Temporal Evidence Disentanglement with Progressive Supervision for Chest X-ray Report Generation

Anonymous ACL submission

Abstract

Chest X-ray report generation aims to produce clinically meaningful reports from radiographic images. Existing methods mainly focus on single examinations, while recent longitudinal approaches typically incorporate historical information through implicit fusion or decoder-level conditioning. Such designs often entangle temporally invariant and visit-specific evidence, making it difficult to reliably model disease progression. In this paper, we propose TDPS, a two-stage framework for chest X-ray report generation. TDPS first learns temporally disentangled image and report representations with dedicated disentanglement and alignment objectives, decomposing longitudinal evidence into invariant, current-specific, and prior-specific features. It then employs progressive supervised fine-tuning with observation labels and clinical entity-relation triples as intermediate training objectives, enabling clinically grounded report generation without additional inference-time steps. Extensive experiments on MIMIC-CXR and MIMIC-ABN show that TDPS outperforms state-of-the-art methods on most metrics while remaining competitive on the others. The code is available at https://anonymous.4open.science/r/TDPS.

1 Introduction

Radiological imaging, such as chest X-rays, plays an indispensable role in clinical diagnosis and treatment (Lambin et al., 2017). Chest X-ray report generation aims to automatically produce clinically meaningful reports from radiographic images, reducing radiologists' workload and improving the efficiency of clinical documentation (Chen et al., 2020). Existing radiology report generation (RRG) methods have made substantial progress by enhancing visual-language alignment, incorporating medical knowledge, and leveraging large language models (Wang et al., 2023; Huang et al., 2023; Li et al., 2023; Zhang et al., 2024; Hou et al., 2025b; Yin et al., 2025). However, most of them generate reports from a single examination, while real-world radiological reporting often requires comparison with prior studies to describe whether findings are stable, newly developed, worsened, improved, or resolved. Such longitudinal reasoning is essential for characterizing disease progression and supporting clinical decision-making.

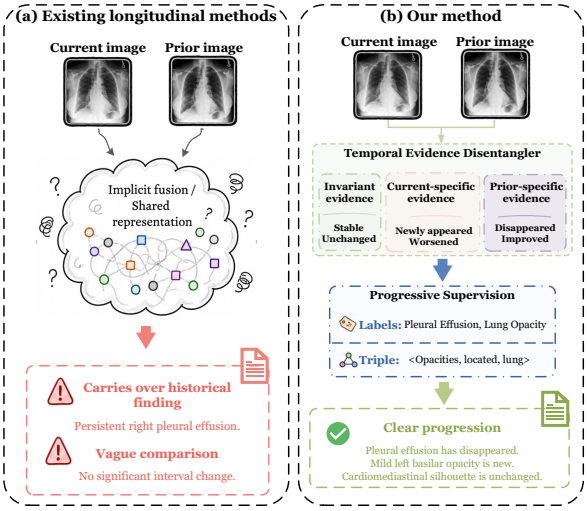


Figure 1: Motivation of TDPS. Existing longitudinal methods implicitly fuse current and prior images, which may carry over historical findings or produce vague comparisons. TDPS disentangles invariant, current-specific, and prior-specific evidence and uses progressive supervision for clear progression descriptions.

Recent studies have explored longitudinal RRG by incorporating prior images, prior reports, indications, or medical history (Karwande et al., 2022; Zhu et al., 2023; Nguyen et al., 2023; Liu et al., 2024b; Mei et al., 2025; Liu et al., 2026). These methods demonstrate the value of historical information, but they usually integrate current and prior evidence through implicit fusion, global alignment, or decoder-level conditioning. As a result, temporally invariant evidence and visit-specific evidence are often entangled in a shared representation, making it difficult to explicitly distinguish stable find-

ings from clinically meaningful changes.

We argue that the key challenge in longitudinal report generation is not merely how to incorporate prior information, but how to distinguish the temporal status of clinical evidence for clinically grounded report generation. Evidence shared across current and prior examinations can support stable findings, whereas evidence specific to either visit may indicate newly appeared, worsened, improved, or resolved abnormalities. Without such temporal disentanglement, a model may incorrectly carry over historical findings or generate vague progression descriptions.

Moreover, temporal evidence disentanglement alone does not ensure clinically accurate report generation. The model must further ground temporal evidence into clinically salient observations and entity-level relations, such as disease presence, uncertainty, and disease-anatomy associations. Direct supervision from free-text reports is often insufficient for this purpose, since reports are heterogeneous and contain expressions not directly tied to abnormal findings. Therefore, effective longitudinal RRG should not only disentangle temporal evidence, but also learn to ground such evidence with structured clinical supervision during training.

To address these challenges, we propose **TDPS** (**T**emporal **D**isentanglement with **P**rogressive **S**upervision), a two-stage framework for chest X-ray report generation. In the first stage, TDPS learns temporally disentangled image and report representations with dedicated disentanglement and alignment objectives, decomposing longitudinal evidence into temporally invariant, current-specific, and prior-specific features. In the second stage, TDPS employs progressive supervised fine-tuning with observation labels and clinical entity-relation triples as intermediate training objectives, enabling the model to learn clinically grounded temporal representations while generating only the final report at inference. Our contributions are summarized as follows:

- We propose TDPS, a two-stage framework for chest X-ray report generation, which models the temporal status of clinical evidence and learns clinically grounded representations using training-time structured supervision.
- We design temporal feature disentanglement modules and targeted training objectives for TDPS, including contrastive and reconstruction constraints to separate longitudinal representations, and progressive supervised fine-tuning with observation labels and clinical entity-relation triples for clinical grounding.
- Experiments on MIMIC-CXR and MIMIC-ABN demonstrate that TDPS achieves strong overall performance, with ablation studies validating the benefits of longitudinal information, temporal disentanglement, and progressive clinical supervision.

2 Related Work

2.1 Radiology Report Generation

Radiology report generation (RRG) has evolved from early CNN-RNN architectures (**?**Wang et al., 2018, 2021) to Transformer-based and large language model-based frameworks (Wang et al., 2023; Liu et al., 2024a; Chen et al., 2024; Hou et al., 2025a). Recent studies improve report quality through visual-language alignment (Chen et al., 2021; Shen et al., 2024), retrieval augmentation (Liu et al., 2024c; Jeong et al., 2024; Huang et al., 2025b), and medical knowledge integration (Huang et al., 2023; Zhang et al., 2024; Hou et al., 2025b; Yin et al., 2025). Despite these advances, most methods generate reports from a single examination and thus cannot explicitly model disease progression across visits.

To better reflect real-world radiological practice, longitudinal RRG incorporates prior images, prior reports, indications, or medical history into report generation (Karwande et al., 2022; Zhu et al., 2023; Nguyen et al., 2023; Liu et al., 2024b; Mei et al., 2025; Liu et al., 2025, 2026). These methods demonstrate the value of historical information, but usually incorporate current and prior evidence through implicit fusion, alignment, or decoder-level conditioning. In contrast, our work explicitly decomposes longitudinal features into temporally invariant, current-specific, and prior-specific features, enabling more direct modeling of the temporal status of clinical evidence.

2.2 Structured Clinical Knowledge for Report Generation

Structured clinical knowledge has been widely used to improve the clinical accuracy of radiology report generation. Prior studies introduce disease labels, medical entities, knowledge graphs to enhance abnormality modeling, or optimize factual correctness (Zhang et al., 2020; Li et al., 2023; Bu

et al., 2024; Huang et al., 2025a). This information provides compact information about disease presence and clinical relations. However, existing methods mainly use them for single-examination report generation or general knowledge enhancement. Our method instead uses observation labels and clinical entity-relation triples as training-time intermediate objectives to ground temporally disentangled representations, improving longitudinal report generation without adding extra inference-time generation steps.

3 Method

3.1 Problem Formulation

Given a current chest X-ray image x^c, a prior image x^p, a prior report r^p, and clinical context c, longitudinal chest X-ray report generation aims to generate the current report r^c. In this work, the decoder generates the current report conditioned on the current image, prior image, and clinical context:

$$p(r^c \mid x^c, x^p, c). \tag{1}$$

During training, the prior report r^p is additionally used to provide longitudinal textual supervision for temporal representation learning, while the current report r^c serves as the generation target. We also derive observation labels Y^{label} and clinical triples Y^{triple} from r^c using CheXpert (Irvin et al., 2019) and RadGraph (Jain et al., 2021), respectively, as training-time intermediate supervision. At inference, TDPS uses x^c, x^p, and c as inputs and directly generates r^c without using the current report, labels, or triples.

