LONGITUDINAL RADIOLOGY REPORT GENERATION VIA PROGRESSION ALIGNMENT AND PATHOLOGY SEMANTIC LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Longitudinal radiology report generation requires representations that capture both clinical findings and their changes across examinations. To address these complementary requirements, we propose a unified framework for cross-modal progression alignment and pathology semantic enhancement. We represent progression as the residual between current and prior representations within each modality. Through temporal direction alignment, we encourage consistent directions of change between visual and textual representations, while auxiliary static alignment maintains image–report consistency within each examination. Cross-modal consistency alone, however, does not explicitly associate visual features with clinical findings. To establish these associations, a label-induced inverse optimal transport objective jointly learns pathology prototypes and a visual adapter, providing semantic guidance to the features used for generation. The resulting representations combine current visual evidence with prior image–report context to condition a frozen language model. Experiments on Longitudinal-MIMIC demonstrate strong performance in report generation and clinical agreement, with ablation studies and representation analyses supporting the complementary roles of progression alignment and explicit pathology semantic supervision.

1 INTRODUCTION

Radiology reports describe the findings of an imaging examination in the context of a patient's clinical history. When prior examinations are available, comparison with the current images helps distinguish persistent findings from newly developed, improved, or resolved abnormalities (Dalla Serra et al., 2023; Nicolson et al., 2024). Therefore, longitudinal radiology report generation incorporates this comparative context by using current images together with prior images and reports (Zhu et al., 2023), which requires representations that capture both the clinical state of each examination and the changes between examinations.

Existing approaches have explored multimodal fusion and explicit temporal reasoning to incorporate longitudinal evidence into report generation. For example, Zhu et al. (2023) integrate current images with prior images and reports through cross-attention and a memory-based decoder. More recent methods model disease evolution directly: BiOTPrompt (Liu et al., 2026) uses bidirectional optimal transport to identify evolving regions and construct prompts, while MARE aligns lesion-level features and regularizes visual and textual evolution through analogical reasoning (Gao et al., 2026b). These developments highlight the importance of representing clinical change across examinations. Building on this direction, we investigate how examination-level visual and textual representations can be jointly organized to express both consistent temporal changes and explicit pathology semantics.

As illustrated in Fig. 1, our central hypothesis is that visual and textual representations of the same longitudinal examination pair should exhibit consistent directions of change in a learned shared space. Although images and reports provide different observations, both characterize the clinical evolution between the prior and current examinations (Zhang et al., 2022; Wang et al., 2022). To modeling this consistency, we represent progression within each modality as the residual between its current and prior global representations and introduce a contrastive objective that aligns the paired

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.

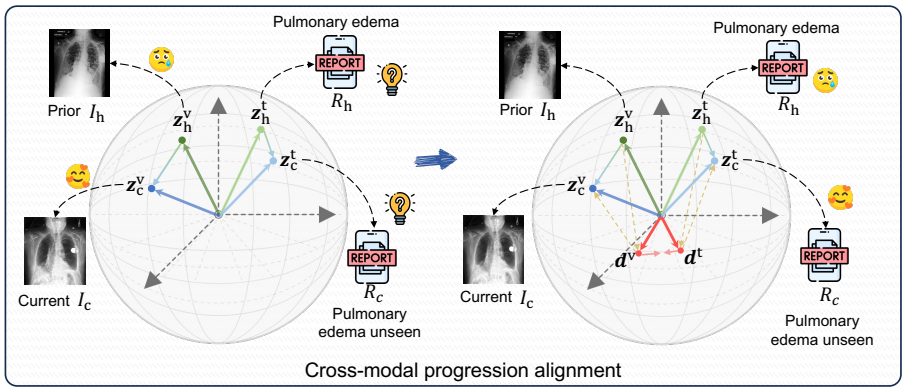


Figure 1: Motivation for cross-modal progression alignment. Images and reports describe the same clinical evolution between prior and current examinations. We therefore encourage their intra-modal temporal residuals to align in direction, promoting consistent representations of clinical change across modalities.

residual directions. Focusing on direction allows the two modalities to encode changes with different magnitudes. A complementary static contrastive objective aligns images and reports for each examination, jointly supervising examination-level semantics and cross-modal progression.

Beyond temporal consistency, the clinical basis of generated descriptions must also be considered. As we know, end-to-end image-to-report generation alone offers limited insight into the clinical findings underlying its predictions (Jain et al., 2021; Delbrouck et al., 2024). Studies using anatomical regions and structured clinical entities and relations have demonstrated the value of intermediate evidence for improving report quality and interpretability (Tanida et al., 2023; Dalla Serra et al., 2022). These findings motivate explicit modeling of individual clinical observations within the visual representation, providing a semantic basis for the descriptions in the final report. Accordingly, we propose a data-driven pathology semantic enhancement strategy that uses the 14 observation labels provided with MIMIC-CXR to supervise the learning of corresponding pathology prototypes. Through an inverse optimal transport (IOT) objective (Stuart & Wolfram, 2020; Li et al., 2019b), the model learns image-to-prototype assignments that match the observed labels, encouraging visual features to encode explicit pathology-level semantics for report generation.

Together, progression alignment and pathology semantic enhancement form a unified representation-learning framework for longitudinal radiology report generation, termed **PAPE**, linking cross-modal consistency with explicit clinical semantics. Within this framework, compact current and prior representations are integrated through temporal attention and provided to a frozen language model, enabling report generation to draw on both current visual evidence and longitudinal context. Experiments on Longitudinal-MIMIC evaluate the framework's effectiveness, while ablation studies and representation analyses examine how the proposed objectives contribute to report quality, temporal alignment, and pathology-related semantics. The main contributions are as follows:

- We propose a longitudinal cross-modal progression alignment method that represents clinical change as intra-modal temporal residuals. Direction-level contrastive learning aligns these residuals across images and reports, while static alignment maintains semantic consistency within each examination.
- We propose a data-driven pathology semantic enhancement strategy in which label-supervised prototypes associate visual representations with clinical observation categories. To do this, we establish an inverse optimal transport objective to learn disease prototypes, providing explicit pathology-level supervision alongside progression alignment.
- Experiments on Longitudinal-MIMIC demonstrate the effectiveness of the combined framework for report generation. Complementary ablation studies and representation analyses characterize the roles of the auxiliary objectives in temporal alignment, pathology semantics, and report quality.

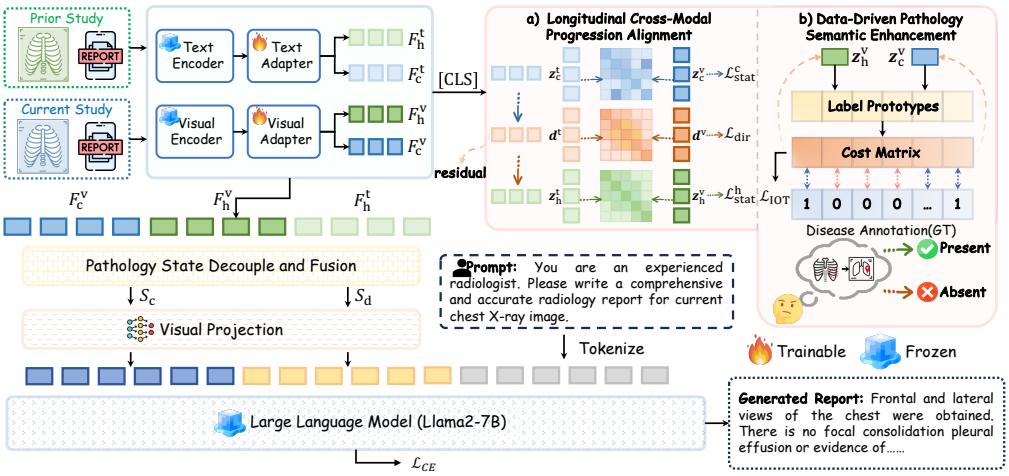


Figure 2: Overall framework of the proposed PAPE, which comprises longitudinal cross-modal progression alignment, data-driven pathological semantic enhancement, and pathology state decoupling and fusion module for radiology report generation task.

