively mitigating order sensitivity and temporal shortcuts. Finally, anonymized region-level contrastive learning is introduced to align modalities, preventing shortcut matching via anatomical names and enforcing authentic pathology learning. Extensive experiments on two large public datasets demonstrate that VEE-RG exhibits more stable temporal modeling capabilities and outperforms competitive methods, achieving a 1.7% improvement in CheXbert F1-score on the MIMIC-CXR dataset and a 4.1% improvement in F1-score on the Chest ImaGenome dataset. Our code is available at `https://anonymous.4open.science/r/VEE-RG-C1F4`.

1 INTRODUCTION

Chest X-ray (CXR) is one of the most commonly used medical imaging examinations in clinical practice and is widely utilized for the evaluation and diagnosis of structures such as the lungs and heart (Johnson et al., 2019; Beddiar et al., 2023). Radiologists typically need to examine anatomical regions in the images and subsequently use standardized text to describe their clinical findings and diagnostic opinions (Goergen et al., 2013). However, manually writing reports is highly dependent on professional experience and is time-consuming (Hosny et al., 2018). Doctors not only need to describe whether abnormalities exist in the images, but also need to accurately articulate the anatomical regions, extent, and severity of the abnormalities, and, when necessary, compare them with previous visits to track disease progression (Zhou et al., 2026). As the volume of medical imaging examinations continues to increase, automated radiology report generation (RRG) has begun to be used to alleviate the workload of radiology departments (Rennie et al., 2017). RRG generates diagnostic reports by automatically analyzing medical imaging data, thereby optimizing clinical workflows and helping to improve the diagnostic efficiency of radiologists.

Recent methods have attempted to utilize previous images for longitudinal radiology report generation (LRRG) (Bannur et al., 2023; Zhu et al., 2023; Yun et al., 2025). Existing mainstream LRRG methods usually integrate current and previous visits through techniques such as feature concatenation and attention fusion, while some methods further use prior reports or clinical contexts to assist the model in generation. However, these methods often rely on image-level visual features to generate reports, making it difficult to accurately focus on subtle longitudinal changes in abnormal regions between multiple visits, which often indicate the trajectory of disease evolution. Although a small

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.

number of methods have attempted to introduce local visual features into longitudinal models, they fail to construct stably paired regional visual representations (Dalla Serra et al., 2023; Gao et al., 2026).

Existing longitudinal report generation methods often struggle with visit-order variations. Constrained by fixed chronological inputs, they are prone to relying on template shortcuts and superficial data biases rather than capturing true temporal disease dynamics. Therefore, reliable methods typically require stronger temporal discrimination capabilities to faithfully track pathology progression. To achieve such temporal sensitivity, exchange equivariance offers a promising paradigm, as preserving the effects of temporal transformations has proven crucial in video representations (Jenni & Jin, 2021) and longitudinal disease prediction (Emre et al., 2024) over strict invariance.

Inspired by this, we propose **VEE-RG**: **V**isit-order **E**xchange **E**quivariance Modeling for Region-Level Longitudinal **R**eport **G**eneration. VEE-RG first extracts regional visual features from both historical and current images using an anatomical detector. Then, it leverages shared anatomical queries, paired detection gating, and a global feature fallback mechanism to construct stable paired-region visual representations. On this basis, we introduce an exchange-equivariant change factorization mechanism, performing symmetric and antisymmetric factorization on the paired-region visual features, thereby imposing an exchange-equivariant inductive bias on the longitudinal change representations. Subsequently, we employ anonymized longitudinal region-level contrastive learning, combined with bidirectional InfoNCE constraints, to align the visual and textual representations of the same anatomical region. Finally, VEE-RG generates and assembles descriptions for each region according to a fixed anatomical region order, thereby forming a complete longitudinal radiology report.

Our main contributions are summarized as follows:

- We propose VEE-RG, a region-level longitudinal report generation framework that generates visit-order stable reports.
- To establish reliable regional visual correspondence and model longitudinal disease changes, we design a paired-region stabilization module and a visit-order exchange-equivariant change factorization module. Cross-modal alignment is designed via an anonymized region-level contrastive learning strategy to prevent shortcut matching via anatomical names and enforce authentic pathology learning.
- Extensive experiments on the MIMIC-CXR and Chest ImaGenome datasets show that VEE-RG improves the clinical F1-score by 1.7% and 4.1%, respectively. Exchange equivariance analyses further validate VEE-RG's stable temporal modeling capabilities.

2 Related Work

2.1 Regional Report Generation

Region-related report generation includes region-aware methods that exploit anatomical cues and methods that explicitly generate descriptions at the region level. ITA uses a multi-head transformer, with each head generating a sentence for a specific anatomical region (Wang et al., 2022). COMG uses multi-organ masks and organ-specific disease priors to guide visual feature fusion, while MCA-RG uses anatomy and pathology concepts to guide report generation; both generate a complete report rather than separate region-wise descriptions (Gu et al., 2024; Xing et al., 2025). In contrast, RGRG detects predefined regions and generates descriptions that are concatenated into a full report, offering interactivity and interpretability (Tanida et al., 2023). AP-ISG further uses iterative scene-graph generation and attribute prototype learning to model relationships between anatomical structures and disease findings (Zhang et al., 2024). Our method is most closely related to these region-level methods. However, they are limited to single-visit settings and cannot capture longitudinal changes between corresponding anatomical regions across consecutive visits.

2.2 Longitudinal Radiology Report Generation

Longitudinal report generation leverages prior patient visits to provide temporal context for the current visit. Early methods like CheXRelNet and BioViL-T modeled cross-temporal disease evolution

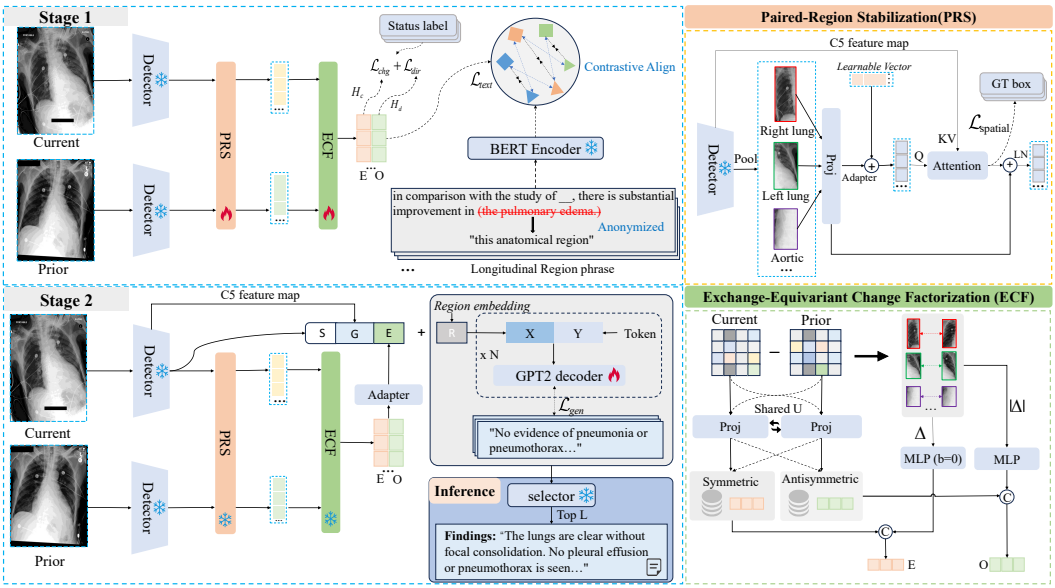


Figure 1: Overview of VEE-RG. Stage 1: Pre-trains longitudinal regional representations via contrastive vision-language alignment. Stage 2: With the ECF and PRS modules frozen, trains GPT-2 decoder to generate and assemble region-level descriptions into the final findings. Snowflake : frozen; fire : trainable.