3.2 Overview of TDPS

As shown in Figure 2, TDPS is a two-stage framework for longitudinal chest X-ray report generation. Stage 1 learns temporal feature disentanglement from current-prior pairs, decomposing image and report representations into invariant, current-specific, and prior-specific features. It uses temporal disentanglement losses and global instance-level alignment to learn separable yet modality-consistent temporal representations. Stage 2 integrates the disentangled temporal visual features and clinical context into the decoder, and progressively fine-tunes the model with observation labels, clinical triples, and final reports. The structured objectives are used only during training, and TDPS directly generates the final report at inference.

3.3 Stage 1: Temporal Disentanglement and Modality Alignment Pre-training

Stage 1 aims to decompose each current-prior pair into three temporal features: invariant, current-specific, and prior-specific. The invariant feature captures information shared by the two studies, while the current-specific and prior-specific features capture information unique to the current and prior studies, respectively. These features are learned from both image and report modalities, with an additional global alignment objective to maintain image-report consistency. The current report is only used during training for representation learning and is not available at inference.

Feature Extraction. We use a vision encoder $E_v(\cdot)$ and a text encoder $E_t(\cdot)$ to extract image and report representations:

$$\mathbf{V^c} = E_v(x^c), \mathbf{V^p} = E_v(x^p), \tag{2}$$

$$\mathbf{T^c} = E_t(r^c), \mathbf{T^p} = E_t(r^p), \mathbf{T^{ctx}} = E_t(c), \tag{3}$$

where V^c and V^p denote the current and prior visual features. T^c, T^p, and T^{ctx} denote the current report, prior report, and clinical context features. These features provide paired longitudinal evidence from both image and report modalities.

Temporal Feature Disentanglement. To obtain these temporal features, TDPS applies temporal feature disentanglement to both image and report representations. The intuition is that the information shared between current and prior studies forms the invariant feature, while the residual information from each study forms the current-specific or prior-specific feature. For visual features, we estimate shared longitudinal evidence through bidirectional cross-attention:

$$\begin{aligned} \mathbf{H_v^{c\leftarrow p}} &= \mathrm{Attn}(\mathbf{V^c}, \mathbf{V^p}, \mathbf{V^p}), \\ \mathbf{H_v^{p\leftarrow c}} &= \mathrm{Attn}(\mathbf{V^p}, \mathbf{V^c}, \mathbf{V^c}), \end{aligned} \tag{4}$$

where $\mathrm{Attn}(Q, K, V)$ denotes cross-attention with query Q and key-value pairs (K, V). We then fuse the bidirectional interaction features to obtain the temporal invariant visual features V^{inv}:

$$V^{inv} = \mathrm{MLP}\big([\mathbf{H_v^{c\leftarrow p}}; \mathbf{H_v^{p\leftarrow c}}]\big), \tag{5}$$

where $[\cdot;\cdot]$ denotes concatenation, and MLP is implemented as a linear projection followed by a GELU activation function. The current-specific

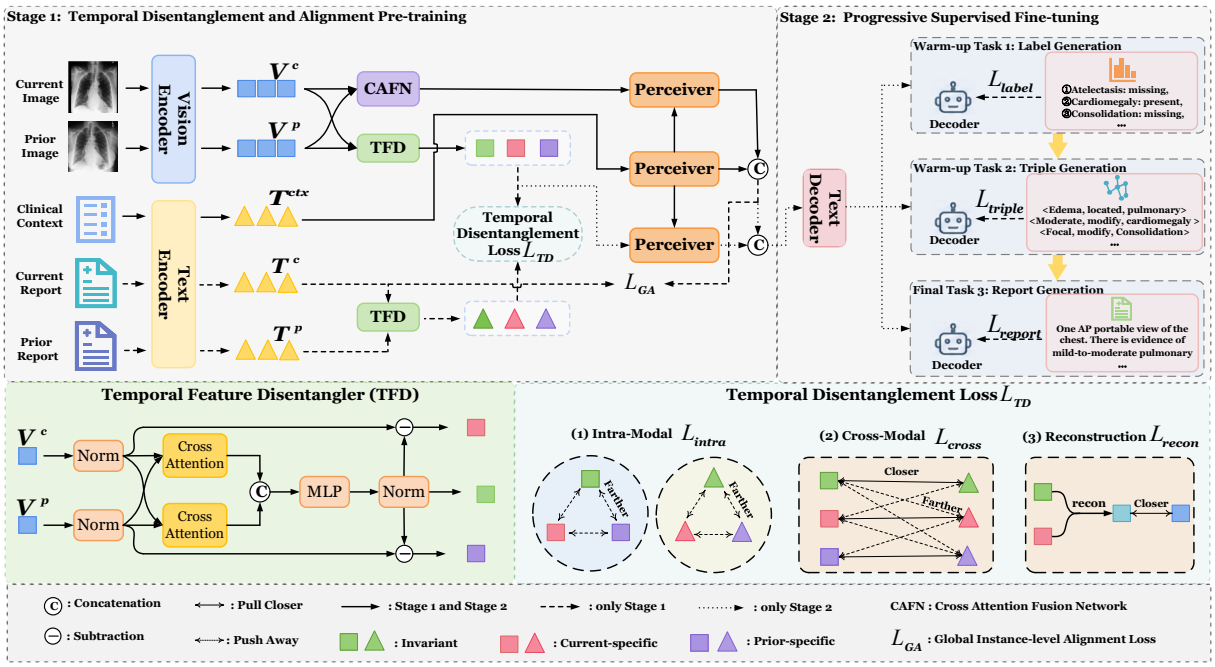


Figure 2: Overview of our proposed TDPS. Stage 1 learns invariant, current-specific, and prior-specific features with temporal disentanglement and global instance-level alignment losses. Stage 2 integrates temporal visual features and clinical context for progressive supervised fine-tuning.

and prior-specific visual features are computed as residual features:

$$\mathbf{V}^{\mathbf{cur}} = \mathbf{V}^{\mathbf{c}} - \mathbf{V}^{\mathbf{inv}}, \mathbf{V}^{\mathbf{pri}} = \mathbf{V}^{\mathbf{p}} - \mathbf{V}^{\mathbf{inv}}. \quad (6)$$

The report temporal feature disentanglement module follows the same procedure with independent parameters, producing T^{inv}, T^{cur}, and T^{pri} from T^c and T^p. The invariant features are encouraged to encode stable findings shared across studies, while the current-specific and prior-specific features capture evidence unique to each visit, supporting the modeling of newly appeared, worsened, improved, or resolved findings.

Temporal Disentanglement Loss. To ensure that the three temporal features are distinguishable, semantically aligned across modalities, and information-preserving, we introduce a temporal disentanglement loss with intra-modal disentanglement, cross-modal disentanglement, and reconstruction constraints.

1) Intra-modal disentanglement. To make the three temporal features capture distinct information, we reduce their pairwise similarity within each modality. For the visual modality, the intra-modal disentanglement loss is:

$$\mathcal{L}^{v}_{intra} = \frac{1}{N}\sum_{i=1}^{N}\big(\mathrm{sim}^2(\mathbf{v}_i^{inv}, \mathbf{v}_i^{cur}) + \mathrm{sim}^2(\mathbf{v}_i^{inv}, \mathbf{v}_i^{pri}) + \mathrm{sim}^2(\mathbf{v}_i^{cur}, \mathbf{v}_i^{pri})\big), \quad (7)$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity. The textual intra-modal loss $\mathcal{L}^{t}_{intra}$ is computed in the same way using $\mathbf{t}^{inv}$, $\mathbf{t}^{cur}$, and $\mathbf{t}^{pri}$. The overall intra-modal disentanglement loss is:

$$\mathcal{L}_{intra} = \mathcal{L}^{v}_{intra} + \mathcal{L}^{t}_{intra}. \quad (8)$$