2 METHODS

2.1 FRAMEWORK OVERVIEW AND PROBLEM FORMULATION

A longitudinal radiology report should describe the current findings and how they relate to the patient's prior examination. Given current chest radiographs I_{c} and a prior examination $(I_{\mathrm{h}}, R_{\mathrm{h}})$ from the same patient, we formulate this task as

$$\widetilde{R}_{\mathrm{c}} = f_\theta(I_{\mathrm{c}}, I_{\mathrm{h}}, R_{\mathrm{h}}), \tag{1}$$

where f_θ combines current visual evidence with historical clinical context to generate the current report. During training, we adopt the ground-truth report R_{c} both as the generation target and to supervise representation learning, whereas inference requires only the current images and the prior examination. As illustrated in Fig. 2, the framework builds on visual and textual representations extracted by frozen modality-specific encoders and refined through lightweight MLP-based adapters. The resulting token sequences, denoted by $\boldsymbol{F}_r^m$, support both auxiliary representation learning and downstream report generation, with the first token of each sequence serving as its global representation $\boldsymbol{z}_r^m$; the subscript $r \in \{\mathrm{h}, \mathrm{c}\}$ indexes the prior and current examinations, while the superscript $m \in \{\mathrm{v}, \mathrm{t}\}$ denotes the visual and textual modalities, respectively. Longitudinal cross-modal progression alignment then encourages the image and report representations to agree on both the state represented by each examination and the direction of change between examinations. Because this agreement alone does not explicitly connect the learned features to clinical findings, we introduce label-supervised prototypes through pathology semantic enhancement, providing the visual representations with additional guidance about what those findings are.

For report generation, separate Perceiver modules compress the current visual features $\boldsymbol{F}_{\mathrm{c}}^{\mathrm{v}}$ and the prior image–report features $(\boldsymbol{F}_{\mathrm{h}}^{\mathrm{v}}, \boldsymbol{F}_{\mathrm{h}}^{\mathrm{t}})$ into compact latent representations, after which temporal cross-attention conditions the current representation on the prior examination. Concatenating the current-state and history-conditioned representations preserves current visual evidence alongside longitudinal context. The resulting sequence is projected into the language model's embedding space and, together with a fixed prompt, conditions the frozen language model to generate the current report.

2.2 LONGITUDINAL CROSS-MODAL PROGRESSION ALIGNMENT

For a longitudinal pair of examinations from the same patient, the change in each modality's representation characterizes the progression from the prior examination to the current one. Using the

global representations $\boldsymbol{z}_r^m$, we express these intra-modal temporal residuals as

$$\boldsymbol{d}^{\mathrm{v}} = \boldsymbol{z}_{\mathrm{c}}^{\mathrm{v}} - \boldsymbol{z}_{\mathrm{h}}^{\mathrm{v}}, \qquad \boldsymbol{d}^{\mathrm{t}} = \boldsymbol{z}_{\mathrm{c}}^{\mathrm{t}} - \boldsymbol{z}_{\mathrm{h}}^{\mathrm{t}}. \tag{2}$$

Similarly, for the same examination, the cross-modal residual $\boldsymbol{e}_r, r \in \{\mathrm{h}, \mathrm{c}\}$ is defined as

$$\boldsymbol{e}_r = \boldsymbol{z}_r^{\mathrm{t}} - \boldsymbol{z}_r^{\mathrm{v}}, r \in \{\mathrm{h}, \mathrm{c}\} \tag{3}$$

Then, we subtract the visual residual from the textual residual and regrouping the terms:

$$\begin{aligned} \boldsymbol{d}^{\mathrm{t}} - \boldsymbol{d}^{\mathrm{v}} &= (\boldsymbol{z}_{\mathrm{c}}^{\mathrm{t}} - \boldsymbol{z}_{\mathrm{c}}^{\mathrm{v}}) - (\boldsymbol{z}_{\mathrm{h}}^{\mathrm{t}} - \boldsymbol{z}_{\mathrm{h}}^{\mathrm{v}}) \\ &= \boldsymbol{e}_{\mathrm{c}} - \boldsymbol{e}_{\mathrm{h}}, \end{aligned} \tag{4}$$

or, equivalently, $\boldsymbol{d}^{\mathrm{t}} - \boldsymbol{d}^{\mathrm{v}} - \boldsymbol{e}_{\mathrm{c}} + \boldsymbol{e}_{\mathrm{h}} = \mathbf{0}$. Thus, the difference between the modality-specific progression residuals is determined by the change in cross-modal residual between examinations.

Corollary (Progression consistency).
If the cross-modal residual remains unchanged between the prior and current examinations, i.e., $\boldsymbol{e}_{\mathrm{c}} = \boldsymbol{e}_{\mathrm{h}}$, then $\boldsymbol{d}^{\mathrm{t}} = \boldsymbol{d}^{\mathrm{v}}$. For nonzero residuals, their normalized directions therefore coincide:

$$\frac{\boldsymbol{d}^{\mathrm{t}}}{\|\boldsymbol{d}^{\mathrm{t}}\|_2} = \frac{\boldsymbol{d}^{\mathrm{v}}}{\|\boldsymbol{d}^{\mathrm{v}}\|_2}. \tag{5}$$

This corollary follows directly from Eq. (4); the identity itself does not guarantee directional agreement when the cross-modal residual varies. Since the images and reports describe the same underlying clinical evolution, we adopt directional consistency as a modeling objective for their progression residuals. This is a weaker requirement than residual equality, allowing the two modalities to encode changes with different magnitudes.

Temporal direction alignment. To compare residual directions, we use stabilized normalization $\phi(\boldsymbol{x}) = \boldsymbol{x}/\max(\|\boldsymbol{x}\|_2, \eta)$, where $\eta > 0$ is a small constant that keeps zero and near-zero residuals well-defined. For paired collections $\mathcal{X} = (\boldsymbol{x}_i)_{i=1}^N$ and $\mathcal{Y} = (\boldsymbol{y}_i)_{i=1}^N$ in a batch of N longitudinal examination pairs, define

$$\ell_{\mathrm{NCE}}(\mathcal{X} \to \mathcal{Y}) = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\big(\phi(\boldsymbol{x}_i)^\top \phi(\boldsymbol{y}_i)/\tau\big)}{\sum_{j=1}^{N} \exp(\phi(\boldsymbol{x}_i)^\top \phi(\boldsymbol{y}_j)/\tau)}. \tag{6}$$

In Eq. (6), matching indices identify positive pairs, while other batch members serve as negatives. The temperature $\tau > 0$ is learnable, initialized to 0.07, and shared by the direction and static alignment objectives. Using the symmetric loss $\mathcal{C}(\mathcal{X}, \mathcal{Y}) = \frac{1}{2}[\ell_{\mathrm{NCE}}(\mathcal{X} \to \mathcal{Y}) + \ell_{\mathrm{NCE}}(\mathcal{Y} \to \mathcal{X})]$, we align the residual collections $\mathcal{D}^m = (\boldsymbol{d}_i^m)_{i=1}^N$ through

$$\mathcal{L}_{\mathrm{dir}} = \mathcal{C}(\mathcal{D}^{\mathrm{v}}, \mathcal{D}^{\mathrm{t}}). \tag{7}$$

This objective encourages cross-modal agreement in progression direction for each examination pair while distinguishing it from other pairs in the batch.