(Karwande et al., 2022; Bannur et al., 2023), while subsequent generation methods typically employ attention fusion, multimodal constraints, or temporal encoding. For example, HERGen models extended patient histories via a causal Transformer (Wang et al., 2024); MLRG fuses historical data using multi-view contrastive learning (Liu et al., 2025); Diff-RRG guides the language model with extracted image differences (Yun et al., 2025); and PriorRG combines clinical context-guided contrastive learning with coarse-to-fine decoding (Liu et al., 2026). Libra further models temporal differences between current and prior images through a Temporal Alignment Connector (Zhang et al., 2025). However, these methods neglect region-level longitudinal representations. While MARE explicitly aligns lesion-level visual features across time (Gao et al., 2026) and Dalla Serra et al. concatenate equivalent anatomical regions across scans (Dalla Serra et al., 2023), these methods inadequately process fine-grained longitudinal changes and fail to ensure model stability against input variations.

3 Method

3.1 Overview

Figure 1 illustrates the VEE-RG network architecture. We divide the network into two core representation modules: (1) the Paired-Region Stabilization module (PRS), and (2) the Exchange-equivariant Change Factorization module (ECF). Given a patient's prior X-ray image x^p (if available) and current X-ray image x^c, let $p_r = h_{p,r}$ and $c_r = h_{c,r}$ denote the PRS representations of region r in the prior and current visits, respectively. VEE-RG aims to model paired anatomical regions, generate a sentence for each region through a shared decoder, and combine them in a fixed anatomical order to form the final report, incorporating visit-order exchange equivariance modeling to ensure generation stability.

3.2 Paired-Region Stabilization Module (PRS)

Object Detector. Following RGRG (Tanida et al., 2023), we use Faster R-CNN with a ResNet-50 backbone to extract region-level visual features from chest X-ray images. The detector predicts bounding boxes, confidences, and ROI features for predefined anatomical regions, which are ar-

ranged in RGRG's predefined order. We also retain the high-level C5 feature map of the detector. The detector remains frozen during subsequent training.

Standard detectors are vulnerable to localization errors and missed detections. Furthermore, independent processing of consecutive images often causes spatial misalignments of identical regions due to patient posture or imaging variations. To ensure stable representations, PRS treats predicted boxes as soft spatial priors on the global feature map. Specifically, each region has a learnable query q_r to compute global spatial attention. For $t \in \{p, c\}$, let $f_{t,r}$ be the ROI feature and $F_{t,k}$ the C5 feature at cell k. Let $B_{t,r,k}$ be the fraction of cell k covered by region r's bounding box, and let $g_r \in \{0, 1\}$ indicate valid detection in both visits. We set $B_{t,r,k} = 0$ if $g_r = 0$. The spatial attention is:

$$a_{t,r,k} = \mathrm{softmax}_k \left(\frac{(q_r + g_r W_q f_{t,r})^\top (W_{\mathrm{C5}} F_{t,k} + P_k)}{\sqrt{d}} + \lambda B_{t,r,k} \right), \tag{1}$$

Here, P_k is a learnable positional encoding; W_q and W_r map the ROI feature to query and residual spaces. W_{C5} projects C5 for attention keys/values, and λ is a learned box-bias coefficient. The regional representation aggregates global attention and the ROI residual:

$$h_{t,r} = \mathrm{LN} \left(\sum_k a_{t,r,k} W_{\mathrm{C5}} F_{t,k} + g_r W_r f_{t,r} \right), \tag{2}$$

If $g_r = 0$, PRS generates a fallback representation via the region query and global feature map. Otherwise, the ROI feature modulates the query and provides a local residual. Shared PRS parameters across visits ensure regional representations lie in the same feature space.

3.3 Exchange-Equivariant Change Factorization Module (ECF)

Given stably aligned prior and current regional features from PRS, we define their signed feature difference, explicit change magnitude, and direction features for region r:

$$\Delta_r = c_r - p_r, \qquad m_r = M(|\Delta_r|), \qquad d_r = D(\Delta_r), \tag{3}$$

Here, M and D are MLP encoders; D uses bias-free linear layers and a Tanh activation. Thus, the direction representation negates upon visit exchange. Since feature differences alone discard the original state and compositional information, we process the ordered pairs (p_r, c_r) and (c_r, p_r) through a shared encoder $U(\cdot)$ to obtain contextual representations u_r^{pc} and u_r^{cp}. These are decomposed into symmetric and antisymmetric contexts:

$$s_r = \frac{u_r^{pc} + u_r^{cp}}{2}, \qquad a_r = \frac{u_r^{pc} - u_r^{cp}}{2}. \tag{4}$$

We concatenate these into an exchange-symmetric component $E_r = [m_r, s_r]$ and an exchange-antisymmetric component $O_r = [d_r, a_r]$. To supervise them, two lightweight MLP heads process E_r and O_r for change and direction classification on valid region sets $\mathcal{V}_{\mathrm{chg}}$ and $\mathcal{V}_{\mathrm{dir}}$, respectively:

$$\mathcal{L}_{\mathrm{chg}} = \frac{1}{|\mathcal{V}_{\mathrm{chg}}|} \sum_{r \in \mathcal{V}_{\mathrm{chg}}} \mathrm{BCE}(z_r^{\mathrm{chg}}, y_r^{\mathrm{chg}}), \qquad \mathcal{L}_{\mathrm{dir}} = \frac{1}{|\mathcal{V}_{\mathrm{dir}}|} \sum_{r \in \mathcal{V}_{\mathrm{dir}}} \mathrm{BCE}(z_r^{\mathrm{dir}}, y_r^{\mathrm{dir}}), \tag{5}$$

where y_r^{chg} and y_r^{dir} are the corresponding ground-truth labels. Finally, E and O are passed to the report decoder to provide longitudinal change and direction information.