2) Cross-modal disentanglement. While intra-modal disentanglement separates temporal features within each modality, cross-modal disentanglement encourages the disentangled image and report representations to preserve consistent temporal semantics. We first aggregate the disentangled features of each sample:

$$\begin{aligned}\mathbf{v}^{td} &= \mathrm{Pool}([\mathbf{V}^{\mathbf{inv}}; \mathbf{V}^{\mathbf{cur}}; \mathbf{V}^{\mathbf{pri}}]),\\ \mathbf{t}^{td} &= \mathrm{Pool}([\mathbf{T}^{\mathbf{inv}}; \mathbf{T}^{\mathbf{cur}}; \mathbf{T}^{\mathbf{pri}}]),\end{aligned} \quad (9)$$

where $[\cdot;\cdot]$ denotes concatenation. Given a mini-batch of N image-report pairs, the image-to-report disentanglement loss is defined as:

$$\mathcal{L}^{td}_{v2t} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\mathrm{sim}(\mathbf{v}_i^{td}, \mathbf{t}_i^{td})/\tau)}{\sum_{j=1}^{N}\exp(\mathrm{sim}(\mathbf{v}_i^{td}, \mathbf{t}_j^{td})/\tau)}, \quad (10)$$

where τ is the temperature parameter. The report-to-image disentanglement loss $\mathcal{L}^{td}_{t2v}$ is defined symmetrically. The cross-modal disentanglement loss is:

$$\mathcal{L}_{cross} = \frac{1}{2}\left(\mathcal{L}^{td}_{v2t} + \mathcal{L}^{td}_{t2v}\right). \quad (11)$$

3) Reconstruction constraint. To prevent temporal feature disentanglement from discarding useful information, we reconstruct the original current and prior features from the invariant and visit-specific features:

$$\widetilde{\mathbf{V}}^{\mathbf{c}} = \mathrm{MLP}([\mathbf{V}^{\mathbf{inv}}; \mathbf{V}^{\mathbf{cur}}]), \quad \widetilde{\mathbf{V}}^{\mathbf{p}} = \mathrm{MLP}([\mathbf{V}^{\mathbf{inv}}; \mathbf{V}^{\mathbf{pri}}]),$$
$$\widetilde{\mathbf{T}}^{\mathbf{c}} = \mathrm{MLP}([\mathbf{T}^{\mathbf{inv}}; \mathbf{T}^{\mathbf{cur}}]), \quad \widetilde{\mathbf{T}}^{\mathbf{p}} = \mathrm{MLP}([\mathbf{T}^{\mathbf{inv}}; \mathbf{T}^{\mathbf{pri}}]). \tag{12}$$

Then the reconstruction loss is defined as:

$$\mathcal{L}_{recon} = \frac{1}{N}\sum_{i=1}^{N}(\|\widetilde{\mathbf{V}}^{\mathbf{c}}_{\mathbf{i}} - \mathbf{V}^{\mathbf{c}}_{\mathbf{i}}\|_2^2 + \|\widetilde{\mathbf{V}}^{\mathbf{p}}_{\mathbf{i}} - \mathbf{V}^{\mathbf{p}}_{\mathbf{i}}\|_2^2 + \|\widetilde{\mathbf{T}}^{\mathbf{c}}_{\mathbf{i}} - \mathbf{T}^{\mathbf{c}}_{\mathbf{i}}\|_2^2 + \|\widetilde{\mathbf{T}}^{\mathbf{p}}_{\mathbf{i}} - \mathbf{T}^{\mathbf{p}}_{\mathbf{i}}\|_2^2). \tag{13}$$

Finally, the temporal disentanglement loss is:

$$\mathcal{L}_{TD} = \mathcal{L}_{intra} + \mathcal{L}_{cross} + \mathcal{L}_{recon}. \tag{14}$$

Global Instance-level Alignment Loss. We further adopt a global instance-level alignment loss to preserve the overall correspondence between longitudinal visual evidence and the current report. Specifically, we first integrate current and prior visual features through a cross-attention fusion network:

$$\begin{aligned}\bar{\mathbf{V}}^{\mathbf{l}} &= \mathbf{V}^{\mathbf{c}} + \mathrm{Attn}(\mathbf{V}^{\mathbf{c}}, \mathbf{V}^{\mathbf{p}}, \mathbf{V}^{\mathbf{p}}),\\ \mathbf{V}^{\mathbf{l}} &= \bar{\mathbf{V}}^{\mathbf{l}} + \mathrm{MLP}(\bar{\mathbf{V}}^{\mathbf{l}}),\end{aligned} \tag{15}$$

where V^l denotes the longitudinal visual feature that integrates current and prior image information.

To further incorporate clinical context, we use a Perceiver (Jaegle et al., 2021) to condense the encoded clinical context and fuse it with the longitudinal visual feature:

$$\begin{aligned}\widetilde{\mathbf{T}}^{\mathbf{ctx}} &= \mathrm{Perceiver}(\mathbf{M}, \mathbf{T}^{\mathbf{ctx}}),\\ \widetilde{\mathbf{V}}^{\mathbf{l}} &= \mathrm{Perceiver}(\widetilde{\mathbf{T}}^{\mathbf{ctx}}, \mathbf{V}^{\mathbf{l}}),\end{aligned} \tag{16}$$

where M is a learnable matrix, T^{ctx} denotes the clinical context feature, and $\widetilde{V}^l$ is the final longitudinal visual feature enriched with clinical context.

We then apply global average pooling to obtain instance-level visual and textual representations:

$$\mathbf{v}^{GA} = \mathrm{Pool}(\widetilde{\mathbf{V}}^{\mathbf{l}}), \quad \mathbf{t}^{GA} = \mathrm{Pool}(\mathbf{T}^{\mathbf{c}}). \tag{17}$$

Given a mini-batch of N image-report pairs, the image-to-report global alignment loss is:

$$\mathcal{L}^{GA}_{v2t} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\mathrm{sim}(\mathbf{v}^{GA}_i, \mathbf{t}^{GA}_i)/\tau)}{\sum_{j=1}^{N}\exp(\mathrm{sim}(\mathbf{v}^{GA}_i, \mathbf{t}^{GA}_j)/\tau)}. \tag{18}$$

The report-to-image loss $\mathcal{L}^{GA}_{t2v}$ is defined symmetrically. The global instance-level alignment loss is:

$$\mathcal{L}_{GA} = \frac{1}{2}\left(\mathcal{L}^{GA}_{v2t} + \mathcal{L}^{GA}_{t2v}\right). \tag{19}$$

Overall Objective. The overall objective of Stage 1 combines temporal disentanglement and global instance-level alignment:

$$\mathcal{L}_{stage1} = (1-\lambda)\mathcal{L}_{TD} + \lambda\mathcal{L}_{GA}, \tag{20}$$

where λ is implemented as a learnable loss-balancing coefficient.

3.4 Stage 2: Progressive Supervised Fine-tuning

Stage 2 aims to transfer the temporal evidence learned in Stage 1 to report generation and further guide the model with structured clinical supervision. It consists of two parts: integrating temporally disentangled visual features into the decoder, and progressively fine-tuning the model with observation label generation, clinical triple generation, and report generation.

Temporal Feature Integration. We first concatenate the three disentangled temporal visual features along the sequence dimension:

$$\mathbf{V}^{\mathbf{td}} = [\mathbf{V}^{\mathbf{inv}}; \mathbf{V}^{\mathbf{cur}}; \mathbf{V}^{\mathbf{pri}}]. \tag{21}$$

To condition temporal evidence on clinical context, we fuse the condensed context feature $\widetilde{T}^{ctx}$ with V^{td} using a Perceiver:

$$\widetilde{\mathbf{V}}^{\mathbf{td}} = \mathrm{Perceiver}(\widetilde{\mathbf{T}}^{\mathbf{ctx}}, \mathbf{V}^{\mathbf{td}}). \tag{22}$$

Here, $\widetilde{V}^{td}$ denotes temporal visual feature enriched with clinical context. Finally, we concatenate the condensed clinical context feature $\widetilde{T}^{ctx}$, the longitudinal visual feature $\widetilde{V}^l$, and the temporal visual feature $\widetilde{V}^{td}$ as decoder memory:

$$\mathbf{Z}^{\mathbf{dec}} = [\widetilde{\mathbf{T}}^{\mathbf{ctx}}; \widetilde{\mathbf{V}}^{\mathbf{l}}; \widetilde{\mathbf{V}}^{\mathbf{td}}]. \tag{23}$$

Progressive Supervised Fine-tuning. Free-text reports are heterogeneous and often contain sentences not directly tied to abnormal findings. To help the model first learn compact clinical concepts, we use observation labels and clinical triples as training-time intermediate supervision before final report generation. The observation labels and clinical triples are extracted from ground-truth reports using CheXpert (Irvin et al., 2019) and RadGraph (Jain et al., 2021), respectively.