Auxiliary static alignment. Alongside direct progression alignment, examination-level image–report alignment provides a complementary constraint on the endpoint representations. Eq. (4) implies

$$\|\boldsymbol{d}^{\mathrm{t}} - \boldsymbol{d}^{\mathrm{v}}\|_2 \le \|\boldsymbol{e}_{\mathrm{c}}\|_2 + \|\boldsymbol{e}_{\mathrm{h}}\|_2, \tag{8}$$

which reflects cross-modal residuals at the two endpoints. Our static contrastive objective compares normalized representations, providing angular semantic alignment rather than directly minimizing the unnormalized cross-modal residuals. Static alignment therefore provides complementary examination-level supervision, while the direction objective explicitly aligns temporal changes across modalities. Considering a representation collection of N samples, $\mathcal{Z}_r^m = (\boldsymbol{z}_{r,i}^m)_{i=1}^N$, we define

$$\mathcal{L}_{\mathrm{stat}}^r = \mathcal{C}(\mathcal{Z}_r^{\mathrm{v}}, \mathcal{Z}_r^{\mathrm{t}}), \tag{9}$$

and average the losses for the prior and current examinations:

$$\mathcal{L}_{\mathrm{stat}} = \tfrac{1}{2}(\mathcal{L}_{\mathrm{stat}}^{\mathrm{h}} + \mathcal{L}_{\mathrm{stat}}^{\mathrm{c}}). \tag{10}$$

Together, these objectives encourage consistent progression directions across modalities while maintaining semantic agreement within each examination.

2.3 DATA-DRIVEN PATHOLOGY SEMANTIC ENHANCEMENT

The objectives in Section 2.2 encourage cross-modal agreement in examination states and temporal changes, but do not explicitly associate visual representations with individual clinical findings. To establish these semantic correspondences, we introduce $K = 14$ learnable prototypes $\boldsymbol{P} = (\boldsymbol{p}_j)_{j=1}^{K}$, one for each supplied observation category, including "No Finding" and "Support Devices". We look forward to introducing fine-grained chest disease prototypes modeling to enhance the semantic validity of the overall image report. Given the observed fine-grained disease labels, the prototype learning problem naturally matches the bi-level optimization problem. Inspired by inverse optimal transport (IOT), in this section, we propose a data-driven fine-grained pathological semantic representation enhancement strategy.

Image-to-prototype assignments. For examination $r \in \{\mathrm{h}, \mathrm{c}\}$, the transport cost $[\boldsymbol{C}_r]_{i,j}$ of assigning the i-th visual representation to prototype j is defined by their negative cosine similarity:

$$[\boldsymbol{C}_r]_{i,j} = -\phi(\boldsymbol{z}_{r,i}^{\mathrm{v}})^{\top}\phi(\boldsymbol{p}_j). \tag{11}$$

Let $y_{r,i,j}$ denote the supplied numerical label and $\mathcal{G}_{r,i} = \{j : y_{r,i,j} > 0.5\}$ the set of positive labels. To exclude examples without positive labels from this auxiliary supervision, we assign each example a row mass $a_{r,i}/N$, where $a_{r,i} = \mathbf{1}[|\mathcal{G}_{r,i}| > 0]$, N is the batch size, and $\mathbf{1}[A] = 1$ if condition A holds true. Then, the entropy-regularized optimal transport problem is

$$\begin{aligned} \boldsymbol{T}_r^* &= \underset{\boldsymbol{T}\in\mathbb{R}_+^{N\times K}}{\arg\min} \big\{\langle \boldsymbol{C}_r, \boldsymbol{T}\rangle - \epsilon H(\boldsymbol{T})\big\}, \\ \text{subject to} \quad & \sum_{j=1}^{K}[\boldsymbol{T}]_{i,j} = \frac{a_{r,i}}{N}, \qquad i = 1, \ldots, N, \end{aligned} \tag{12}$$

where $H(\boldsymbol{T}) = -\sum_{i,j}[\boldsymbol{T}]_{i,j}\log[\boldsymbol{T}]_{i,j}$ with $0\log 0 = 0$, and $\epsilon > 0$ controls entropy regularization. The total mass assigned to each prototype is not fixed in advance, the problem therefore separates across examples and admits the closed-form solution

$$[\boldsymbol{T}_r^*]_{i,j} = \frac{a_{r,i}}{N} q_{r,i,j}, \qquad q_{r,i,j} = \frac{\exp(-[\boldsymbol{C}_r]_{i,j}/\epsilon)}{\sum_{k=1}^{K}\exp(-[\boldsymbol{C}_r]_{i,k}/\epsilon)}. \tag{13}$$

Here $q_{r,i,j}$ is the predicted assignment probability for prototype j and rows corresponding to examples without positive labels carry zero mass.

Label-induced inverse optimal transport. Given observed labels, we can define an empirical assignment $\boldsymbol{T}_r^{\mathrm{e}}$ that distributes each example's mass equally among its positive categories:

$$t_{r,i,j} = \frac{\mathbf{1}[j \in \mathcal{G}_{r,i}]}{\max(|\mathcal{G}_{r,i}|, 1)}, \qquad [\boldsymbol{T}_r^{\mathrm{e}}]_{i,j} = \frac{1}{N} t_{r,i,j}. \tag{14}$$

This construction gives $\boldsymbol{T}_r^{\mathrm{e}}$ and $\boldsymbol{T}_r^*$ identical row masses, including zero rows for examples without positive labels. Given these target assignments, inverse transport learns the parameters that determine $\boldsymbol{C}_r$, rather than optimizing a free cost matrix. Specifically, the visual adapter and prototypes are trained to minimize

$$D_{\mathrm{KL}}(\boldsymbol{T}_r^{\mathrm{e}} \,\|\, \boldsymbol{T}_r^*) = \sum_{i,j}[\boldsymbol{T}_r^{\mathrm{e}}]_{i,j}\log\frac{[\boldsymbol{T}_r^{\mathrm{e}}]_{i,j}}{[\boldsymbol{T}_r^*]_{i,j}}, \tag{15}$$

where terms with zero empirical mass contribute zero. Substituting Eqs. 13 and 14 and discarding parameter-independent terms yields

$$\mathcal{L}_{\mathrm{IOT}}^{r} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} t_{r,i,j}\log q_{r,i,j}, \qquad \mathcal{L}_{\mathrm{IOT}} = \tfrac{1}{2}\big(\mathcal{L}_{\mathrm{IOT}}^{\mathrm{h}} + \mathcal{L}_{\mathrm{IOT}}^{\mathrm{c}}\big). \tag{16}$$

We can find that the row-constrained IOT objective is thus equivalent to cross-entropy against normalized positive-label targets and requires no iterative transport solver. The overall IOT optimization objective naturally consists of two parts that are from priori and current examinations. Note that examples without positive labels contribute zero loss, while normalization remains over the full batch. Please refer to Appendix A.1 for detailed derivation of IOT.

By matching image-to-prototype assignments to the observed labels, our proposed IOT objective encourages visual representations to reflect the clinical findings associated with each examination, providing explicit pathology-level supervision for report generation.