Exchange Equivariance. PRS and ECF jointly enable visit-order exchange equivariance: for an ordered regional pair (p_r, c_r), ECF decomposes the longitudinal representation to satisfy $E(c_r, p_r) = E(p_r, c_r)$ and $O(c_r, p_r) = -O(p_r, c_r)$. To enforce this, ECF shares parameters across both visit orders, uses a bias-free odd mapping for directional representations, and forms components via averaging and differencing. The direction head computes the half-difference of its outputs on O and $-O$ to ensure classification equivariance. **Importantly, this exchange-equivariant design is not intended for real-world clinical scenarios where images are manually swapped, as clinical inputs strictly follow chronological order. Instead, it acts as a crucial structural inductive bias for regularization.** Without such a constraint, longitudinal models are prone to learning spurious positional shortcuts—generating temporal descriptions based on input slots rather than genuine visual comparisons of anatomical regions. The symmetric-antisymmetric decomposition forces the model

to derive change existence and direction purely from pairwise visual evidence. Consequently, our subsequent swap-input experiments demonstrate that models without this constraint suffer a large performance gap upon swapping inputs. This gap reveals a heavy reliance on positional shortcuts, which leads to fragile generalization under real-world distribution shifts (e.g., noisy prior scans or inter-hospital variations) even when the chronological order is correctly maintained.

3.4 Anonymized Longitudinal Region-level Contrastive Learning

For each region r with valid paired detections and an explicit change label, we use its corresponding regional text as supervision. To prevent the model from directly matching visual regions through anatomical names, we anonymize the text by replacing specific names with a generic phrase ("this anatomical region"). We then extract the text embedding t_j using a frozen BERT (Liu et al., 2021) followed by a projection layer. For the visual modality, we concatenate the ECF outputs (E_r and O_r) and project them into a regional visual embedding v_r. Given a temperature τ, the matching probability between v_r and t_j is calculated as:

$$q_{rj}^{v\to t} = \frac{\exp(v_r^\top t_j/\tau)}{\sum_{\ell}\exp(v_r^\top t_\ell/\tau)}. \tag{6}$$

For a batch of M regional pairs, we construct a multi-positive contrastive loss according to the canonical ID:

$$\mathcal{L}_{v\to t} = -\frac{1}{M}\sum_{r=1}^{M}\sum_{j=1}^{M} p_{rj}\log q_{rj}^{v\to t}, \tag{7}$$

Here, p_{rj} is the positive target distribution; text j is positive for region r when the two samples share a canonical region ID. Repeated anonymized sentences are assigned frequency-normalized weights when constructing this distribution. The final objective averages this with a symmetrically computed text-to-image loss:

$$\mathcal{L}_{\text{text}} = \frac{1}{2}\left(\mathcal{L}_{v\to t} + \mathcal{L}_{t\to v}\right). \tag{8}$$

The training objective in the longitudinal pretraining stage (Stage 1) consists of the contrastive loss, change auxiliary supervision, and the spatial consistency constraint of PRS:

$$\mathcal{L} = \lambda_T\mathcal{L}_{\text{text}} + \lambda_C\mathcal{L}_{\text{chg}} + \lambda_D\mathcal{L}_{\text{dir}} + \lambda_S\mathcal{L}_{\text{spatial}}, \tag{9}$$

where $\mathcal{L}_{\text{spatial}}$ supervises the corresponding spatial attention using the normalized coverage distribution of valid bounding boxes on the C5 grid, and is computed exclusively for these valid boxes; λ_T, λ_C, λ_D, and λ_S are scalar loss weights.

3.5 Region-level Report Generation

Following the RGRG setting (Tanida et al., 2023), we use the 355M-parameter GPT-2 Medium model fine-tuned on PubMed abstracts as our report generator. To generate the report, we project the 1024-dimensional current ROI features to match the GPT-2 hidden dimension. Meanwhile, we globally pool the C5 features and add learnable region identity encodings. The ECF representation is converted into an EVENT residual by a trainable EventAdapter. Thus, the EVENT token of each region r is defined as:

$$\begin{aligned} EVENT_r &= W_o\left[W_e\mathrm{LN}(E_r);\ kW_d\mathrm{LN}(O_r)\right], \\ M_r &= [S_r, EVENT_r, G, R_r] \in \mathbb{R}^{4\times 1024}, \end{aligned} \tag{10}$$

Here, S_r, G, and R_r denote the projected current ROI feature, pooled C5 feature, and learnable region identity/position encoding, respectively; W_e, W_d, and W_o are trainable projections implementing the EventAdapter, and k scales the order-sensitive component. If a valid prior examination is unavailable, we duplicate the current image as a dummy prior, zeroing out and masking the EVENT token. For each region with textual supervision, the text sequence begins with the BOS token, and the loss is computed against the ground-truth regional phrase:

$$\mathcal{L}_{\text{gen}} = -\sum_{t}\log p(y_t \mid y_{<t}, M_r). \tag{11}$$

During inference, to reduce descriptions of irrelevant regions, we employ a region selector (initialized with RGRG) to select the high-confidence top-L regions and feed them into GPT-2 to generate regional descriptions. These descriptions are then concatenated and deduplicated to form the final report following RGRG.

4 EXPERIMENTS

4.1 SETTINGS

Datasets. We evaluate our method on two datasets: (1) MIMIC-CXR (Johnson et al., 2019), the largest public dataset for report generation; and (2) Chest ImaGenome (Wu et al., 2021), which contains automatically constructed scene graphs for the MIMIC-CXR images. Each scene graph maps a frontal chest X-ray to 29 anatomical bounding boxes with corresponding report sentences when available. For paired visits, it further provides region-level longitudinal phrases and change-state labels. Following official splits, the MIMIC dataset comprises 159,434/1,268/2,422 (train/val/test) samples, while the ImaGenome split, processed following RGRG, contains 113,915/16,498/32,711 samples. For longitudinal pretraining and ablations, we extract patients with $\geq$ 2 visits from ImaGenome, maintaining official splits. The final non-empty "Findings" section serves as the reference text. All images are resized to a maximum dimension of 512 and converted to grayscale.

Metrics. Model performance is evaluated using standard NLG metrics, including BLEU (Papineni et al., 2002), METEOR (Denkowski & Lavie, 2014), and ROUGE-L (Lin, 2004). For clinical efficacy (CE), we report precision, recall, and F1-scores across 14 and 5 disease categories using CheXbert (Smit et al., 2020). We also measure Temporal Entity F1 (Tempo F1) (Zhang et al., 2025), which calculates the intersection of predefined temporal radiological keywords between predictions and references. To ensure fair baseline comparisons, text cleaning is applied exclusively to the MIMIC split.

Implementation Details. Following RGRG, the 29 predefined anatomical regions are adopted, with each regional phrase limited to 64 tokens. The hidden dimensions are set to 256 for PRS, 512 for ECF, and 1,024 for the decoder. Both training stages employ the AdamW optimizer with a learning rate of $1e{-}4$ and cosine decay. Stage 1 trains for 12 epochs with a batch size of 32, while stage 2 trains for 8 epochs with a batch size of 6. During inference, beam search with a beam width of 3 is employed, and the top $L = 8$ regions are selected to assemble the final report. The loss weights are configured as $(\lambda_T, \lambda_C, \lambda_D, \lambda_S) = (0.4, 1.0, 0.7, 0.15)$, and the scaling factor is set to $k = 0.3$.