For any target sequence $Y = \{y_k\}_{k=1}^{K}$, we define the generation loss as:

$$\mathcal{L}_{gen}(\mathbf{Y}) = -\sum_{k=1}^{K}\log p(y_k \mid y_{<k}, \mathbf{Z}^{\mathbf{dec}}). \tag{24}$$

The model is progressively fine-tuned as:

$$\mathcal{L}_{gen}(\mathbf{Y}^{\mathbf{label}}) \rightarrow \mathcal{L}_{gen}(\mathbf{Y}^{\mathbf{triple}}) \rightarrow \mathcal{L}_{gen}(\mathbf{Y}^{\mathbf{report}}), \tag{25}$$

where Y^{label}, Y^{triple}, and Y^{report} denote the textualized observation labels, clinical triples, and ground-truth report, respectively. During inference, TDPS does not generate labels or triples and directly generates the final report.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate TDPS on two datasets: MIMIC-CXR (Johnson et al., 2019) and MIMIC-ABN (Ni et al., 2020). MIMIC-CXR is the largest publicly available chest X-ray report generation dataset, and we curate longitudinal samples with available prior studies. MIMIC-ABN is a subset of MIMIC-CXR focusing on abnormal findings. We follow the official data splits for all experiments. Detailed dataset statistics are provided in Appendix A.1.

Evaluation Metrics. Following prior work (Chen et al., 2020; Goswami et al., 2025; Liu et al., 2026), we evaluate generated reports using both natural language generation (NLG) and clinical efficacy (CE) metrics. For NLG metrics, we report BLEU-n (B-n), METEOR (MTR), and ROUGE-L (R-L). For CE metrics, we use CheXpert (Irvin et al., 2019) to annotate 14 observations and compute micro-average Precision (P), Recall (R), and F1-score (F1).

Implementation Details. We use RAD-DINO (Pérez-García et al., 2024) as the frozen image encoder, CXR-BERT (Boecking et al., 2022) as the text encoder, and DistilGPT2 (Sanh et al., 2019) as the text decoder. We set the feature dimension to $d = 768$, the number of Perceiver latents to $m = 128$, the maximum generation length to $L_{\max} = 100$, and the beam size to 3. On MIMIC-CXR, TDPS is trained for 10 epochs in Stage 1 and 40 epochs in Stage 2, where Stage 2 consists of label, triple, and report generation fine-tuning for 10, 10, and 20 epochs, respectively. Additional training details are provided in Appendix A.1.

Baselines. We compare TDPS with representative state-of-the-art RRG methods, including general chest X-ray report generation models R2Gen (Chen et al., 2020), CMN (Chen et al., 2021), KiUT (Huang et al., 2023), R2genGPT (Wang et al., 2023), MAN (Shen et al., 2024), DCG (Liu et al., 2024c), BT-LLM (Liu et al., 2024a), SEI (Liu et al., 2024b), Medi-VLM (Goswami et al., 2025), CmEAA (Huang et al., 2025b), and longitudinal RRG models HERGen (Wang et al., 2024), MLRG (Liu et al., 2025), PriorRG (Liu et al., 2026). The general RRG baselines include memory-based, knowledge-enhanced, retrieval-augmented, and large language model-based approaches, while the longitudinal baselines incorporate historical information such as prior images, prior reports, indications, or medical history. Detailed descriptions of all baselines are provided in Appendix A.3.

4.2 Main Results

Table 1 reports the comparison between TDPS and existing state-of-the-art methods on MIMIC-CXR and MIMIC-ABN. Overall, TDPS achieves the best performance on most metrics across both datasets and remains competitive on the remaining ones, demonstrating its strong effectiveness for longitudinal chest X-ray report generation.

On MIMIC-CXR, TDPS consistently improves both NLG and CE metrics over prior methods. The gains on CE metrics indicate that TDPS better captures clinically relevant observations, while the improvements on NLG metrics show that the generated reports remain linguistically close to reference reports. This suggests that temporal feature disentanglement and progressive clinical supervision help the model generate reports that are both clinically informative and textually coherent.

On MIMIC-ABN, TDPS also achieves strong overall performance, obtaining the best results on most metrics. Since MIMIC-ABN focuses on abnormal cases, these results further demonstrate the advantage of explicitly modeling longitudinal evidence and grounding it with structured clinical supervision. Although TDPS is slightly lower than the best competing method on one metric for each dataset, it shows more balanced performance across NLG and CE metrics, indicating better overall report generation quality.

Fine-grained results 14 CheXpert observations are provided in Appendix B.1, where TDPS shows consistent improvements on most abnormalities.

4.3 LLM-based Evaluation

In addition to NLG and CE metrics, we conduct an LLM-based evaluation to assess the clinical consistency and expert alignment of generated reports. Following GREEN, we use the pre-

Dataset	Method	Venue	NLG Metrics ↑						CE Metrics ↑		
			B-1	B-2	B-3	B-4	MTR	R-L	P	R	F1
MIMIC -CXR	KiUT	CVPR'23	0.393	0.243	0.159	0.113	0.160	0.285	0.371	0.318	0.321
	R2GenGPT	Meta-Radio'23	**0.411**	0.267	0.186	0.134	0.160	0.297	0.392	0.387	0.389
	MAN	AAAI'24	0.396	0.244	0.162	0.115	0.151	0.274	0.411	0.398	0.389
	DCG	MM'24	0.397	0.258	0.166	0.126	0.162	0.295	0.441	0.414	0.404
	BT-LLM	AAAI'24	0.402	0.262	0.180	0.128	0.175	0.291	0.465	0.482	0.473
	HERGen	ECCV'24	0.395	0.248	0.169	0.122	0.156	0.285	-	-	-
	MediVLM	EMNLP'25	0.377	0.246	0.158	0.123	0.149	0.281	-	-	-
	CmEAA	COLING'25	0.407	0.255	0.174	0.126	-	0.281	0.505	0.330	0.399
	MLRG	CVPR'25	**0.411**	0.277	0.204	0.158	0.176	0.320	0.549	0.468	0.505
	PriorRG†	AAAI'26	0.399	0.279	0.211	0.167	0.185	0.319	0.538	0.481	0.508
	TDPS (Ours)	-	0.408	**0.286**	**0.220**	**0.176**	**0.190**	**0.330**	**0.555**	**0.483**	**0.516**
	Δ (‰) ↑	-	-3	+7	+9	+9	+5	+10	+6	+2	+8
MIMIC -ABN	R2Gen	EMNLP'20	0.253	0.144	0.092	0.063	0.106	0.229	0.444	0.425	0.434
	CMN	ACL'21	0.256	0.147	0.095	0.066	0.110	0.230	0.466	**0.454**	0.460
	SEI	MICCAI'24	0.267	0.157	0.104	0.073	0.114	0.231	0.460	0.408	0.433
	PriorRG†	AAAI'26	0.287	0.170	0.114	0.082	0.126	0.231	0.485	0.424	0.453
	TDPS (Ours)	-	**0.301**	**0.184**	**0.125**	**0.091**	**0.133**	**0.241**	**0.495**	0.446	**0.469**
	Δ (‰) ↑	-	+14	+14	+11	+9	+7	+10	+10	-8	+9

Table 1: Comparison of report generation performance with state-of-the-art methods on MIMIC-CXR and MIMIC-ABN datasets. †denotes reproduced result. Δ indicates the performance difference between TDPS and the best baseline. The best and second-best values are in bold and underlined, respectively.