2.4 Longitudinal Semantic State Compression and Decoding

Longitudinal report generation requires current visual evidence and context from the prior examination (Liu et al., 2026; Gao et al., 2026b). Motivated by work on visual-token redundancy and adaptive visual processing (Fan et al., 2026; Zhu et al., 2026), we use separate Perceiver modules to compress the current image features $\boldsymbol{F}_{\mathrm{c}}^{\mathrm{v}}$ and prior image–report features $(\boldsymbol{F}_{\mathrm{h}}^{\mathrm{v}}, \boldsymbol{F}_{\mathrm{h}}^{\mathrm{t}})$ into compact states $\boldsymbol{S}_{\mathrm{c}}$ and $\boldsymbol{S}_{\mathrm{h}}$. Temporal cross-attention then conditions $\boldsymbol{S}_{\mathrm{c}}$ on $\boldsymbol{S}_{\mathrm{h}}$, producing a history-conditioned current representation $\boldsymbol{S}_{\mathrm{d}}$. Concatenating $\boldsymbol{S}_{\mathrm{c}}$ and $\boldsymbol{S}_{\mathrm{d}}$ preserves current visual evidence alongside longitudinal context and the combined sequence is projected into the language model's embedding space and normalized to form the conditioning sequence $\boldsymbol{S}_{\mathrm{v}}$. Together with a fixed report-generation prompt $\mathbf{p}$, $\boldsymbol{S}_{\mathrm{v}}$ conditions the frozen language model to generate the current report. We optimize the token-level negative log-likelihood over the non-padding target-report positions $\mathcal{T}$ in a training batch:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{|\mathcal{T}|} \sum_{(i,l)\in\mathcal{T}} \log P_{\mathrm{LM}}\big(w_{\mathrm{c},i}^{l} \mid w_{\mathrm{c},i}^{<l}, \boldsymbol{S}_{\mathrm{v},i}, \mathbf{p}\big)\,, \tag{17}$$

where $w_{\mathrm{c},i}^{l}$ denotes the l-th token of sample i's current report and $w_{\mathrm{c},i}^{<l}$ its preceding tokens. P_{LM} denotes the conditional next-token probability distribution defined by the frozen language model, given the preceding report tokens, the conditioning sequence $\boldsymbol{S}_{\mathrm{v},i}$, and the fixed prompt $\mathbf{p}$. The overall training objective combines report generation with the auxiliary representation-learning losses:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda_1 \mathcal{L}_{\mathrm{stat}} + \lambda_2 \mathcal{L}_{\mathrm{dir}} + \lambda_3 \mathcal{L}_{\mathrm{IOT}}, \tag{18}$$

where λ_1, λ_2, and λ_3 control the contributions of the auxiliary objectives. The modality-specific encoders and language model remain frozen during training. Detailed module specifications are provided in Appendix A.2.

3 Experiments

3.1 Experimental Settings

Datasets.

We conduct our experiments on the Longitudinal-MIMIC dataset (Zhu et al., 2023), which is derived from MIMIC-CXR (Johnson et al., 2019). MIMIC-CXR contains 377,110 chest radiographs and 227,835 radiology reports from 65,379 patients. Longitudinal-MIMIC consists of longitudinal visit pairs from the same patient, with each sample containing two consecutive visits. Following the official split, the dataset contains 92,374 training samples from 26,156 patients, 737 validation samples from 203 patients, and 2,058 test samples from 266 patients.

Evaluation Metrics.

To comprehensively evaluate the quality of the generated reports, we assess model performance using natural language generation (NLG) metrics and clinical efficacy (CE) metrics. For NLG metrics, we use BLEU-1 (**B-1**) through BLEU-4 (**B-4**) (Papineni et al., 2002), ROUGE-L (**R-L**) (Lin, 2004), and METEOR (**MTR**) (Banerjee & Lavie, 2005). For CE metrics, we use CheXbert (Smit et al., 2020) to extract 14 clinical observation labels from the generated reports and compare them with the ground-truth labels to calculate precision (**P**), recall (**R**), and F1-score (**F1**).

Implementation Details.

In our study, all experiments are conducted on two NVIDIA H100 GPUs. We use Rad-DINO (Pérez-García et al., 2025) as the image encoder, CXR-BERT (Boecking et al., 2022) as the text encoder, and LLaMA2-7B (Touvron et al., 2023) as the large language model. All of λ_1, λ_2, and λ_3 are initially set to 0.01. We train the model for 3 epochs using the AdamW optimizer with an initial learning rate of 3×10^{-4}, a batch size of 22, and a cosine annealing learning-rate schedule. During inference, we employ beam search with a beam size of 3 to balance computational efficiency and generation quality.

3.2 Comparison with State-of-the-Art Methods

The comparison covers single-examination methods that generate reports from current imaging and longitudinal methods that incorporate prior examination data. Together, these baselines include

Table 1: Experimental results on the Longitudinal-MIMIC dataset. The best results are **bolded**, and the second-best results are underlined. TIM-I and TIM-II denote the two stages of TIM, reported separately in the original paper.

Input	Method	Year	Venue	NLG Metrics						CE Metrics		
				B-1	B-2	B-3	B-4	MTR	R-L	P	R	F1
Single-examination	R2Gen(Chen et al., 2020b)	2020	EMNLP	0.353	0.218	0.145	0.103	0.142	0.277	0.333	0.273	0.276
	R2GenCMN(Chen et al., 2021)	2021	ACL	0.353	0.218	0.148	0.106	0.142	0.278	0.334	0.275	0.278
	R2GenGPT(Wang et al., 2023)	2023	Meta-Rad	0.365	0.237	0.163	0.117	0.136	0.277	0.341	0.312	0.325
	PromptMRG(Jin et al., 2024)	2024	AAAI	0.398	-	-	0.112	0.157	0.268	0.501	**0.509**	0.476
	EKAGen(Bu et al., 2024)	2024	CVPR	0.415	0.254	0.166	0.117	0.154	0.285	0.517	0.483	0.499
	GMoD(Xiang et al., 2024)	2024	MICCAI	0.398	0.251	0.172	0.124	0.166	0.286	-	-	-
	SRRG(Hou et al., 2025)	2025	ICCV	0.416	0.260	0.178	0.128	0.162	**0.294**	0.473	0.499	0.486
Longitudinal data	Prefilling(Zhu et al., 2023)	2023	MICCAI	0.343	0.210	0.140	0.099	0.137	0.271	0.538	0.434	0.480
	HERGen(Wang et al., 2025)	2025	ECCV	0.395	0.248	0.169	0.122	0.156	0.285	0.415	0.301	0.317
	HC-LLM(Liu et al., 2025)	2025	AAAI	0.404	0.260	0.178	0.128	0.160	0.287	0.417	0.357	0.357
	Diff-RRG(Yun et al., 2026)	2026	MICCAI	0.405	0.251	0.169	0.120	0.164	0.276	0.528	0.430	0.474
	STREAM(Yang et al., 2025)	2025	TMI	0.420	0.267	0.184	0.133	0.164	0.291	0.531	0.407	0.460
	BiOTPrompt(Liu et al., 2026)	2026	CVPR	0.397	0.253	0.174	0.126	0.155	0.285	0.471	0.424	0.417
	TIM-I(Dong et al., 2026)	2026	CVPR	0.421	0.259	0.171	0.118	0.172	0.275	0.537	0.466	0.483
	TIM-II(Dong et al., 2026)	2026	CVPR	**0.430**	0.265	0.179	0.124	**0.185**	0.287	**0.563**	0.505	**0.511**
	MARE(Gao et al., 2026b)	2026	AAAI	0.409	0.265	0.184	0.133	0.161	0.291	0.433	0.378	0.375
	Ours	2026	–	0.427	**0.271**	**0.186**	**0.134**	0.171	0.290	0.561	0.468	0.510

Table 2: Ablation of the auxiliary objectives on the Longitudinal-MIMIC dataset. All variants include $\mathcal{L}_{\mathrm{CE}}$. The best results are **bolded**.