4.2 BASELINES

We compare our model with 13 competitive methods: (1) SI (single-image input): RGRG (Tanida et al., 2023), AP-ISG (Zhang et al., 2024), CRRG-CLIP (Xu et al., 2024), PromptMRG (Jin et al., 2024), DDGIP (Huang et al., 2025), CvT-DistilGPT2 (Nicolson et al., 2023), MAN (Shen et al., 2024), MCA-RG (Xing et al., 2025), FMVP (Liu et al., 2023). (2) LI (longitudinal input): CXRMate (Nicolson et al., 2024), HERGen (Wang et al., 2024), MLRG (Liu et al., 2025), and PriorRG (Liu et al., 2026).

4.3 MAIN RESULTS

Comparison with SOTA methods. As shown in Tables 1 and 2, under the official Chest ImaGenome split, VEE-RG achieves the best results on most metrics. VEE-RG outperforms RGRG (with the same backbone) by 1% in BLEU-4 and 4.2% in clinical F1, though RGRG yields a higher ROUGE-L, likely because it does not employ top-L region filtering. Compared with DDGIP, it improves clinical F1 by 4.1% and surpasses the SOTA models in 5-class clinical precision, F1-5, and Tempo F1. On the MIMIC-CXR split, VEE-RG achieves the highest BLEU-1, Recall, and F1. Specifically, it surpasses DDGIP by 4.6% in overall Recall (and 11% in 5-class Recall), while outperforming PriorRG by 1.7% in clinical F1. Furthermore, VEE-RG slightly underperforms PriorRG on Tempo F1 and BLEU-4, likely because PriorRG's additional clinical context and coarse-to-fine decoding favor temporal keywords and n-gram overlap, whereas VEE-RG prioritizes regional clinical coverage and exchange stability.

Table 1: Comparison with representative methods on the Chest ImaGenome and MIMIC-CXR test splits. Best results are in bold. B-1/B-4: BLEU-1/BLEU-4; MTR: METEOR; R-L: ROUGE-L; P_{14}, R_{14}, $F1_{14}$: 14-class CheXbert metrics; *: reproduced results.

Dataset	Input	Method	Venue	NLG Metrics ↑				CE Metrics ↑		
				B-1	B-4	MTR	R-L	P_{14}	R_{14}	$F1_{14}$
Chest ImaGenome	SI	RGRG*	CVPR23	0.350	0.120	0.171	**0.283**	0.476	0.528	0.501
	SI	AP-ISG*	TMI24	0.374	0.115	0.156	0.249	0.400	0.529	0.456
	SI	CRRG-CLIP	ArXiv24	0.241	0.078	0.109	0.239	-	-	-
	SI	DDGIP*	NAACL25	**0.401**	0.090	0.172	0.280	0.452	0.564	0.502
	SI	CvT-DistilGPT2*	AIM23	0.338	0.111	0.151	0.286	0.518	0.420	0.464
	LI	**VEE-RG (Ours)**	-	0.390	**0.130**	**0.174**	0.268	**0.519**	**0.569**	**0.543**
MIMIC-CXR	SI	PromptMRG	AAAI24	0.398	0.112	0.157	0.268	0.501	0.509	0.476
	SI	MAN	AAAI24	0.396	0.115	0.151	0.274	0.411	0.398	0.389
	SI	MCA-RG	MICCAI25	0.411	0.128	0.163	0.300	0.472	0.406	0.408
	SI	DDGIP*	NAACL25	0.411	0.094	0.156	0.275	0.502	0.560	0.529
	SI	FMVP	TMM23	0.389	0.108	0.150	0.284	0.332	0.383	0.336
	LI	CXRMate	IMU24	0.361	0.108	0.159	0.263	0.495	0.367	0.422
	LI	HERGen*	ECCV24	0.374	0.112	0.149	0.282	0.557	0.425	0.482
	LI	MLRG	CVPR25	0.411	0.158	0.176	0.320	0.549	0.468	0.505
	LI	PriorRG*	AAAI26	0.395	**0.167**	**0.185**	**0.324**	**0.576**	0.501	0.536
	LI	**VEE-RG (Ours)**	-	**0.412**	0.116	0.164	0.275	0.508	**0.606**	**0.553**

Table 2: Comparison of five-class clinical and temporal performance on the official Chest ImaGenome and MIMIC-CXR test splits. Best results are in bold. P_5, R_5, and $F1_5$: CheXbert metrics over five disease categories; * indicates our reproduced results.

Dataset	Input	Method	Venue	CE-5 Metrics ↑			Temporal ↑
				P_5	R_5	$F1_5$	Tempo F1
Chest ImaGenome	SI	RGRG*	CVPR23	0.491	0.617	0.547	0.379
	SI	AP-ISG*	TMI24	0.536	0.535	0.535	0.380
	SI	DDGIP*	NAACL25	0.423	**0.689**	0.524	0.331
	LI	**VEE-RG (Ours)**	-	**0.550**	0.630	**0.588**	**0.407**
MIMIC-CXR	SI	DDGIP*	NAACL25	0.543	0.627	0.582	0.243
	LI	HERGen*	ECCV24	0.590	0.467	0.521	0.239
	LI	PriorRG*	AAAI26	**0.603**	0.555	0.578	**0.282**
	LI	**VEE-RG (Ours)**	-	0.540	**0.737**	**0.623**	0.268

4.4 Ablation Study

We conduct ablation studies on the MIMIC-CXR test split to evaluate our model. Additionally, we ablate the auxiliary tasks and antisymmetric component on the ImaGenome subset (≥ 2 visits) under the same split.

Effect of the structures and generator inputs. As shown in Table 3, removing PRS leads to consistent drops in both BLEU-1 and BLEU-4, validating its contribution to stable region-level modeling. Replacing both PRS and ECF with an MLP further degrades BLEU-4 to 0.108, Precision to 0.498, and Tempo F1 to 0.260. This indicates that without explicit regional alignment and interaction, visual representations struggle to retain local lesion evidence. Removing the global C5 branch reduces Recall and clinical F1, indicating that C5 features help prevent omissions of clinical findings. Removing the EVENT branch causes a severe 3.9% absolute drop in Tempo F1 (from 0.268 to 0.229), confirming its critical role in capturing temporal changes. Finally, omitting the current-image ROI branch degrades BLEU-1/4 and F1 but leaves Tempo F1 unchanged, suggesting it primarily provides stable static visual evidence rather than directional change information.

Effect of auxiliary tasks. As shown in Table 4, removing $\mathcal{L}_{\text{chg}}$ lowers change accuracy and F1 from 0.632 and 0.682 to 0.440 and 0.505, respectively. Removing $\mathcal{L}_{\text{dir}}$ has little effect on change recognition but reduces Direction F1 from 0.783 to 0.521. Excluding $\mathcal{L}_{\text{text}}$ raises change accuracy

Table 3: Ablation study of the model structure and generator inputs on the MIMIC-CXR test.

