Learning Directional Semantic Transitions for Longitudinal Chest X-ray Analysis

Zhangfeng Hu[1], Zefan Yang[1], Ge Wang[1], Tanveer Syeda-Mahmood[2], Anushree Burade[3], Mannudeep K. Kalra[3], and Pingkun Yan[1]✉

[1] Rensselaer Polytechnic Institute, Troy, NY, USA
yanp2@rpi.edu
[2] Stanford University, Stanford, CA, USA
[3] Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

Abstract. Chest X-ray (CXR) interpretation often requires longitudinal comparison to assess disease progression. Existing approaches typically rely on temporal feature fusion or inter-study discrepancy modeling, yet remain limited in capturing subtle progression semantics and overlook the inherently directional nature of disease trajectories. In this paper, we propose ProTrans, a novel vision-language pretraining framework that formulates disease progression as a directional semantic transition between paired CXR studies. ProTrans leverages radiology reports to anchor individual CXR representations within interpretable disease states, and introduces a learnable progression feature map to explicitly encode semantic shifts between states, aligned with report-derived progression descriptions. To enforce direction-aware perception, ProTrans incorporates a reversed temporal modeling process and imposes bidirectional reconstruction consistency across states and transitions, thereby disentangling directional semantics and promoting coherent trajectory modeling. Extensive experiments on longitudinal downstream tasks, including disease progression classification and progression captioning, demonstrate that ProTrans consistently outperforms existing methods, establishing a unified pretraining framework for longitudinal CXR understanding. https://github.com/RPIDIAL/ProTrans

Keywords: Chest X-ray · Longitudinal analysis · Vision-Language Pretraining · Contrastive learning · Progression.

1 Introduction

Recent advances in vision–language pretraining (VLP) have substantially improved chest X-ray (CXR) analysis, demonstrating superior performance in tasks such as abnormality localization, classification, and report generation [5,9,19,24–26]. By aligning visual representations with clinical concepts derived from paired radiology reports, these models capture rich semantic information essential for CXR interpretation. Despite these successes, existing approaches primarily focus on single-study analysis. In clinical practice, however, radiologists routinely

compare the current CXR with prior examinations to identify subtle temporal changes and assess disease progression [31]. Such longitudinal assessment is pivotal to diagnosis, treatment, and prognosis.

To bridge this gap, recent efforts have begun extending vision architectures to capture longitudinal information, which can be broadly categorized into two groups. The first group includes temporal fusion modules to aggregate multi-time-point features, including cross-attention encoders [4, 8, 14, 27] to exploit temporal correlations, temporal alignment connectors to integrate multi-scale context from priors [30], and group causal transformers to model temporally ordered dependencies in long sequences [18]. The second category models inter-study discrepancies by extracting patch-, region-, or multi-level feature differences [17, 20, 29]. Despite promising results, it remains unclear whether these methods can accurately and reliably distill clinically meaningful progression semantics. Two key challenges persist. First, consecutive CXRs are dominated by invariant anatomical structures, where true progression cues are subtle and easily overwhelmed. Meanwhile, pseudo-changes caused by inconsistent imaging conditions can confound both naive feature fusion and differentiation modeling. Second, disease progression is inherently directional. In addition to change detection, clinically useful models should characterize their trajectories (*e.g.*, improving versus worsening). However, this directional nature of progression remains largely underexplored in existing approaches.

To address these limitations, we propose **ProTrans**, a novel VLP framework for progression representation learning that models disease **Pro**gression as directional semantic **Trans**itions between consecutive CXR studies. Rather than capturing temporal change through feature-level fusion or subtraction, ProTrans formulates progression as a learnable transformation in a semantic space. Specifically, we introduce a learnable progression feature map that resides between two longitudinal CXRs, as a bridge to encode the semantic transition between prior and current states. To achieve this goal, we first design a state contrastive alignment strategy that leverages associated radiology reports to ground each CXR representation in a report-based semantic state. Then, a transition contrastive constraint aligns progression representations with disease transition cues derived from paired reports, encouraging them to capture clinical state changes. To further enforce temporal directionality, ProTrans introduces a reversed temporal learning process to estimate reversed progression representations. Then we impose a bidirectional reconstruction consistency ensuring that the current state can be reconstructed from the prior state and progression, and vice versa using the reversed progression. By jointly learning forward and backward transitions, the model disentangles directional temporal semantics and improves sensitivity to trajectory-specific disease evolution. Together, these components enable a unified pretraining framework for semantically-sensitive and direction-aware progression representation learning.