Model	$\mathcal{L}_{\mathrm{stat}}$	$\mathcal{L}_{\mathrm{dir}}$	$\mathcal{L}_{\mathrm{IOT}}$	NLG Metrics						CE Metrics		
				B-1	B-2	B-3	B-4	MTR	R-L	P	R	F1
Baseline	×	×	×	0.409	0.257	0.174	0.124	0.160	0.278	0.526	0.426	0.471
w/ $\mathcal{L}_{\mathrm{stat}}$	✓	×	×	0.418	0.261	0.177	0.128	0.164	0.281	0.546	0.437	0.485
w/ $\mathcal{L}_{\mathrm{dir}}$	×	✓	×	0.421	0.264	0.180	0.130	0.165	0.283	0.539	0.424	0.475
w/ $\mathcal{L}_{\mathrm{IOT}}$	×	×	✓	0.417	0.262	0.179	0.129	0.164	0.284	0.549	0.446	0.492
w/ $\mathcal{L}_{\mathrm{stat}} + \mathcal{L}_{\mathrm{dir}}$	✓	✓	×	0.425	0.268	0.183	0.132	0.166	0.286	0.550	0.431	0.484
Ours	✓	✓	✓	**0.427**	**0.271**	**0.186**	**0.134**	**0.171**	**0.290**	**0.561**	**0.468**	**0.510**

memory-based, prompt-guided, and explicit disease-evolution modeling approaches, providing a broad comparison of report-generation performance. As shown in Table 1, our method achieves the highest BLEU-2, BLEU-3, and BLEU-4 scores, together with the second-highest BLEU-1, clinical precision, and clinical F1 scores overall. Its clinical F1 of 0.510 is close to the 0.511 achieved by TIM-II, which includes an additional iterative report-refinement stage. These results demonstrate strong performance in both textual overlap and clinical agreement, supporting the joint use of longitudinal context and explicit semantic supervision for report generation.

Comparative Analysis. As shown in Table 1, STREAM integrates regional semantic retrieval with temporal modeling (Yang et al., 2025) and achieves strong higher-order BLEU and ROUGE-L scores, whereas TIM combines pathology–progression decoupling with report refinement (Dong et al., 2026) and leads in clinical precision and F1. These differing strengths show that textual overlap and clinical agreement capture distinct aspects of report quality. Together, they suggest the value of integrating explicit clinical semantics with temporal modeling. Our controlled ablations in Table 2 provide more direct evidence for this complementarity: IOT improves clinical performance both as the sole auxiliary objective and alongside direction and static alignment.

3.3 Ablation Study

In Table 2, we evaluate the contributions of the three auxiliary objectives in Eq. (18) within the same longitudinal report-generation framework. For all variants, the inputs, generation pathway, and $\mathcal{L}_{\mathrm{CE}}$ remain unchanged; the baseline is trained without auxiliary supervision. All variants in Figs. 3 and 4 likewise retain $\mathcal{L}_{\mathrm{CE}}$. Relative to this baseline, the model with $\mathcal{L}_{\mathrm{stat}}$ improves across all reported NLG and clinical metrics, highlighting the value of image–report alignment within each examination. With $\mathcal{L}_{\mathrm{dir}}$, the model also achieves gains in all NLG metrics, clinical precision, and F1 through supervision of the paired progression directions. Under joint static and direction

supervision, all NLG scores exceed those of either objective alone, while all three clinical metrics improve over the variant with direction supervision alone.

Beyond cross-modal alignment, $\mathcal{L}_{\mathrm{IOT}}$ provides explicit pathology guidance through label-supervised image-to-prototype assignments. With this objective as the sole auxiliary loss, clinical F1 increases from 0.471 to 0.492. A further improvement from 0.484 to 0.510 is observed when IOT complements both alignment objectives. Across the reported metrics, the full model performs best, supporting the complementarity of examination-level alignment, progression-direction alignment, and pathology semantic supervision.

Effect of Longitudinal Cross-Modal Progression Alignment.

To examine progression-direction alignment, we measure the distance between paired normalized visual and textual progression residuals on training batches: $D_{\mathrm{pos}} = \frac{1}{N}\sum_{i=1}^{N} \|\phi(\boldsymbol{d}_i^{\mathrm{v}}) - \phi(\boldsymbol{d}_i^{\mathrm{t}})\|_2^2$, where N is the number of longitudinal examination pairs in the batch and ϕ is the stabilized normalization defined in Section 2.2. A lower D_{pos} indicates closer agreement between the paired normalized progression residuals, consistent with the direction-level objective.

As shown in Fig. 3, the distance remains high and nearly constant without auxiliary supervision. With $\mathcal{L}_{\mathrm{dir}}$, it declines early in training before settling at a lower level. In combination with static alignment, the curve shifts further downward, while the full model with IOT maintains the lowest distance.

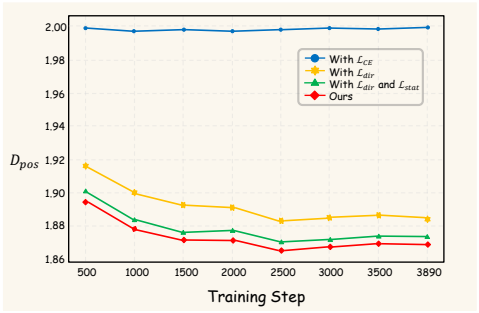


Figure 3: Distance between paired normalized visual and textual progression residuals during training.

Effect of Data-driven Pathology Semantic Enhancement.

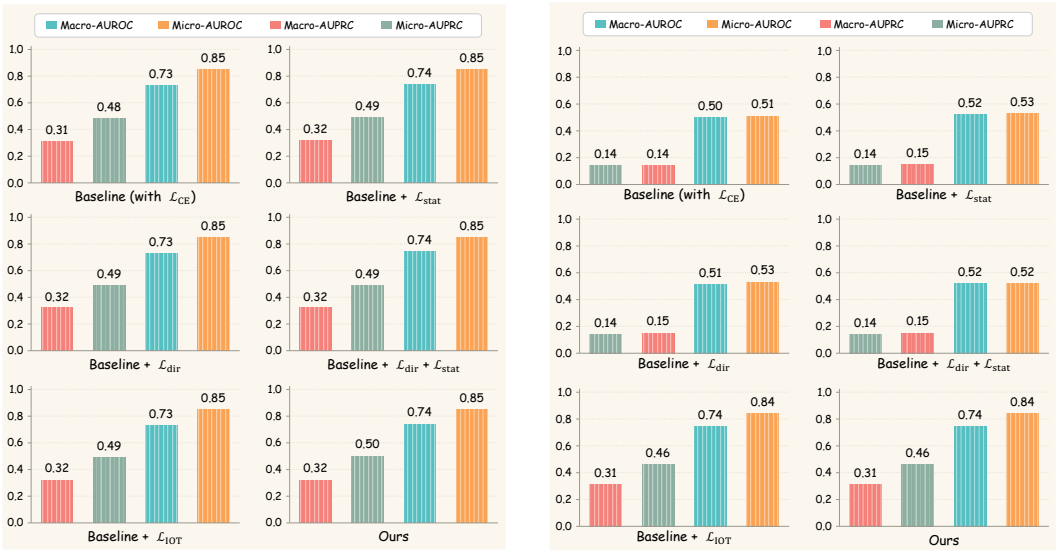
To examine whether IOT enables the prototypes to learn clinically meaningful pathology semantics, we compare linear probing of visual representations with direct prototype-based recognition. Under the linear-probing protocol (Chen et al., 2020a; Grill et al., 2020; Caron et al., 2021), the similar performance across variants in Fig. 4(a) suggests that pathology information is already recoverable from the baseline representations with a trained classifier. In direct prototype-based recognition, the variants with IOT supervision perform substantially better than those without it (Fig. 4(b)). These results support the role of IOT in enabling the prototypes to capture semantics associated with their assigned clinical observation categories.

Through joint optimization of the prototypes and the shared visual adapter, IOT also provides pathology-level supervision to the visual representations used for report generation. Its contribution to generation therefore operates through the adapted visual features. The clinical gains from IOT in Table 2 support the benefit of this supervision. With IOT as the sole auxiliary objective, prototype-based recognition is already comparable to that of the full model, yet the full model generates better reports. This suggests complementary roles for pathology semantic supervision and cross-modal alignment: the former establishes explicit associations with clinical categories, while the latter promotes consistency within examinations and in progression directions across modalities.