Group	Variant	B-1	B-4	P_{14}	R_{14}	$F1_{14}$	$F1_5$	Tempo F1
Struct.	w/o PRS	0.401	0.109	0.504	0.604	0.549	0.621	0.264
	w/o PRS+ECF	0.406	0.108	0.498	0.603	0.545	0.618	0.260
Input	w/o C5	0.406	0.109	0.507	0.596	0.548	0.622	0.265
	w/o EVENT	0.405	0.110	0.502	0.583	0.539	0.620	0.229
	w/o ROI	0.370	0.098	0.499	0.537	0.517	0.589	**0.268**
	Full	**0.412**	**0.116**	**0.508**	**0.606**	**0.553**	**0.623**	**0.268**

Table 4: Auxiliary-tasks ablation on the ImaGenome subset ($\geq$ 2 visits) of the MIMIC test split. State F1 is the macro-F1 for final regional-state prediction. Chg: Change; Dir: Direction.

Variant	Chg Acc.	Chg F1	Dir F1	State F1
w/o $\mathcal{L}_{\text{chg}}$	0.440	0.505	0.776	0.412
w/o $\mathcal{L}_{\text{dir}}$	0.612	0.681	0.521	0.454
w/o $\mathcal{L}_{\text{text}}$	**0.653**	0.656	0.778	0.579
Full	0.632	**0.682**	**0.783**	**0.588**

Table 5: Antisymmetric component interventions on the ImaGenome subset ($\geq$ 2 visits) of the MIMIC test split. T-F1: Tempo F1.

Method	B-1	B-4	$F1_{14}$	$F1_5$	T-F1
PriorRG	0.363	**0.115**	0.549	0.595	**0.270**
HERGen	0.353	0.094	0.489	0.525	0.209
VEE-RG	**0.407**	0.108	**0.573**	**0.644**	0.265
$O = 0$	0.406	0.106	0.566	0.639	0.250
$\lvert O \rvert$	0.404	0.106	0.566	0.640	0.254
$-O$	0.401	0.104	0.566	0.640	0.237

but lowers Change F1 and State F1, indicating that the contrastive objective supports robust regional-state discrimination.

Effect of the exchange-antisymmetric component. Table 5 shows that VEE-RG outperforms HERGen across all reported metrics and improves over PriorRG in BLEU-1, F1-14, and F1-5, though PriorRG retains higher BLEU-4 and Tempo F1. Removing O reduces F1-14 and Tempo F1. Retaining only its magnitude partially recovers Tempo F1, whereas reversing its sign yields the lowest BLEU-1, BLEU-4, and Tempo F1, underscoring the importance of directional consistency. Additionally, PriorRG exhibits a substantial drop in NLG metrics on the ImaGenome longitudinal subset ($\geq$ 2 visits), indicating its high sensitivity and instability on longitudinal data.

4.5 Exchange Equivariance Analysis

Given forward inputs, models generate the current report; upon swapping the inputs, they generate the prior report. We conduct experiments on common valid visit pairs in the official MIMIC test split, comparing VEE-RG against its No-EVENT variant, PriorRG, and HERGen (with the last two visits in its sequence swapped).

Exchange stability. We define the Absolute Gap as the absolute difference in metric scores between forward and reverse generations. A smaller value indicates that the model is more stable after the input is exchanged. As shown in Figure 2(a), VEE-RG yields most balanced gaps. Notably, incorporating EVENT halves the Tempo F1 gap (from 2.91% to 1.48%), indicating it significantly mitigates temporal perturbations caused by input swapping. In contrast, HERGen and PriorRG suffer from severe and irregular performance fluctuations, highlighting VEE-RG's structural stability against visit-order changes.

Exchange specificity. We define Temporal Reference Discrimination (TRD; detailed in Appendix B.3), which yields higher scores against the reference from the intended time point than against the alternative. As shown in Figure 2(b), VEE-RG outperforms the No-EVENT baseline by 1.83%, 2.05%, and 1.60% in F1-14, F1-5, and Tempo F1, respectively. It also consistently surpasses HERGen and PriorRG across all metrics, indicating more accurate time-point attribution. We further explore the role of EVENT in temporal modeling, with details provided in Appendix B.1.

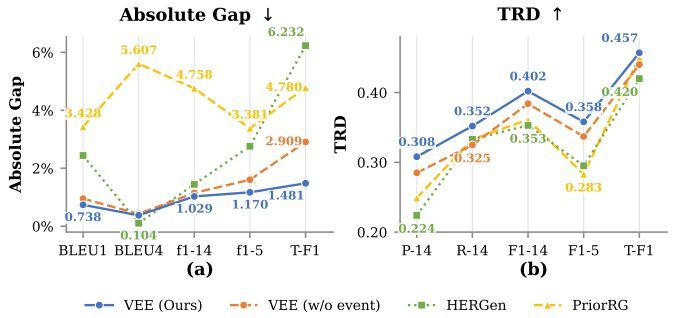


Figure 2: Visit-order exchange analysis on common valid pairs. (a) Absolute Gap between forward and reverse generations; (b) Temporal Reference Discrimination (TRD).

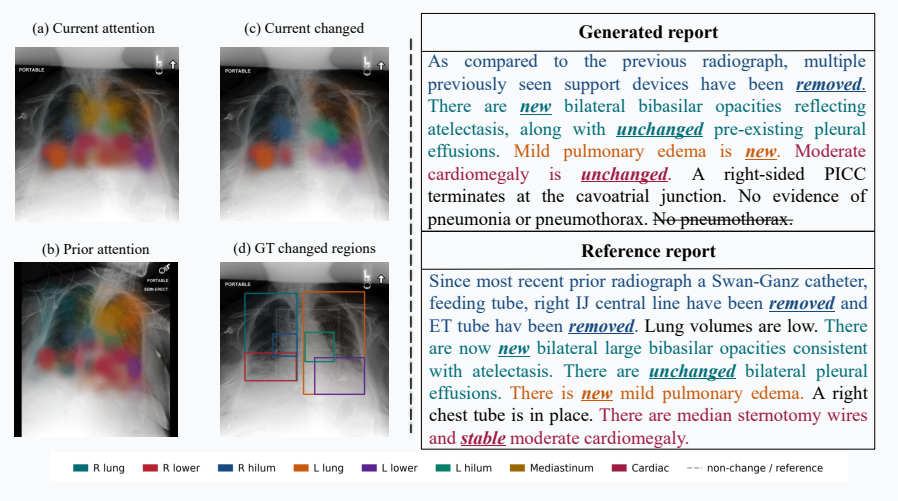


Figure 3: Qualitative case study. (a,b) ECF attention maps for the current and prior visits; (c) regions selected by the predicted change gate; (d) GT changed-region support boxes. The right panel compares the generated and reference reports. Text colors indicate anatomical regions, and underlined terms highlight temporal cues.