	All Stage		Stage 1		Stage 2		NLG Metrics ↑						CE Metrics ↑		
	LD	TFD	$\mathcal{L}_{\text{GA}}$	$\mathcal{L}_{\text{TD}}$	Label	Triple	B-1	B-2	B-3	B-4	MTR	R-L	P	R	F1
(a)	✓						0.378	0.270	0.202	0.162	0.180	0.319	0.534	0.401	0.458
(b)			✓				0.348	0.216	0.148	0.110	0.151	0.268	0.526	0.385	0.445
(c)	✓		✓				0.389	0.273	0.208	0.166	0.183	0.320	0.533	0.453	0.490
(d)	✓		✓		✓		0.395	0.280	0.212	0.168	0.185	0.322	0.535	0.473	0.502
(e)	✓		✓			✓	0.399	0.281	0.214	0.169	0.186	0.324	0.543	0.468	0.502
(f)	✓		✓		✓	✓	0.402	0.282	0.215	0.171	0.187	0.326	0.547	0.474	0.503
(g)	✓	✓	✓		✓	✓	0.400	0.280	0.213	0.170	0.186	0.322	0.544	0.464	0.501
(all)	✓	✓	✓	✓	✓	✓	**0.408**	**0.286**	**0.220**	**0.176**	**0.190**	**0.330**	**0.555**	**0.483**	**0.516**

Table 2: Ablation study on different components on MIMIC-CXR. LD denotes Longitudinal Data. ✓ indicates the component is used; blank entries indicate it is not. The best results are in bold.

trained GREEN-RadLlama2-7B (Ostmeier et al., 2024) to compute two metrics: #Matched Findings and GREEN score. #Matched Findings measures the number of clinically relevant findings in the generated report that match the reference report, while the GREEN score further accounts for clinical significance errors and provides an expert-aligned assessment of report quality. We compare TDPS with representative baselines, including R2Gen (Chen et al., 2020), CMN (Chen et al., 2021), R2GenGPT (Wang et al., 2023), SEI (Liu et al., 2024b), and PriorRG (Liu et al., 2026). As shown in Figure 4, TDPS outperforms all baselines on both metrics, indicating that it generates reports with better clinical consistency and stronger agreement with expert preferences.

4.4 Ablation Study

Table 2 reports ablation studies of TDPS. First, we examine the effect of **longitudinal information and global alignment**. Comparing variant (b) with variant (c), removing longitudinal inputs leads to clear drops in both NLG and CE metrics, showing that prior studies provide useful evidence for longitudinal report generation. Comparing variant (a) with variant (c), removing the global instance-level alignment loss $\mathcal{L}_{GA}$ also degrades performance, indicating that global image-report alignment in Stage 1 helps learn better representations for downstream generation.

Second, we evaluate **temporal feature disentanglement**. Comparing variant (f) with the full model, the TFD modules improve report quality by explicitly modeling invariant, current-specific, and prior-specific features. However, when the temporal disentanglement loss $\mathcal{L}_{TD}$ is removed, variant (g) performs worse than the full model and can even underperform variant (f), suggesting that TFD requires explicit disentanglement constraints to learn meaningful temporal features rather than introducing noisy representations.

Third, we analyze **progressive supervised fine-tuning**. Comparing variants (c), (d), (e), and (f), performance improves as observation label genera-

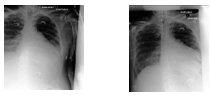
Input Data	Radiologist-written Report	PriorRG-generated Report	TDPS-generated Report (Ours)
Case1 Current Image Prior Image **Clinical Context:** _-year-old male with EKG changes and fever.	Single portable view of the chest is compared to previous exam from ___. Tracheostomy tube **is again noted**. Left PICC tip is not clearly delineated on the current exam. Again there is mild pulmonary vascular congestion. Streaky opacities at the lung bases suggestive of atelectasis ; however infection cannot be excluded. Cardiomediastinal silhouette **is stable** as are the osseous and soft tissue structures.	Single portable view of the chest demonstrates interval placement of a tracheostomy tube, which terminates approximately 2.5 cm above the carina. There is retrocardiac opacity and silhouetting of the descending thoracic aorta. The cardiac silhouette is enlarged but stable in configuration. There is no pneumothorax.	Single portable view of the chest is compared to previous exam from ___. Tracheostomy tube **is again noted.** A left PICC has tip terminating in the mid SVC. There is engorgement of the pulmonary vasculature with indistinct pulmonary vascular markings seen. Cardiomediastinal silhouette **is stable** as are the osseous and soft tissue structures.
Case2  Current Image Prior Image **Clinical Context:** Patient with endotracheal tube placement.	Comparison is made to previous study from ___. There is an **endotracheal tube** whose distal tip is 6.2 cm above the carina appropriately sited. There is a left-sided IJ line with distal lead tip in the mid SVC. There is a nasogastric tube whose tip and sideport are below the GE junction. There is a persistent left retrocardiac opacity. There is some atelectasis at the left lung base.	Comparison is made to previous study from ___. There is a left-sided IJ line with distal lead tip in distal SVC. There is a nasogastric tube whose side port is above the GE junction. This could be advanced several centimeters for more optimal placement. There is unchanged cardiomegaly. There is a left retrocardiac opacity and left-sided pleural effusion.	Comparison is made to prior study from ___. There is an **endotracheal tube** whose distal tip is 5.8 cm above the carina. There is a left-sided central venous catheter with distal lead tip in the mid SVC. There is a nasogastric tube whose tip and side port are below the GE junction. The heart size is enlarged. There is a persistent left retrocardiac opacity. There are no pneumothoraces.

Figure 3: Qualitative comparison of generated reports on MIMIC-CXR. Sentences in the radiologist-written report are color-coded to match corresponding descriptions in AI-generated reports. Correct disease progression patterns and responses to clinical context are underlined.

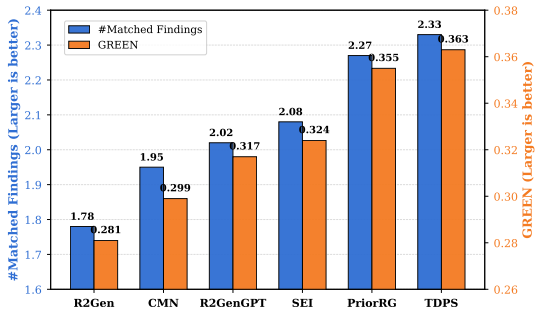


Figure 4: LLM-based expert-aligned evaluation on the MIMIC-CXR dataset. Higher values indicate better clinical consistency and stronger expert alignment.

tion and clinical triple generation are progressively introduced before report generation. This demonstrates that structured clinical signals provide useful intermediate supervision and help the model learn clinically grounded representations.

Overall, the full TDPS achieves the best performance among all variants, validating the complementary contributions of longitudinal information, global alignment, temporal feature disentanglement, temporal disentanglement loss, and progressive clinical supervision. Additional analyses, including detailed ablations of $\mathcal{L}_{TD}$, and the effects of different longitudinal data, are provided in Appendix B.2 and Appendix B.3.

4.5 Qualitative Study

Figure 3 presents qualitative examples on MIMIC-CXR, comparing TDPS with PriorRG (Liu et al., 2026). The more diverse colors in the example reports indicate broader coverage of clinical findings. Compared with PriorRG, TDPS generates reports that better cover clinically relevant findings and more accurately reflect longitudinal changes. In Case 1, TDPS captures stable disease progression cues with expressions such as "again noted" and "stable", showing its ability to compare current and prior studies. In Case 2, TDPS also incorporates clinical context by describing the endotracheal tube, which is omitted or less accurately described by PriorRG. These examples suggest that TDPS can better model disease progression and generate clinically grounded reports. Additional qualitative examples are provided in Appendix B.4.

5 Conclusion

In this paper, we proposed TDPS, a two-stage framework for longitudinal chest X-ray report generation. TDPS first learns temporally disentangled representations from current-prior pairs, decomposing longitudinal evidence into invariant, current-specific, and prior-specific features with dedicated disentanglement, reconstruction, and alignment objectives. It then integrates the disentangled temporal features into report generation and uses observation labels and clinical triples as training-time intermediate supervision for progressive supervised fine-tuning. Experiments on MIMIC-CXR and MIMIC-ABN demonstrate that TDPS achieves strong overall performance, obtaining the best results on most metrics while remaining competitive on the others. These results validate the benefits of temporal feature disentanglement and progressive clinical supervision for generating clinically accurate longitudinal radiology reports.

Limitations

Despite the strong performance of our framework, several limitations remain. First, although TDPS-generated reports are evaluated for clinical consistency via large language models simulating domain experts, their correctness has not been validated by actual medical professionals. Second, the evaluation of disease progression descriptions lacks automated quantitative metrics; developing such metrics would significantly advance longitudinal chest X-ray report generation. In future work, we plan to develop automated metrics for disease progression to further improve and validate the model's clinical relevance.

Ethical Considerations

This work studies longitudinal chest X-ray report generation using publicly available research datasets. Although TDPS improves report generation by modeling temporal evidence and incorporating structured clinical supervision, automatically generated reports may still contain incorrect, incomplete, or hallucinated findings. Therefore, the proposed method should be used only as an assistive tool and must not replace professional radiologists in clinical diagnosis or decision-making.