Case Study.

To illustrate the model's use of longitudinal information, Fig. 5 presents a report-generation example together with attention visualizations for the prior and current examinations. The highlighted regions include the lung bases, cardiac silhouette, and areas containing medical devices, offering a qualitative view of the spatial attention associated with this example.

The generated reports illustrate differences in the description of clinically relevant findings. R2GenGPT identifies the pacemaker but incorrectly states that no pleural effusion is present, contradicting the reference report, and omits the associated atelectasis. Our model describes both bilateral pleural effusions and basal atelectasis, while its statement of no relevant interval change agrees with the stability reported in the reference. Although the pacemaker is omitted, the report captures the major abnormalities and their stability relative to the prior examination, illustrating how current visual evidence and longitudinal context can jointly support clinically consistent descriptions.



(a) Linear probing (b) Prototype-based recognition

Figure 4: Pathology recognition under two complementary evaluation protocols: (a) linear probing of visual representations and (b) direct prediction from image-to-prototype associations, without an additional classifier.

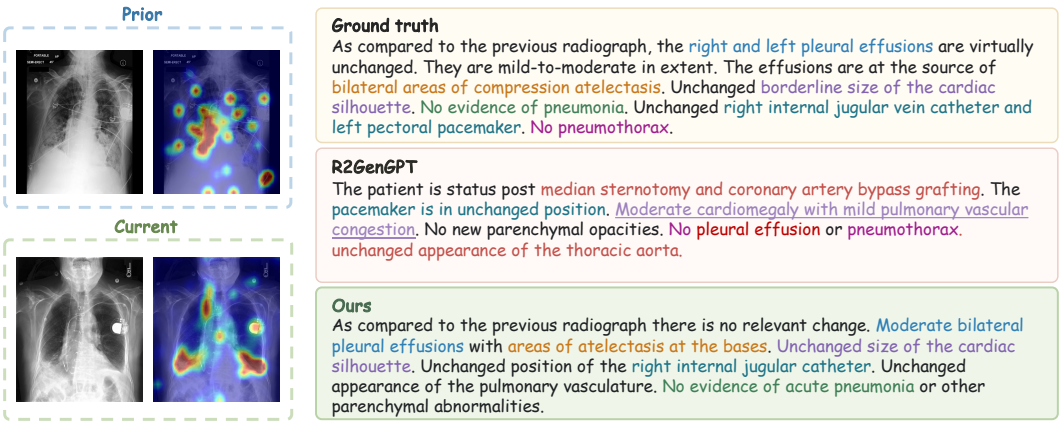


Figure 5: Case study. The left side shows the prior and current images with their corresponding attention maps. The right side compares the reference report with generated reports. Red highlights descriptions that contradict or are not mentioned in the reference report, while the other colors highlight distinct clinical findings.

4 Conclusion

In this paper, we propose a unified representation-learning framework (PAPE) for longitudinal radiology report generation that combines cross-modal progression alignment with explicit pathology semantic supervision. Temporal direction alignment encourages visual and textual representations to capture consistent directions of clinical change, while static alignment preserves semantic agreement within each examination. Besides, a label-induced inverse optimal transport objective further associates visual representations with clinical observation prototypes. These complementary objectives support report generation from current visual evidence and longitudinal context. Experiments on Longitudinal-MIMIC demonstrate improvements in report quality and clinical accuracy, while ablation studies and representation analyses support the contributions of progression alignment and pathology semantic enhancement. The current framework requires prior images and a corresponding report, and future work will extend it to accommodate both longitudinal and single-examination inputs, including cases in which historical information is unavailable.

AI USE STATEMENT

We used generative AI tools for language polishing of the manuscript, including improving grammar, clarity, and readability. This use was limited to editing the presentation of the paper. The authors take full responsibility for the final content of this work, including all text, claims, and conclusions.

REFERENCES

Francisco Andrade, Gabriel Peyré, and Clarice Poon. Sparsistency for inverse optimal transport. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 16575–16606, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/4761fab863f0900d90cf601fce6d5155-Paper-Conference.pdf.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pp. 1–21. Springer, 2022.

Shenshen Bu, Taiji Li, Yuedong Yang, and Zhiming Dai. Instance-level expert knowledge and aggregate discriminative attention for radiology report generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14194–14204. IEEE, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020a.

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, 2020b.

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.

Francesco Dalla Serra, William Clackett, Hamish MacKinnon, Chaoyang Wang, Fani Deligianni, Jeff Dalton, and Alison Q O'Neil. Multimodal generation of radiology reports using knowledge-grounded extraction of entities and relations. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 615–624, 2022.

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.

Jean-Benoit Delbrouck, Pierre Chambon, Zhihong Chen, Maya Varma, Andrew Johnston, Louis Blankemeier, Dave Van Veen, Tan Bui, Steven Truong, and Curtis Langlotz. Radgraph-xl: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12902–12915, 2024.

Yiheng Dong, Yi Lin, Shilong Huang, Xiyan Yang, and Xin Yang. Tim: Temporal decoupling with iterative mutual-refinement model for longitudinal radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6951–6961, 2026.

Yingqi Fan, Junlong Tong, Anhao Zhao, and Xiaoyu Shen. What do visual tokens really encode? uncovering sparsity and redundancy in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11987–11997, June 2026.

Jiechao Gao, Chang Liu, and Yuangang Li. S2d-align: Shallow-to-deep auxiliary learning for anatomically-grounded radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30780–30788, 2026a.

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

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *Advances in neural information processing systems*, volume 33, pp. 21271–21284, 2020.

Xuege Hou, Yali Li, and Shengjin Wang. Knowledge-driven query network with adaptive cross-view attention for structured radiology report generation. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1245–1254. IEEE, 2025.

Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. Kiut: Knowledge-injected u-transformer for radiology report generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19809–19818. IEEE, 2023.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021.

Jiayi Ji, Yunpeng Luo, Xiaoshuai Sun, Fuhai Chen, Gen Luo, Yongjian Wu, Yue Gao, and Rongrong Ji. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 1655–1663, 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.

Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 6666–6673, 2019a.

Ruilin Li, Xiaojing Ye, Haomin Zhou, and Hongyuan Zha. Learning to match via inverse optimal transport. *Journal of machine learning research*, 20(80):1–37, 2019b.

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

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

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.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(6):5595–5603, 2025. doi: 10.1609/aaai.v39i6.32596. URL https://doi.org/10.1609/aaai.v39i6.32596.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13755–13765, 2026.

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.

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

Richard Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

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.

Andrew M Stuart and Marie-Therese Wolfram. Inverse optimal transport. *SIAM Journal on Applied Mathematics*, 80(1):599–619, 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.

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.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 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, 2025.

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

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3876–3887, 2022.

ZhiPeng Xiang, ShaoGuo Cui, CaoZhi Shang, Jingfeng Jiang, and Liqiang Zhang. Gmod: graph-driven momentum distillation framework with active perception of disease severity for radiology report generation. In *International conference on medical image computing and computer-assisted intervention*, pp. 295–305. Springer, 2024.

Yuan Xue, Tao Xu, L Rodney Long, Zhiyun Xue, Sameer Antani, George R Thoma, and Xiaolei Huang. Multimodal recurrent model with attention for automated radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 457–466. Springer, 2018.

Yan Yang, Xiaoxing You, Ke Zhang, Zhenqi Fu, Xianyun Wang, Jiajun Ding, Jiamei Sun, Zhou Yu, Qingming Huang, Weidong Han, et al. Spatio-temporal and retrieval-augmented modeling for chest x-ray report generation. *IEEE Transactions on Medical Imaging*, 44(7):2892–2905, 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 James C. Gee, Daniel C. Alexander, Jaesung Hong, Juan Eugenio Iglesias, Carole H. Sudre, Archana Venkataraman, Polina Golland, Jong Hyo Kim, and Jinah Park (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15966 of *Lecture Notes in Computer Science*, pp. 152–161, Cham, 2026. Springer Nature Switzerland. doi: 10.1007/978-3-032-04981-0_15. URL https://doi.org/10.1007/978-3-032-04981-0_15.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine learning for healthcare conference*, pp. 2–25. PMLR, 2022.