Our contributions are three-fold. **(1)** We introduce ProTrans, a novel progression representation learning framework that formulates disease progression as directional semantic transitions, enhanced by reversed temporal modeling and

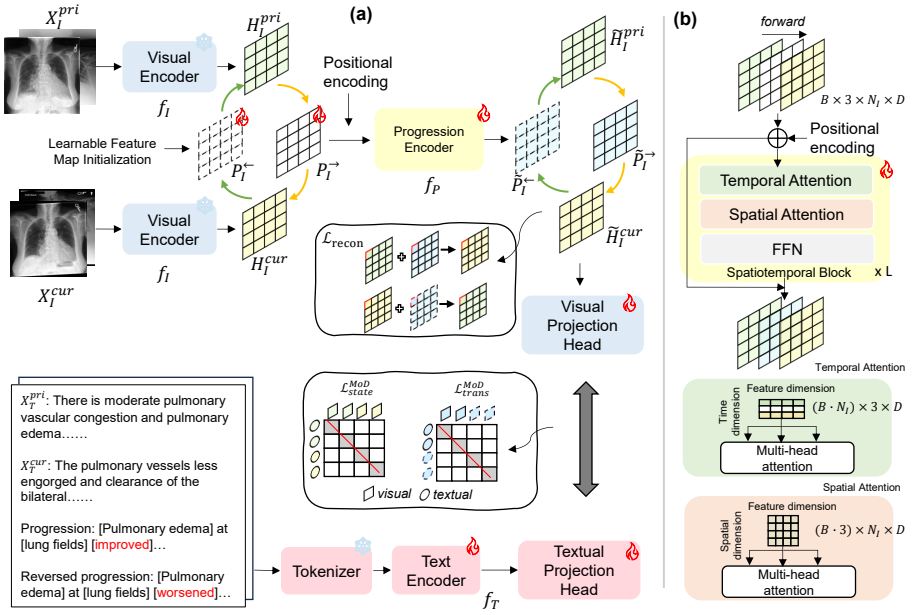

Fig. 1. Overview of ProTrans. (a) Pretraining pipeline for progression representation learning, and (b) Architecture of the progression encoder.

bidirectional reconstruction consistency to capture a coherent disease trajectory. **(2)** We propose joint state and transition contrastive alignment that grounds bi-temporal clinical states and their interval transitions within a shared semantic space based on radiology reports, enabling well-grounded and interpretable encoding of disease evolution. **(3)** We demonstrate that ProTrans improves performance across longitudinal downstream tasks, including disease progression classification and progression captioning, providing a unified pretraining foundation for longitudinal CXR understanding.

2 Method

Fig.1(a) illustrates the framework of ProTrans with details showing how disease progression is modeled as a directional semantic transition between longitudinal CXR studies. The framework learns state representations and progression transformations jointly, integrating bidirectional temporal modeling and reconstruction constraints to enforce trajectory-aware representation learning.

Bidirectional Progression Representation Modeling. Given a pair of prior and current CXR images X_I^{pri}, X_I^{cur} and their corresponding reports X_T^{pri}, X_T^{cur}, we first use visual encoder f_I and text encoder f_T to extract visual and textual embeddings as $\mathbf{H}_I^t \in \mathbb{R}^{N_I \times d}$, $\mathbf{H}_T^t \in \mathbb{R}^{N_T \times d}$, where $t \in \{\text{pri}, \text{cur}\}$. N_I, N_T denote the numbers of visual and textual tokens, and d is the token dimension.

Then to explicitly model progression semantics, we introduce a learnable progression feature map $\mathbf{P}_I^{\rightarrow} \in \mathbb{R}^{N_I \times d}$, which is inserted between prior and current image embeddings to form a directed temporal sequence as $\mathbf{S}^{\rightarrow} = [\mathbf{H}_I^{pri}; \mathbf{P}_I^{\rightarrow}; \mathbf{H}_I^{cur}] \in \mathbb{R}^{3 \times N_I \times d}$. Positional encoding is integrated to preserve the ordering information within the sequence. The sequence is then processed by a progression encoder f_P, as depicted in Fig.1(b), composed of L spatiotemporal blocks. Each block contains a temporal attention layer to model cross-time dependencies, a spatial attention layer to capture anatomical correspondence, and a feed-forward network (FFN), all equipped with residual connections. After processing by f_P, we obtain a set of progression-aware representations as $[\tilde{\mathbf{H}}_I^{pri}, \tilde{\mathbf{P}}_I^{\rightarrow}, \tilde{\mathbf{H}}_I^{cur}] = f_P(\mathbf{S}^{\rightarrow})$. To enhance sensitivity to temporal directionality, we model the reversed progression by constructing a backward sequence as $\mathbf{S}^{\leftarrow} = [\mathbf{H}_I^{cur}; \mathbf{P}_I^{\leftarrow}; \mathbf{H}_I^{pri}]$, where $\mathbf{P}_I^{\leftarrow}$ is a learnable reversed progression map and produces the corresponding reversed progression representation $\tilde{\mathbf{P}}_I^{\leftarrow}$.