The model is evaluated on MIMIC-CXR and MIMIC-ABN, and its performance may not fully generalize to other institutions, patient populations, imaging protocols, or reporting styles. Biases in the training data may also be reflected in the generated reports. Before any real-world deployment, further validation by medical professionals and evaluation across diverse clinical settings are required. We will release only research code and experimental configurations, without any patient-identifiable information, and users should comply with relevant data-use agreements and clinical safety regulations.

References

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, and 1 others. 2022. Making the most of text semantics to improve biomedical vision–language processing. In *Proceedings of the 17th European conference on computer vision*, volume 1, pages 1–21.

Shenshen Bu, Taiji Li, Yuedong Yang, and Zhiming Dai. 2024. Instance-level expert knowledge and aggregate discriminative attention for radiology report generation. In *Proceedings of the 35th IEEE/CVF conference on computer vision and pattern recognition*, volume 1, pages 14194–14204.

Wenting Chen, Linlin Shen, Jingyang Lin, Jiebo Luo, Xiang Li, and Yixuan Yuan. 2024. Fine-grained image-text alignment in medical imaging enables explainable cyclic image-report generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 9494–9509.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. 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, pages 5904–5914.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, volume 1, pages 1439–1449.

Debanjan Goswami, Ronast Subedi, and Shayok Chakraborty. 2025. Medivlm: A vision language model for radiology report generation from medical images. In *Proceedings of the 30th Findings of the association for computational linguistics: EMNLP 2025*, volume 1, pages 10287–10304.

Wenjun Hou, Yi Cheng, Kaishuai Xu, Heng Li, Yan Hu, Wenjie Li, and Jiang Liu. 2025a. Radar: Enhancing radiology report generation with supplementary knowledge injection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 26366–26381.

Xuege Hou, Yali Li, and Shengjin Wang. 2025b. Knowledge-driven query network with adaptive cross-view attention for structured radiology report generation. In *Proceedings of the 20th IEEE/CVF International Conference on Computer Vision*, volume 1, pages 1234–1243.

Xiaofei Huang, Wenting Chen, Jie Liu, Qisheng Lu, Xiaoling Luo, and Linlin Shen. 2025a. Damper: A dual-stage medical report generation framework with coarse-grained mesh alignment and fine-grained hypergraph matching. In *Proceedings of the 39th AAAI Annual Conference on Artificial Intelligence*, volume 39, pages 3769–3778.

Xiyang Huang, Yingjie Han, Runzhi Li, Pengcheng Wu, Kunli Zhang, and 1 others. 2025b. Cmeaa: Cross-modal enhancement and alignment adapter for radiology report generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, volume 1, pages 8546–8556.

Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. 2023. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the 34th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 1, pages 19809–19818.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, and 1 others. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the 33th AAAI Annual conference on artificial intelligence*, volume 33, pages 590–597.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 1, pages 4651–4664.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, and 1 others. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.

Jaehwan Jeong, Katherine Tian, Andrew Li, Sina Hartung, Subathra Adithan, Fardad Behzadi, Juan Calle, David Osayande, Michael Pohlen, and Pranav Rajpurkar. 2024. Multimodal image-text matching improves retrieval-based chest x-ray report generation. In *Proceedings of the 7nd International Conference on Medical imaging with deep learning*, volume 1, pages 978–990.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(01):317.

Gaurang Karwande, Amarachi B Mbakwe, Joy T Wu, Leo A Celi, Mehdi Moradi, and Ismini Lourentzou. 2022. Chexrelnet: An anatomy-aware model for tracking longitudinal relationships between chest x-rays. In *Proceedings of the 25th International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 1, pages 581–591.

Philippe Lambin, Ralph TH Leijenaar, Timo M Deist, Jurgen Peerlings, Evelyn EC De Jong, Janita Van Timmeren, Sebastian Sanduleanu, Ruben THM Larue, Aniek JG Even, Arthur Jochems, and 1 others. 2017. Radiomics: The bridge between medical imaging and personalized medicine. *Nature reviews Clinical oncology*, 14(12):749–762.

Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. 2023. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the 34th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 1, pages 3334–3343.

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

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

Kang Liu, Zhuoqi Ma, Xiaolu Kang, Yunan Li, Kun Xie, Zhicheng Jiao, and Qiguang Miao. 2025. Enhanced contrastive learning with multi-view longitudinal data for chest x-ray report generation. In *Proceedings of the 36th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 1, pages 10348–10359.

Kang Liu, Zhuoqi Ma, Xiaolu Kang, Zhusi Zhong, Zhicheng Jiao, Grayson Baird, Harrison Bai, and Qiguang Miao. 2024b. Structural entities extraction and patient indications incorporation for chest x-ray report generation. In *Proceedings of the 27th International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 38, pages 433–443.

Rui Liu, Mingjie Li, Shen Zhao, Ling Chen, Xiaojun Chang, and Lina Yao. 2024c. In-context learning for zero-shot medical report generation. In *Proceedings of the 32nd ACM international conference on multimedia*, volume 1, pages 8721–8730.

Xin Mei, Libin Yang, Dehong Gao, Xiaoyan Cai, Junwei Han, and Tianming Liu. 2025. Adaptive medical topic learning for enhanced fine-grained cross-modal alignment in medical report generation. *IEEE Transactions on Multimedia*, 27(01):5050–5061.

Dang Nguyen, Chacha Chen, He He, and Chenhao Tan. 2023. Pragmatic radiology report generation. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 1, pages 385–402.

Jianmo Ni, Chun-Nan Hsu, Amilcare Gentili, and Julian McAuley. 2020. Learning visual-semantic embeddings for reporting abnormal findings on chest x-rays. In *Findings of the association for computational linguistics: EMNLP 2020*, volume 1, pages 1954–1960.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and 1 others. 2024. Green: Generative radiology report evaluation and error notation. In *Findings of the association for computational linguistics: EMNLP 2024*, volume 1, pages 374–390.

F Pérez-García, H Sharma, S Bond-Taylor, K Bouzid, V Salvatelli, M Ilse, and 1 others. 2024. Rad-dino: Exploring scalable medical image encoders beyond text supervision. *arXiv preprint arXiv:2401.10815*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version

of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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

Fuying Wang, Shenghui Du, and Lequan Yu. 2024. Hergen: Elevating radiology report generation with longitudinal data. In *Proceedings of the 18th European Conference on Computer Vision*, volume 1, pages 183–200.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the 29th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 1, pages 9049–9058.

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

Zhanyu Wang, Luping Zhou, Lei Wang, and Xiu Li. 2021. A self-boosting framework for automated radiographic report generation. In *Proceedings of the 32th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 1, pages 2433–2442.

Heng Yin, Shanlin Zhou, Pandong Wang, Zirui Wu, and Yongtao Hao. 2025. Kia: Knowledge-guided implicit vision-language alignment for chest x-ray report generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, volume 1, pages 4096–4108.

Ke Zhang, Hanliang Jiang, Jian Zhang, Qingming Huang, Jianping Fan, Jun Yu, and Weidong Han. 2024. Semi-supervised medical report generation via graph-guided hybrid feature consistency. *IEEE Transactions on Multimedia*, 26(01):904–915.

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. 2020. When radiology report generation meets knowledge graph. In *Proceedings of the 34th AAAI Annual conference on artificial intelligence*, volume 34, pages 12910–12917.

Qingqing Zhu, Tejas Sudharshan Mathai, Pritam Mukherjee, Yifan Peng, Ronald M Summers, and Zhiyong Lu. 2023. Utilizing longitudinal chest x-rays and reports to pre-fill radiology reports. In *Proceedings of the 26th International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 1, pages 189–198.

Appendix Contents

A Experimental Settings **11**
A.1 Detailed Dataset Statistics 11
A.2 Implementation Details 11
A.3 Baselines Details 12
B Additional Experiments. **12**
B.1 More detailed study on the clinical accuracy of the 14 observations. 12
B.2 More detailed ablation study of Temporal Disentanglement Loss $\mathcal{L}_{TD}$.. 13
B.3 More detailed ablation study on the effectiveness of longitudinal data. 13
B.4 More detailed qualitative study.. 13

A Experimental Settings

A.1 Detailed Dataset Statistics

Statistic	MIMIC-CXR			MIMIC-ABN		
	Train	Val	Test	Train	Val	Test
#Image	239,998	2,113	3,852	69,641	586	844
#Sample	239,998	2,113	3,852	69,641	586	844
#Report	150,957	1,182	2,343	34,763	263	378
%PI	60.5	61.3	87.7	52.7	50.6	89.2
%IND	66.4	65.4	57.3	64.6	62.7	56.3
%HIS	30.6	31.5	37.4	34.9	37.3	43.4

Table 3: Dataset statistics for MIMIC-CXR and MIMIC-ABN. #Image, #Sample, and #Report denote the number of images, samples, and reports, respectively. %PI, %IND, and %HIS represent the percentage of reports containing prior imaging, indication, and history information, respectively.