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.

Xingyu Zhu, Kesen Zhao, Liang Yi, Shuo Wang, Zhicai Wang, Beier Zhu, Hanwang Zhang, and Xiangnan He. Look carefully: Adaptive visual reinforcements in multimodal large language models for hallucination mitigation. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 81235–81256, 2026.

A Appendix

A.1 Derivation of Temporal Inverse Optimal Transport

We derive the closed-form transport plan and the label-induced objective in 16, and then characterize their gradients. For each examination $r \in \{\mathrm{h}, \mathrm{c}\}$, let $\mathcal{A}_r = \{i : |\mathcal{G}_{r,i}| > 0\}$ denote the examples with at least one positive label. The visual adapter and prototypes, collectively parameterized by ϑ, determine the cost matrix $\boldsymbol{C}_r(\vartheta)$ through Eq. 11. The inverse problem learns these parameters by matching the optimal transport plan to the empirical label assignments. The same adapter and prototypes are shared across prior and current examinations.

Closed-form solution of the inner problem. Because Eq. 12 constrains only the row masses, its optimization separates across examples. For $i \notin \mathcal{A}_r$, nonnegativity and zero row mass force $[\boldsymbol{T}_r^*]_{i,j} = 0$ for every j. For the remaining rows, introduce multipliers $\lambda_{r,i}$ and write

$$\begin{aligned}\mathcal{J}_r(\boldsymbol{T}, \lambda_r) = &\sum_{i \in \mathcal{A}_r} \sum_{j=1}^{K} ([\boldsymbol{C}_r]_{i,j}[\boldsymbol{T}]_{i,j} + \epsilon[\boldsymbol{T}]_{i,j} \log[\boldsymbol{T}]_{i,j}) \\ &- \sum_{i \in \mathcal{A}_r} \lambda_{r,i} \left(\sum_{j=1}^{K} [\boldsymbol{T}]_{i,j} - \frac{1}{N} \right).\end{aligned} \tag{19}$$

For finite costs and $\epsilon > 0$, the entropy-regularized objective has a unique minimizer with strictly positive entries on each active row. The stationarity condition is therefore

$$\begin{aligned}\frac{\partial \mathcal{J}_r}{\partial [\boldsymbol{T}]_{i,j}} &= [\boldsymbol{C}_r]_{i,j} + \epsilon\big(\log[\boldsymbol{T}]_{i,j} + 1\big) - \lambda_{r,i} = 0, \\ [\boldsymbol{T}_r^*]_{i,j} Eq. &= \exp\left(\frac{\lambda_{r,i}}{\epsilon} - 1\right) \exp\left(-\frac{[\boldsymbol{C}_r]_{i,j}}{\epsilon}\right).\end{aligned} \tag{20}$$

Define $Z_{r,i} = \sum_{k=1}^{K} \exp(-[\boldsymbol{C}_r]_{i,k}/\epsilon)$. Enforcing the row constraint gives $\exp(\lambda_{r,i}/\epsilon - 1) = 1/(NZ_{r,i})$. Combining the active and zero-mass rows yields

$$[\boldsymbol{T}_r^*]_{i,j} = \frac{a_{r,i}}{N} \frac{\exp(-[\boldsymbol{C}_r]_{i,j}/\epsilon)}{Z_{r,i}} = \frac{a_{r,i}}{N} q_{r,i,j}, \tag{21}$$

which recovers Eq. (13). No iterative transport solver is required under these row constraints.

Reduction of the inverse problem. As defined in Eq. (14), positive categories receive uniform target weights $t_{r,i,j}$, while all other categories receive zero target mass. Thus $\sum_j t_{r,i,j} = a_{r,i}$ and $[\boldsymbol{T}_r^{\mathrm{e}}]_{i,j} = t_{r,i,j}/N$. The empirical and optimal plans have the same total mass $|\mathcal{A}_r|/N$, which need not equal one. Consequently, the linear terms in the generalized KL divergence cancel, leaving Eq. (15). Evaluating that expression on the support of the empirical plan gives

$$\begin{aligned}D_{\mathrm{KL}}(\boldsymbol{T}_r^{\mathrm{e}} \,\|\, \boldsymbol{T}_r^*) &= \frac{1}{N} \sum_{i \in \mathcal{A}_r} \sum_{j \in \mathcal{G}_{r,i}} t_{r,i,j} \log \frac{t_{r,i,j}}{q_{r,i,j}} \\ &= -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} t_{r,i,j} \log q_{r,i,j} + \frac{1}{N} \sum_{i \in \mathcal{A}_r} \sum_{j \in \mathcal{G}_{r,i}} t_{r,i,j} \log t_{r,i,j} \\ &= \mathcal{L}_{\mathrm{IOT}}^r - \frac{1}{N} \sum_{i \in \mathcal{A}_r} \log |\mathcal{G}_{r,i}|.\end{aligned} \tag{22}$$

The second term in the last line depends only on the supplied labels. Minimizing the inverse transport objective is therefore equivalent to minimizing $\mathcal{L}_{\mathrm{IOT}}^r$ with respect to ϑ.

Supervision across examinations. Substituting the cosine cost and averaging the two examination-specific objectives produces the explicit temporal IOT loss

$$\mathcal{L}_{\mathrm{IOT}} = -\frac{1}{2N} \sum_{r \in \{\mathrm{h},\mathrm{c}\}} \sum_{i \in \mathcal{A}_r} \frac{1}{|\mathcal{G}_{r,i}|} \sum_{j \in \mathcal{G}_{r,i}} \log \frac{\exp\big(\phi(\boldsymbol{z}_{r,i}^{\mathrm{v}})^\top \phi(\boldsymbol{p}_j)/\epsilon\big)}{\sum_{k=1}^{K} \exp\big(\phi(\boldsymbol{z}_{r,i}^{\mathrm{v}})^\top \phi(\boldsymbol{p}_k)/\epsilon\big)}. \tag{23}$$

The normalization includes all K observation categories. Examples without positive labels contribute zero, and the loss remains normalized by the full batch size N. Sharing the prototypes applies a common set of clinical semantic targets to both examinations.

Gradients for representation and prototype learning. Let $s_{r,i,j} = -[\boldsymbol{C}_r]_{i,j}/\epsilon$, so that $q_{r,i,j} = \exp(s_{r,i,j})/\sum_k \exp(s_{r,i,k})$. Using $\partial \log q_{r,i,k}/\partial s_{r,i,j} = \mathbf{1}[k=j] - q_{r,i,j}$ gives

$$\begin{aligned} \frac{\partial \mathcal{L}_{\mathrm{IOT}}^{r}}{\partial s_{r,i,j}} &= -\frac{1}{N}\sum_{k=1}^{K} t_{r,i,k}\big(\mathbf{1}[k=j] - q_{r,i,j}\big) = \frac{a_{r,i}q_{r,i,j} - t_{r,i,j}}{N}, \\ \frac{\partial \mathcal{L}_{\mathrm{IOT}}}{\partial [\boldsymbol{C}_r]_{i,j}} &= \frac{t_{r,i,j} - a_{r,i}q_{r,i,j}}{2N\epsilon}. \end{aligned} \tag{24}$$

Accordingly, the gradient with respect to the shared learnable parameters is

$$\nabla_\vartheta \mathcal{L}_{\mathrm{IOT}} = \frac{1}{2N\epsilon}\sum_{r\in\{\mathrm{h},\mathrm{c}\}}\sum_{i=1}^{N}\sum_{j=1}^{K}\big(a_{r,i}q_{r,i,j} - t_{r,i,j}\big)\nabla_\vartheta\left[\phi(\boldsymbol{z}_{r,i}^{\mathrm{v}})^\top \phi(\boldsymbol{p}_j)\right]. \tag{25}$$

For each active example, the assignment error $q_{r,i,j} - t_{r,i,j}$ therefore jointly updates its adapted visual representation and the prototypes through their normalized similarity. Inactive examples have $a_{r,i} = t_{r,i,j} = 0$ and contribute no IOT gradient. This supervision shapes the visual features subsequently used for report generation, while the prototypes serve as learned semantic targets during training.