4.6 Case Analysis

Figure 3 (left) presents anatomy-specific attention maps alongside ground-truth change boxes. Highlighted regions with predicted $p(\text{change}) > 0.5$ closely align with the GT, illustrating that PRS and ECF effectively ground longitudinal evidence within relevant anatomical regions. Figure 3 (right) demonstrates VEE-RG accurately capturing key longitudinal findings—including support device removal, new bibasilar opacities with atelectasis, new mild pulmonary edema, and stable cardiomegaly. However, a repeated pneumothorax statement reflects remaining challenges. We further provide an exchange case analysis in Appendix B.4.

5 Conclusions

We propose VEE-RG, a region-level radiology report generation model that generates stable longitudinal reports. By constructing stably paired regional visual representations and employing an exchange-equivariant change factorization module, VEE-RG effectively decouples longitudinal changes. Extensive experiments demonstrate that VEE-RG produces highly stable reports under visit-order exchange while accurately attributing clinical changes to their proper time points. Future work will explore longer multi-visit histories and more robust region-level modeling.

AI USE STATEMENT

Large language models were leveraged to assist with manuscript drafting and polishing. Specifically, we utilized LLMs to refine wording, enhance readability, and ensure clear presentation across all sections of this paper. The model supported tasks including sentence rephrasing, grammar correction, and improvement of the overall logical flow of the text. We have reviewed all AI-assisted 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 work complies with the ICLR Code of Ethics. Neither human-subject research nor animal experiments are involved in this study. All datasets employed are obtained in accordance with relevant usage guidelines without any privacy violations. During the research process, every effort has been made to mitigate biases and discriminatory outcomes. No personally identifiable information is utilized, and no experiments posing potential privacy or security risks have been conducted. We uphold transparency and research integrity throughout the entire study.

REPRODUCIBILITY STATEMENT

We have made every reasonable effort to guarantee the reproducibility of the experimental results reported in this paper. Our code is publicly available in an anonymous repository to facilitate reproduction. The experimental setup, including training configurations and parameter settings, is elaborated in the main text. We also provide a comprehensive description of VEE-RG to help other researchers replicate our experiments.

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.

Djamila-Romaissa Beddiar, Mourad Oussalah, and Tapio Seppänen. Automatic captioning for medical imaging (mic): a rapid review of literature. *Artificial intelligence review*, 56(5):4019–4076, 2023.

Francesco Dalla Serra, Chaoyang Wang, Fani Deligianni, Jeff Dalton, and Alison O'Neil. Controllable chest x-ray report generation from longitudinal representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 4891–4904, 2023.

Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pp. 376–380, 2014.

Taha Emre, Arunava Chakravarty, Antoine Rivail, Dmitrii Lachinov, Oliver Leingang, Sophie Riedl, Julia Mai, Hendrik PN Scholl, Sobha Sivaprasad, Daniel Rueckert, et al. 3DTINC: Time-equivariant non-contrastive learning for predicting disease progression from longitudinal octs. *IEEE Transactions on Medical Imaging*, 43(9):3200–3210, 2024.

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, 2026.

Stacy K Goergen, Felicity J Pool, Tari J Turner, Jane E Grimm, Mark N Appleyard, Carmel Crock, Michael C Fahey, Michael F Fay, Nicholas J Ferris, Susan M Liew, et al. Evidence-based guideline for the written radiology report: Methods, recommendations and implementation challenges. *Journal of medical imaging and radiation oncology*, 57(1):1–7, 2013.

Tiancheng Gu, Dongnan Liu, Zhiyuan Li, and Weidong Cai. Complex organ mask guided radiology report generation. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7980–7989. IEEE, 2024.

Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H Schwartz, and Hugo JWL Aerts. Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8):500–510, 2018.

Chentao Huang, Guangli Li, Xinjiong Zhou, Yafeng Ren, and Hongbin Zhang. DDGIP: Radiology report generation through disease description graph and informed prompting. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 3884–3894, 2025.

Simon Jenni and Hailin Jin. Time-equivariant contrastive video representation learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9950–9960. IEEE, 2021.

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*, volume 38, 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.

Gaurang Karwande, Amarachi B Mbakwe, Joy T Wu, Leo A Celi, Mehdi Moradi, and Ismini Lourentzou. CheXRelNet: An anatomy-aware model for tracking longitudinal relationships between chest x-rays. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 581–591. Springer, 2022.

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

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13748–13757. IEEE, 2021.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10348–10359. IEEE, 2025.

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.

Zhizhe Liu, Zhenfeng Zhu, Shuai Zheng, Yawei Zhao, Kunlun He, and Yao Zhao. From observation to concept: A flexible multi-view paradigm for medical report generation. *IEEE Transactions on Multimedia*, 26:5987–5995, 2023.

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

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.

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.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1179–1195. IEEE, 2017.

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

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7433–7442. IEEE, 2023.

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.

Lin Wang, Munan Ning, Donghuan Lu, Dong Wei, Yefeng Zheng, and Jie Chen. An inclusive task-aware framework for radiology report generation. In *International conference on medical image computing and computer-assisted intervention*, pp. 568–577. Springer, 2022.

Joy T Wu, Nkechinyere N Agu, Ismini Lourentzou, Arjun Sharma, Joseph A Paguio, Jasper S Yao, Edward C Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. Chest imagenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*, 2021.

Qilong Xing, Zikai Song, Youjia Zhang, Na Feng, Junqing Yu, and Wei Yang. MCA-RG: Enhancing llms with medical concept alignment for radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 380–390. Springer, 2025.

Jianfei Xu, Thanet Markchom, and Huizhi Liang. CRRG-CLIP: Automatic generation of chest radiology reports and classification of chest radiographs. *arXiv preprint arXiv:2501.01989*, 2024.

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.

Ke Zhang, Yan Yang, Jun Yu, Jianping Fan, Hanliang Jiang, Qingming Huang, and Weidong Han. Attribute prototype-guided iterative scene graph for explainable radiology report generation. *IEEE transactions on medical imaging*, 43(12):4470–4482, 2024.

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.

Shaoyang Zhou, Yingshu Li, Yunyi Liu, Lingqiao Liu, Lei Wang, and Luping Zhou. A review of longitudinal radiology report generation: Dataset composition, methods, and performance evaluation. *Medical Image Analysis*, pp. 104304, 2026.

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 Additional Experiments

A.1 Event Scaling

Event scaling. Table 6 evaluates the scaling factor k for the order-sensitive component on the MIMIC-CXR test split. Setting $k = 0$ removes this component, while larger values amplify its contribution. A moderate $k = 0.3$ achieves the best BLEU and clinical F1 scores, striking a favorable balance between static content and longitudinal changes. Increasing to $k = 1$ maximizes Tempo F1 but slightly degrades other metrics, suggesting a trade-off between temporal sensitivity and general report quality.