We adhere to official splits for these datasets and summarize the sample counts for the training, validation, and test set in Table 3.

A.2 Implementation Details

Model architecture. The image encoder (RAD-DINO) (Pérez-García et al., 2024) remains frozen throughout training. In contrast, the text encoder (CXR-BERT) (Boecking et al., 2022), the report generator (DistilGPT2) (Sanh et al., 2019), the Temporal Feature Disentangler (TFD) module, and the Cross Attention Fusion Network (CAFN) are fine-tuned during both training stages. All experiments are conducted on two NVIDIA RTX 3090 GPUs.

Optimizer and scheduling. We adopt the AdamW optimizer for all components. The ReduceLROnPlateau scheduler is employed with a

patience of 5. Early stopping is applied with a patience of 15 epochs to prevent overfitting.

Training schedule. We adopt a two-stage training pipeline. For the MIMIC-CXR (Johnson et al., 2019) dataset, the model is trained for 10 epochs in Stage 1 with a batch size of 24. Stage 2 is then fine-tuned for a total of 40 epochs, consisting of 10 epochs for the label generation task and 10 epochs for the triple generation task, followed by 20 epochs for the report generation task, all with a batch size of 16. In Stage 1, the learning rate is set to 5e-5. In Stage 2, the report generator is optimized with a learning rate of 5e-5, while all other components are trained with a learning rate of 5e-6. For the MIMIC-ABN (Ni et al., 2020) dataset, as a derivative of MIMIC-CXR, the model is initialized with weights pre-trained on MIMIC-CXR and fine-tuned for 40 epochs using the same strategy (Progressive Supervised Fine-tuning), with a learning rate of 5e-6 and a batch size of 16.

For the TDPS checkpoint evaluated on the MIMIC-CXR test set, the learnable loss-balancing coefficient λ converges to 0.376. The reconstruction branch obtains an average cosine similarity of 0.948, suggesting that temporal disentanglement preserves rich representation information. The intermediate label and triple generation tasks achieve F1-scores of 0.836 and 0.591, respectively.

A.3 Baselines Details

The chest X-ray report generation methods selected for comparison are briefly introduced as follows:

1) **R2Gen** (Chen et al., 2020) designs a relational memory to record key information, thereby improving the standardization and efficiency of report generation.

2) **CMN** (Chen et al., 2021) employs shared memory as a medium to implicitly align visual and textual features, thereby enhancing the quality of generated reports.

3) **KiUT** (Huang et al., 2023) integrates a U-shaped connection architecture and an injection-based knowledge distiller to alleviate the insufficient utilization of clinical knowledge in chest X-ray report generation.

4) **R2GenGPT** (Wang et al., 2023) combines a large language model with a visual encoder to achieve chest X-ray report generation.

5) **MAN** (Shen et al., 2024) proposes a more efficient memory-aligned embedding method to help the model align visual and textual features, generating coherent and information-rich reports.

6) **DCG** (Liu et al., 2024c) leverages a context graph to structurally model image features, enhancing the representation of key clinical information to improve report generation.

7) **BT-LLM** (Liu et al., 2024a) mitigates the knowledge gap between general LLMs and the medical domain by retrieving in-domain knowledge examples.

8) **HERGen** (Wang et al., 2024) employs a group causal transformer to effectively integrate longitudinal data from patient visits, compensating for critical temporal information in patient imaging history.

9) **SEI** (Liu et al., 2024b) avoids noise interference from redundant information by employing structural feature injection, improving the model's ability to describe abnormal regions.

10) **MediVLM** (Goswami et al., 2025) incorporates a pre-trained object detector to extract salient anatomical regions from images for report generation.

11) **CmEAA** (Huang et al., 2025b) enhances cross-modal representation learning through cross-modal enhancement and alignment adapters for radiology report generation.

12) **MLRG** (Liu et al., 2025) integrates multi-view longitudinal information to improve chest X-ray report generation.

13) **PriorRG** (Liu et al., 2026) generates reports through prior-guided contrastive pre-training and coarse-to-fine decoding, enhancing disease progression capture and clinical accuracy.

B Additional Experiments

B.1 More detailed study on the clinical accuracy of the 14 observations.

We further conduct a comprehensive evaluation of the clinical accuracy of the 14 observation labels extracted by CheXpert. Table 6 and Table 7 present

Method	NLG Metrics ↑			CE Metrics ↑		
	B-1	B-4	MTR	P	R	F1
TDPS	**0.408**	**0.176**	**0.190**	**0.555**	**0.483**	**0.516**
w/o $\mathcal{L}_{intra}$	0.403	0.172	0.187	0.548	0.460	0.500
w/o $\mathcal{L}_{cross}$	0.405	0.174	0.188	0.549	0.476	0.510
w/o $\mathcal{L}_{recon}$	0.403	0.173	0.188	0.544	0.474	0.507

Table 4: Detailed ablation study of the temporal disentanglement loss $\mathcal{L}_{TD}$. The results show that $\mathcal{L}_{intra}$, $\mathcal{L}_{cross}$, and $\mathcal{L}_{recon}$ all contribute to language quality and clinical accuracy.

Method	NLG Metrics ↑			CE Metrics ↑		
	B-1	B-4	MTR	P	R	F1
TDPS	**0.408**	**0.176**	**0.190**	**0.555**	**0.483**	**0.516**
w/o PI	0.388	0.158	0.176	0.539	0.426	0.476
w/o IND	0.376	0.134	0.169	0.518	0.412	0.460
w/o HIS	0.380	0.137	0.168	0.528	0.418	0.466

Table 5: Results of TDPS with different longitudinal data on the MIMIC-CXR test set, where longitudinal data include Prior Image (PI), Indication (IND), and Medical History (HIS).

the precision, recall, and F1 scores on the MIMIC-CXR and MIMIC-ABN datasets, respectively, comparing TDPS against the state-of-the-art PriorRG model. On the MIMIC-CXR dataset (Table 6), TDPS achieves the best performance on the majority of individual observations and obtains superior overall results, with a micro-averaged F1 of 0.516 and a macro-averaged F1 of 0.373, outperforming PriorRG by clear margins. Notable improvements are observed in clinically critical categories such as Edema, Pneumonia, Atelectasis, Pleural Effusion, and Support Devices. On the MIMIC-ABN dataset (Table 7), which is derived from MIMIC-CXR, TDPS also demonstrates consistent superiority, achieving a micro-averaged F1 of 0.469 and a macro-averaged F1 of 0.320, again surpassing PriorRG across most categories. These gains are particularly evident in Lung Opacity, Lung Lesion, Enlarged Cardiomediastinum. Taken together, these results demonstrate that TDPS consistently improves clinical accuracy across diverse observation categories and datasets, validating the effectiveness of our proposed TDPS framework in generating clinically precise radiology reports.

B.2 More detailed ablation study of Temporal Disentanglement Loss $\mathcal{L}_{TD}$.

We further analyze the effectiveness of each component in the temporal disentanglement loss by removing $\mathcal{L}_{intra}$, $\mathcal{L}_{cross}$, and $\mathcal{L}_{recon}$, respectively. As shown in Table 4, all ablated variants consistently underperform the full TDPS model, confirming the necessity of the three complementary constraints.

Removing $\mathcal{L}_{intra}$ leads to the largest decrease in CE F1, from 0.516 to 0.500, suggesting that intra-modal disentanglement is crucial for reducing redundancy among invariant, current-specific, and prior-specific features. Without this constraint, different temporal components may become less distinguishable within each modality. Removing $\mathcal{L}_{cross}$ also degrades both NLG and CE metrics, indicating that cross-modal temporal alignment helps match corresponding visual and textual temporal evidence. In addition, removing $\mathcal{L}_{recon}$ decreases precision and F1, demonstrating that the reconstruction constraint helps preserve information completeness during feature disentanglement.