A.2 Detailed Implementation

Feature adaptation. Visual and textual tokens are refined by separate residual MLP adapters, each comprising layer normalization, a $768 \to 192$ projection, GELU, and a $192 \to 768$ projection. Adapter parameters are shared across examinations within each modality. For multiple images from the same examination, adapted features are averaged at corresponding token positions. The first token supports the auxiliary objectives, while the full sequence is used for report generation.

Semantic state compression. Two Perceivers with independent parameters summarize current and prior information using separate sets of 128 learnable, 768-dimensional queries. Writing $\mathcal{P}_r(\boldsymbol{Q}, \boldsymbol{F})$ for a Perceiver with queries $\boldsymbol{Q}$ and input features $\boldsymbol{F}$, the compression proceeds as

$$\begin{aligned} \boldsymbol{S}_{\mathrm{c}} &= \mathcal{P}_{\mathrm{c}}(\boldsymbol{Q}_{\mathrm{c}}, \boldsymbol{F}_{\mathrm{c}}^{\mathrm{v}}), \\ \boldsymbol{U}_{\mathrm{h}} &= \mathcal{P}_{\mathrm{h}}(\boldsymbol{Q}_{\mathrm{h}}, \boldsymbol{F}_{\mathrm{h}}^{\mathrm{t}}), \qquad \boldsymbol{S}_{\mathrm{h}} = \mathcal{P}_{\mathrm{h}}(\boldsymbol{U}_{\mathrm{h}}, \boldsymbol{F}_{\mathrm{h}}^{\mathrm{v}}). \end{aligned} \tag{26}$$

The historical branch first summarizes the prior report, then uses this summary to query the prior image features. These two steps share the same Perceiver. Each Perceiver contains three blocks of cross-attention, latent self-attention, and feed-forward processing, with pre-normalization and residual connections. Both attention types use eight heads with 64 dimensions per head; parameters are independent across blocks.

Temporal fusion and decoding. Three temporal cross-attention blocks update the current state using the historical state as keys and values, yielding $\boldsymbol{S}_{\mathrm{d}}$. Each block uses eight heads with 192 dimensions per head, followed by a GELU feed-forward sublayer of hidden width 768, with pre-normalization and residual connections. A final layer normalization produces the output. The current and history-conditioned states are then combined:

$$\boldsymbol{S}_{\mathrm{v}} = \mathrm{LN}(\mathrm{Linear}([\boldsymbol{S}_{\mathrm{c}}; \boldsymbol{S}_{\mathrm{d}}])). \tag{27}$$

Concatenation along the token dimension gives 256 conditioning tokens, and the linear layer maps their dimension from 768 to 4096 for LLaMA2-7B. Together with the fixed prompt, this sequence conditions report generation. Perceiver and temporal-fusion dropout are zero.

The encoders and language model remain frozen, while the adapters, learnable queries, compression and fusion modules, and projection layers are trained jointly. The current report provides training supervision only; generation uses current images and the prior image–report pair. Pathology prototypes supervise the visual adapter through IOT without being included in the decoder's conditioning sequence.

A.3 Related Works

Radiology Report Generation. Radiology report generation (RRG) aims to automatically generate clinically meaningful diagnostic reports from medical images (Ji et al., 2021; Vinyals et al., 2015), thereby reducing the documentation burden on radiologists and improving reporting efficiency. Compared with generic image captioning, RRG requires the generation of fine-grained and clinically structured descriptions, placing greater demands on abnormality localization, clinical semantic understanding, and long-form language generation. Early RRG methods generally adopted encoder–decoder architectures, often augmented with domain knowledge or multimodal alignment mechanisms (Liu et al., 2021; Huang et al., 2023; Xue et al., 2018; Li et al., 2019a). With the rapid development of large language models (LLMs), recent studies have increasingly focused on projecting visual representations into the language space of pretrained LLMs. For example, Bootstrapping LLM (Liu et al., 2024) progressively adapts a general-purpose LLM to radiology-specific vocabulary, report structure, and generation patterns through a staged contrastive decoding strategy. S2D-Align (Gao et al., 2026a) further introduces progressive anatomical alignment by gradually incorporating supervision from image–report pairs, reference reports, and key phrases, while employing a shared memory adapter to facilitate feature transfer across training stages. Despite these advances, most LLM-based methods remain primarily designed for single-study report generation and have therefore received limited attention to the temporal information available in longitudinal examinations.

Recently, several studies have applied LLMs to longitudinal settings. BiOTPrompt (Liu et al., 2026) employs bidirectional optimal transport to identify evolving regions and construct corresponding prompts, whereas MARE (Gao et al., 2026b) uses analogy-based reasoning to align lesion-level features and regularize the evolution of visual and textual representations. These developments highlight the importance of modeling clinical changes across examinations. Building on this direction, we investigate how examination-level visual and textual representations can be jointly organized to express both consistent temporal changes and explicit pathology semantics.

Inverse Optimal Transport. Optimal Transport (OT) theory (Andrade et al., 2024) aims to find a transport plan that minimizes the overall cost of moving mas from one distribution to another. Specifically, OT problem can be formally expressed as the Wasserstein-1 distance:

$$\mathcal{D}_w(\mu,\nu) = \min_{\mathbf{T}\in\Pi(\mu,\nu)} \langle \mathbf{C}, \mathbf{T}\rangle = \sum_{i=1}^{a}\sum_{j=1}^{b} \mathbf{C}_{ij}\mathbf{T}_{ij}, \tag{28}$$

where $\Pi(\mu,\nu)$ represents the set of all possible couplings that convert μ to ν, $\mathbf{T}$ denotes the transport plan and $\mathbf{C}$ is the cost matrix. To facilitate solving the OT problem, an entropic regularization term is typically added to the objective function, making the problem strongly convex and allowing it to be optimized using iterative methods such as the Sinkhorn algorithm (Sinkhorn, 1967). The regularized problem is defined as:

$$\min_{\mathbf{T}\in\Pi(\mu,\nu)} \langle \mathbf{C}, \mathbf{T}\rangle - \epsilon H(\mathbf{T}), \tag{29}$$

where $H(\mathbf{T}) = -\sum_{i,j} T_{ij}\,(\log T_{ij} - 1)$ denotes the entropy of $\mathbf{T}$, ϵ is a regularization parameter.

Compared to the traditional OT, Inverse Optimal Transport (IOT) (Stuart & Wolfram, 2020; Li et al., 2019b) aims to infer the cost matrix based on given optimal transport plan. This can be formulated as a bilevel optimization problem:

$$\min_{\mathbf{C}} \mathcal{L}(\mathbf{T}, \mathbf{T}^*), \tag{30}$$

where $\mathbf{T}^*$ is the given transport plan, and $\mathcal{L}$ is a loss function that measures the discrepancy between $\mathbf{T}$ induced by $\mathbf{C}$ and $\mathbf{T}^*$. In scenarios where the transport cost is difficult to observe directly or must be inferred from data, the IOT framework offers significant practical value. It not only facilitates a richer understanding of the transport process but also further broadens the applicability of optimal transport methods.