Table 6: Effect of the odd-component scaling factor k on the MIMIC-CXR test split.

k	BLEU-1	BLEU-4	$F1_{14}$	$F1_5$	Temporal F1
0	0.406	0.114	0.541	0.619	0.270
0.3	**0.412**	**0.116**	**0.553**	**0.623**	0.268
0.5	0.405	0.114	0.543	0.623	0.265
1	0.408	0.115	0.547	0.621	**0.272**

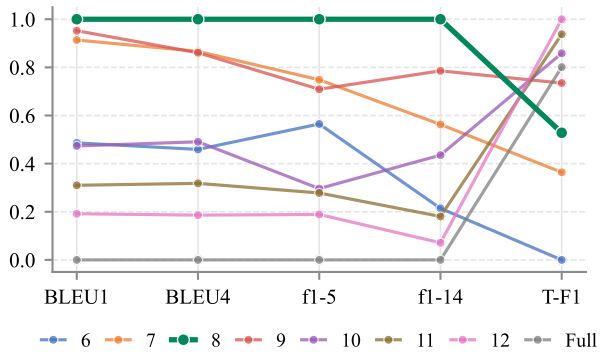


Figure 4: Top-L region-selection ablation on the MIMIC-CXR test split. We vary the number of detector-valid regions selected by the selector and report BLEU-1, BLEU-4, CheXbert F1-14, CheXbert F1-5, and Tempo F1. For visualization, each metric is independently min–max normalized across all settings to $[0, 1]$; the plotted values therefore show relative changes rather than absolute scores.

A.2 Top-L Region Selection

Figure 4 evaluates the number of regions selected during inference. For visualization, each metric is independently min–max normalized across all settings to $[0, 1]$. The $L = 8$ setting reaches the maximum normalized scores for BLEU-1, BLEU-4, F1-14, and F1-5, whereas increasing L beyond 8 generally improves the normalized Tempo F1 but reduces the other metrics. This indicates that $L = 8$ provides the best balance between sufficient regional evidence and irrelevant or redundant inputs.

B Additional Analysis

B.1 Bidirectional Temporal Attribution

To further examine the contribution of EVENT, we evaluate the two visit orders on the ImaGenome subset ($\geq$ 2 visits) of the MIMIC test split. Here, forward denotes the original prior-to-current order and reverse denotes the swapped current-to-prior order. A direction is counted as correct when the generated report obtains a higher Tempo F1 against its corresponding reference than against the other reference. "Forward correct" and "Reverse correct" report the marginal percentages of correct decisions in each order. "Both," "Forward only," "Reverse only," and "Neither" are mutually exclusive pair categories determined by the two decisions; higher values are preferred for the correct categories, whereas lower values are preferred for the one-direction-only and neither categories. As shown in Table 7, EVENT raises the forward and reverse attribution rates by 0.46 and 2.74 percentage points, respectively, and increases the proportion of pairs correctly attributed in both directions from 10.05% to 13.70%. The proportion correct in only one direction decreases from 68.04% to 63.93%, indicating that EVENT makes temporal attribution more consistent under visit-order exchange.

Table 7: Bidirectional temporal attribution under EVENT ablation on the ImaGenome subset ($\geq$ 2 visits) of the MIMIC test split. Entries are percentages. A direction is correct when the corresponding-reference Tempo F1 is higher than the noncorresponding-reference Tempo F1. Arrows indicate the preferred direction.

Variant	Forward correct ↑	Reverse correct ↑	Both correct ↑	Forward only ↓	Reverse only ↓	Neither correct ↓
VEE-RG	**46.12**	**45.21**	**13.70**	**32.42**	**31.51**	22.37
w/o EVENT	45.66	42.47	10.05	35.62	32.42	**21.92**
Δ (pp)	+0.46	+2.74	+3.65	-3.20	-0.91	+0.45

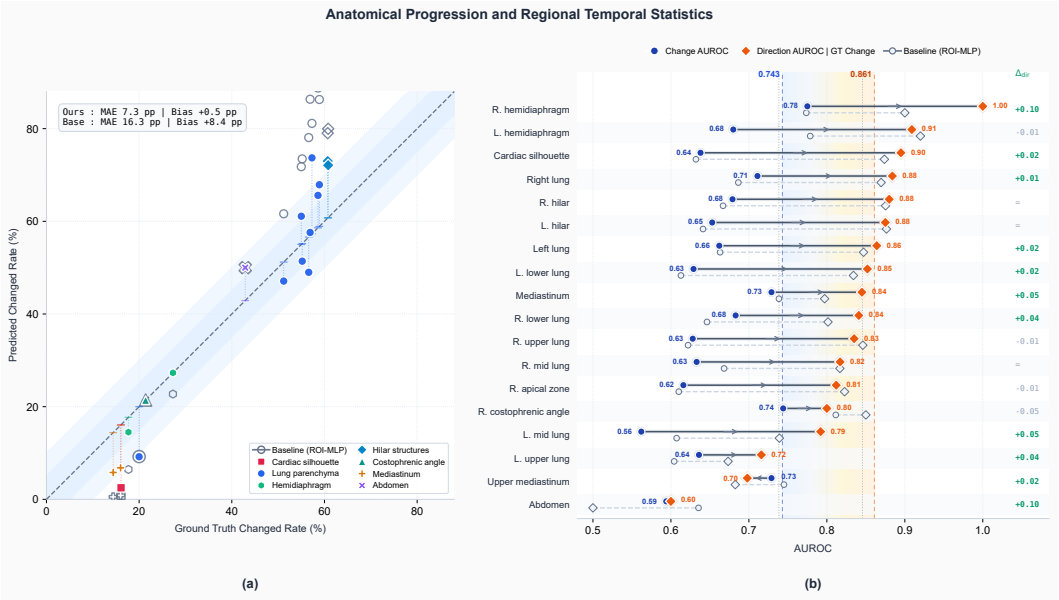


Figure 5: Regional change statistics for 18 common regions. (a) GT and predicted change rates for anatomical regions, with the ROI-MLP baseline shown for comparison. (b) Region-wise change and direction AUROC of VEE-RG and the baseline.

B.2 Regional Change Analysis

To further examine the region-level longitudinal change modeling ability of VEE-RG, we analyze change frequency and discrimination performance across anatomical regions on the ImaGenome subset ($\geq$ 2 visits). Figure 5(a) compares the ground-truth and predicted change rates of each region. VEE-RG obtains a mean absolute error of 7.3 percentage points, compared with 16.3 percentage points for the ROI-MLP baseline, and its mean bias is reduced from $+8.4$ to $+0.5$ percentage points.

Figure 5(b) reports the region-wise AUROC for change detection and direction discrimination. VEE-RG achieves mean AUROCs of 0.743 and 0.861, respectively, exceeding the ROI-MLP baseline values of 0.738 and 0.846. The results also reveal differences across anatomical regions: direction discrimination is relatively strong for structures such as the hemidiaphragm, hilum, and lung parenchyma, whereas fine-grained regions with less support show greater variation. These findings indicate that the model captures regional change patterns while retaining the expected dependence of performance on regional sample size and difficulty.