These results show that the three components of $\mathcal{L}_{TD}$ play complementary roles: $\mathcal{L}_{intra}$ improves intra-modal separability, $\mathcal{L}_{cross}$ enhances cross-modal semantic alignment, and $\mathcal{L}_{recon}$ preserves the completeness of disentangled representations. Together, they enable the model to learn more reliable progression-aware temporal evidence for longitudinal report generation.

B.3 More detailed ablation study on the effectiveness of longitudinal data.

We conduct an ablation study to assess the contribution of each longitudinal data source—Prior Image (PI), Indication (IND), and Medical History (HIS)—to the overall performance of TDPS. Table 5 reports the results on the MIMIC-CXR test set. The full TDPS model consistently achieves the best performance across all NLG and CE metrics. Removing any single source of longitudinal data leads to a decline in both language generation quality and clinical accuracy. These results validate the importance of incorporating all three types of longitudinal information into the TDPS framework.

B.4 More detailed qualitative study.

Additional qualitative analysis is shown in Figure 5. The results show that TDPS outperforms the baseline PriorRG (Liu et al., 2026) in generating radiology reports. In cases with the prior image but no clinical context (Case 3), the model effectively captures disease progression by describing "residual right internal jugular vascular sheath remaining in place" and "Cardiomediastinal contours are sta-

Observation	%	PriorRG ↑			TDPS ↑		
		P	R	F1	P	R	F1
Enlarged Cardiomediastinum	9.9%	0.361	0.399	0.379	**0.371**	**0.435**	**0.401**
Cardiomegaly	14.8%	0.629	**0.634**	**0.631**	**0.632**	0.611	0.621
Lung Opacity	13.8%	0.562	0.348	0.430	**0.618**	**0.358**	**0.453**
Lung Lesion	2.5%	**0.323**	**0.076**	**0.123**	0.244	0.038	0.066
Edema	8.3%	0.504	0.407	0.450	**0.520**	**0.439**	**0.476**
Consolidation	3.3%	**0.302**	**0.150**	**0.201**	0.245	0.139	0.177
Pneumonia	4.4%	0.287	**0.214**	0.245	**0.327**	0.209	**0.255**
Atelectasis	10.9%	0.492	0.477	0.484	**0.521**	**0.491**	**0.505**
Pneumothorax	1.0%	0.289	0.280	0.284	**0.354**	**0.290**	**0.319**
Pleural Effusion	12.4%	0.716	0.613	0.661	**0.741**	**0.630**	**0.681**
Pleural Other	1.6%	0.200	0.065	0.099	**0.231**	**0.071**	**0.109**
Fracture	1.8%	0.145	**0.054**	**0.079**	**0.178**	0.043	0.070
Support Devices	12.8%	0.763	0.749	0.756	**0.772**	**0.753**	**0.762**
No Finding	2.4%	0.223	0.545	0.317	**0.231**	**0.589**	**0.332**
micro avg	-	0.535	0.474	0.502	**0.555**	**0.483**	**0.516**
macro avg	-	0.414	0.358	0.367	**0.427**	**0.364**	**0.373**

Table 6: Clinical accuracy of 14 observations on the MIMIC-CXR dataset. Best values are highlighted in bold.

ble." When only clinical context is available—such as "The pre-existing parenchymal opacities at the right lung" (Case 4) — TDPS generates radiology reports that align with these clinical contexts, accurately reflecting surgical history "status post wedge resection" in Case 4. Even in the absence of both prior images and clinical context (Case 5), our TDPS still produces reports that better align with the clinical information.

Observation	%	PriorRG ↑			TDPS ↑		
		P	R	F1	P	R	F1
Enlarged Cardiomediastinum	9.9%	0.186	0.145	0.163	**0.205**	**0.236**	**0.219**
Cardiomegaly	14.8%	0.497	**0.737**	0.594	**0.518**	0.728	**0.605**
Lung Opacity	13.8%	0.602	0.379	0.465	**0.637**	**0.440**	**0.520**
Lung Lesion	2.5%	0.0	0.0	0.0	**0.300**	**0.032**	**0.057**
Edema	8.3%	**0.500**	0.350	0.412	0.478	**0.401**	**0.437**
Consolidation	3.3%	0.250	**0.215**	**0.231**	**0.323**	0.154	0.208
Pneumonia	4.4%	**0.333**	0.158	**0.214**	0.220	0.158	0.184
Atelectasis	10.9%	0.355	0.569	0.437	**0.414**	**0.574**	**0.481**
Pneumothorax	1.0%	0.0	0.0	0.0	0.0	0.0	0.0
Pleural Effusion	12.4%	**0.674**	**0.705**	**0.689**	0.645	0.630	0.637
Pleural Other	1.6%	0.048	0.016	0.024	**0.174**	**0.065**	**0.094**
Fracture	1.8%	0.0	0.0	0.0	**0.200**	**0.013**	**0.025**
Support Devices	12.8%	0.737	0.709	0.723	**0.738**	**0.738**	**0.738**
No Finding	2.4%	**0.295**	0.310	**0.303**	0.237	0.310	0.269
micro avg	-	0.485	0.424	0.453	**0.495**	**0.446**	**0.469**
macro avg	-	0.320	0.307	0.304	**0.363**	**0.320**	**0.320**

Table 7: Clinical accuracy of 14 observations on the MIMIC-ABN dataset. For each row and metric, the larger value between PriorRG and TDPS is bolded.

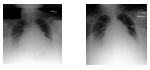
Input Data	Radiologist-written Report	PriorRG-generated Report	TDPS-generated Report (Ours)
Case3  Curren Image Prior Image **Clinical Context:** Null	Right internal jugular sheath ends at upper SVC. A single mediastinal drain tube is present on the right side. The appearance of the post operative widened mediastinum is unchanged since ____. Bilateral, confluent, lung opacities suggesting moderate pulmonary edema has improved asymmetrically on the left side, but unchanged on the right. Pleural effusions, if any, is mild bilaterally.	Interval removal of Swan-Ganz catheter with residual right internal jugular vascular sheath, with no visible pneumothorax. Stable cardiomegaly and persistent pulmonary vascular congestion. Slight worsening of bibasilar atelectasis and persistent small pleural effusions.	Various indwelling support and monitoring devices have been removed with residual right internal jugular vascular sheath remaining in place. Cardiomediastinal contours are stable in the postoperative period. Pulmonary vascular congestion is accompanied by mild interstitial edema and small bilateral pleural effusions, right greater than left.
Case4 Current Image **Clinical Context:** Lung cancer, status post wedge resection. Evaluation for acute process.	As compared to the previous radiograph, there is no relevant change. Unchanged appearance of the left postoperative lung with decrease in size of the hemithorax. Unchanged opacities at the right lung base, potentially caused by atelectasis or, possibly, aspiration. Short-term further radiographic followup should be performed. No larger pleural effusions.	As compared to the previous radiograph, there is no relevant change. Status post tracheostomy. Known parenchymal opacities in the left lung with areas of atelectasis in the retrocardiac lung regions. The presence of a minimal left pleural effusion cannot be excluded. On the right, there is a newly appeared parenchymal opacity at the lung apex, likely atelectatic in origin. Unchanged size of the cardiac silhouette.	As compared to the previous radiograph, the patient has undergone bronchoscopy. There is no evidence of pneumothorax. The pre-existing parenchymal opacities at the right lung base and in the left perihilar areas are unchanged. No larger pleural effusions. Unchanged moderate cardiomegaly and tortuosity of the thoracic aorta.
Case5 Current Image **Clinical Context:** Null	Left transvenous pacemaker leads are in standard position. Top normal heart size, mediastinal and hilar contours are unchanged. New left internal jugular line ends at the mid SVC. Core- valve prosthesis is unchanged in position. Mild bibasilar atelectasis is unchanged. There are no new lung opacities which are concerning for pneumonia.	Interval placement of left internal jugular central venous catheter with tip terminating in the mid superior vena cava, with no visible pneumothorax. The lungs are well expanded and clear. There is no focal consolidation, effusion, or pneumothorax. The heart is moderately enlarged.	Left internal jugular central venous catheter terminates in the lower superior vena cava, with no visible pneumothorax. A left chest wall pacer / AICD has leads terminating in the right atrium and right ventricle. Improved aeration at the lung bases with minimal residual atelectasis remaining. No new areas of consolidation to suggest the presence of pneumonia.

Figure 5: Qualitative comparison of generated reports on MIMIC-CXR. Sentences in the radiologist-written report are color-coded to match corresponding descriptions in AI-generated reports. Correct disease progression patterns and responses to clinical context are underlined.