B.3 Exchange Metrics

Let y^p and y^c denote the reference reports from the prior and current visits, respectively. Given a model G, the forward and reverse inputs are (x^p, x^c) and (x^c, x^p). Their generated reports are

denoted by

$$\hat{y}^c = G(x^p, x^c), \qquad \hat{y}^p = G(x^c, x^p), \tag{12}$$

where the superscript indicates the time point that the generated report is intended to describe. For any metric m for which a larger value indicates better agreement, we define the corresponding-reference scores as

$$s_m^{\mathrm{fwd}} = m(\hat{y}^c, y^c), \qquad s_m^{\mathrm{rev}} = m(\hat{y}^p, y^p). \tag{13}$$

Absolute Gap. The Absolute Gap measures the change in performance after exchanging the visit order:

$$\mathrm{AG}_m = \left| s_m^{\mathrm{fwd}} - s_m^{\mathrm{rev}} \right|. \tag{14}$$

A smaller Absolute Gap indicates greater stability under visit-order exchange.

Temporal Reference Discrimination. To measure whether a generated report favors the reference from its intended time point, we compare each generation with both references. For metric m, we first construct the similarity matrix

$$\mathbf{M}_i = \begin{bmatrix} m(\hat{y}_i^c, y_i^c) & m(\hat{y}_i^c, y_i^p) \\ m(\hat{y}_i^p, y_i^c) & m(\hat{y}_i^p, y_i^p) \end{bmatrix}. \tag{15}$$

Concretely, for each paired visit i, we generate $\hat{y}_i^c$ from (x_i^p, x_i^c) and $\hat{y}_i^p$ from the exchanged input (x_i^c, x_i^p), evaluate both generations against both references, and perform the two comparisons independently before averaging the resulting binary outcomes over all pairs.

For a test set of N paired visits, we define TRD as

$$\mathrm{TRD}_m = \frac{1}{2N} \sum_{i=1}^{N} \left[\mathbb{I}(M_{i,11} > M_{i,12}) + \mathbb{I}(M_{i,22} > M_{i,21}) \right]. \tag{16}$$

Here, $\mathbb{I}(\cdot)$ equals 1 when its condition is true and 0 otherwise. Thus, a higher score for the corresponding reference contributes to TRD, whereas ties and preference for the other reference contribute zero. A larger TRD indicates stronger time-point specificity. We report TRD separately for each metric, with clinical CE and Tempo F1 being more informative than NLG-based TRD, which primarily reflects lexical overlap.

B.4 Exchange Case Analysis

To qualitatively assess temporal behavior under visit-order exchange, we compare VEE-RG with PriorRG on the same paired CXR study from the MIMIC-CXR ImaGenome test subset. As shown in Figure 6, both models receive the prior and current images in chronological and exchanged orders. VEE-RG exhibits four interpretable temporal behaviors: pleural effusion changes from increased to decreased, pulmonary edema is described as improved, while cardiomegaly and basilar opacity/atelectasis remain stable. In contrast, PriorRG shows no clear directional reversal and only one weakly persistent cardiac finding. This case illustrates VEE-RG's ability to separate order-sensitive disease changes from exchange-stable content.

B.5 Dataset and Pair-Coverage Statistics

Tables 8 and 9 summarize the dataset composition and the effective coverage of the visit-order reversal evaluation. The longitudinal-pair counts are reported separately from static-sample counts because the former require patients with at least two visits. The raw MIMIC-CXR longitudinal test list contains 2,260 pair rows; model-specific validity filtering yields the effective counts in Table 9, and all three-model comparisons use their strict 219-pair intersection.

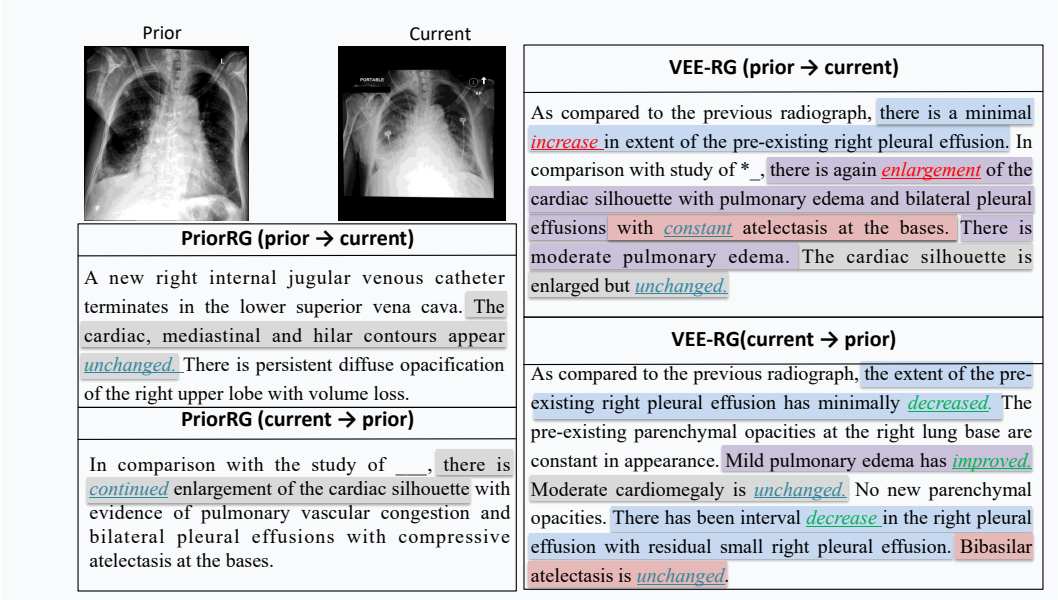


Figure 6: Exchange-case analysis on the ImaGenome subset (≥ 2 visits) of the MIMIC test split. We compare VEE-RG with PriorRG under the original visit order (prior $\rightarrow$ current) and the exchanged visit order (current $\rightarrow$ prior).

Table 8: Distribution of the ImaGenome-derived data under the official MIMIC-CXR and Chest ImaGenome splits. Static samples count individual images, whereas longitudinal pairs are restricted to patients with ≥ 2 visits and count eligible prior–current visit pairs.

Split	Unit	Train	Val.	Test	Total
MIMIC-CXR	Static samples	159,434	1,268	2,422	163,124
	Longitudinal pairs	108,207	934	2,260	111,401
Chest ImaGenome	Static samples	113,915	16,498	32,711	163,124
	Longitudinal pairs	77,945	11,249	22,207	111,401

Table 9: Coverage of the visit-order reversal evaluation on the ImaGenome subset (≥ 2 visits) of the MIMIC test split. A valid reversed pair is a pair for which the model-specific preprocessing provides both visit orders and their corresponding references. The strict common subset is the intersection across all three models and is used for every three-model comparison.

Method	Valid reversed pairs
VEE-RG	1,537
HERGen	254
PriorRG	1,074
Common subset	219