State Contrastive Alignment. To facilitate above progression representation learning, we devise joint state-transition contrastive alignment to ground bi-temporal clinical states and their interval transitions within a shared semantic space based on radiology reports. First, to ensure that each CXR representation corresponds to an interpretable clinical state, we perform cross-modal contrastive alignment between images and their paired reports at each time point. Specifically, given a current image $\tilde{\mathbf{H}}_{I,i}^{cur}$ of sample i, its associated report $\mathbf{H}_{T,i}^{cur}$ is treated as the positive, while all other reports in the mini-batch B—including those from different samples and the prior report of sample i—are treated as negatives. This design enforces fine-grained temporal discrimination, encouraging the model to distinguish subtle semantic differences across longitudinal states. The same alignment is applied to the prior study. Formally, let $\mathbf{h}_{I,i}^t$ and $\mathbf{h}_{T,i}^t$ denote the normalized and projected embeddings from $\tilde{\mathbf{H}}_{I,i}^t$, $\mathbf{H}_{T,i}^t$. The image-to-text similarity logits $p_{I2T}^t \in \mathbb{R}^{B \times B}$ are computed as:

$$p_{I2T,i}^t = \frac{\exp(\text{sim}(\mathbf{h}_{I,i}^t, \mathbf{h}_{T,i}^t)/\tau)}{\sum_{k=1}^{B}[\exp(\text{sim}(\mathbf{h}_{I,i}^t \mathbf{h}_{T,k}^t)/\tau) + \exp(\text{sim}(\mathbf{h}_{I,i}^t, \mathbf{h}_{T,k}^{\bar{t}})/\tau)]}, \tag{1}$$

where τ is a temperature parameter and $\bar{t}$ denotes the complementary timepoint (i.e., $\bar{\text{pri}} = \text{cur}$ and $\bar{\text{cur}} = \text{pri}$), $\text{sim}(\cdot,\cdot)$ denotes a token-wise similarity function [28]. Similarly, the text-to-image similarity logits are described as p_{T2I}^t. Then, the cross-modal alignment loss is defined as:

$$\mathcal{L}_{\text{state}} = -\frac{1}{2B} \sum_{t \in \{pri, cur\}} \sum_{k=1}^{B} (q_k^t \log p_{I2T,k}^t + q_k^t \log p_{T2I,k}^t), \tag{2}$$

where q^t is the one-hot alignment label matrix. Since longitudinal CXR studies often exhibit semantic overlap, such strict one-hot alignment may be overly restrictive. We therefore introduce additional soft targets [13], generated by a momentum model—an exponential-moving-average (EMA) version of the base network. Specifically, momentum image-text similarity logits $q_{I2T}'^t$ and $q_{T2I}'^t$ are computed using momentum embeddings $\mathbf{h}'$ replacing $\mathbf{h}$ in Eq.(1), and then serve as soft supervisory signals replacing q^t in Eq.(2) to obtain a soft alignment loss

$\mathcal{L}_{\text{state}}^{soft}$. The final state alignment loss is defined as:

$$\mathcal{L}_{\text{state}}^{\text{MoD}} = \alpha \mathcal{L}_{\text{state}} + (1-\alpha)\mathcal{L}_{\text{state}}^{soft}, \tag{3}$$

where α controls the trade-off between hard and soft alignment supervision.

Transition Contrastive Alignment. To ensure that learned progression representations $\tilde{\mathbf{P}}_I^{\rightarrow}$, $\tilde{\mathbf{P}}_I^{\leftarrow}$ capture meaningful transition semantics, we further align them with corresponding textual descriptions of disease progression. Specifically, we construct structured transition texts in the form of "[finding] at [position] [progression]", based on the disease progression annotations from Chest ImaGenome dataset [21]. For reversed transitions, we simply use opposite phrasing (e.g., replacing "improved" with "worsened", while "no change" remains unchanged), thereby encoding direction-aware semantics. These structured descriptions are encoded by the text encoder f_T into transition embeddings $\mathbf{P}_T^{\rightarrow}$ and $\mathbf{P}_T^{\leftarrow}$. We then adopt a transition alignment scheme to match progression representations with their corresponding transition embeddings. For each progression representation, its paired description is treated as the positive, while all other transition descriptions—including reversed-direction descriptions—serve as negatives. Similar to Eq. (3), the transition alignment loss is defined as $\mathcal{L}_{\text{trans}}^{\text{MoD}} = \alpha \mathcal{L}_{\text{trans}} + (1-\alpha)\mathcal{L}_{\text{trans}}^{\text{soft}}$.

Bidirectional Reconstruction. While the above alignment objectives ground progression representations in the semantic space, they do not explicitly enforce temporal consistency between the learned transitions and the underlying disease states. To further regularize the progression modeling process, we introduce a bidirectional reconstruction objective that requires progression representations to faithfully describe state evolution across time. Specifically, given $\tilde{\mathbf{P}}_I^{\rightarrow}$ and $\tilde{\mathbf{P}}_I^{\leftarrow}$, we employ a lightweight reconstruction decoder f_R to reconstruct target state embedding at each timepoint from the opposite state and corresponding progression representation, as $\hat{\mathbf{H}}_I^{cur} = f_R([\tilde{\mathbf{H}}_I^{pri}; \tilde{\mathbf{P}}_I^{\rightarrow}]), \hat{\mathbf{H}}_I^{pri} = f_R([\tilde{\mathbf{H}}_I^{cur}; \tilde{\mathbf{P}}_I^{\leftarrow}])$, where $\hat{\mathbf{H}}_I^{cur}$ and $\hat{\mathbf{H}}_I^{pri}$ denote reconstructed current and prior embeddings respectively. To enforce fine-grained consistency between original and reconstructed features, we adopt an InfoNCE [16]-style token-level reconstruction objective. For each reconstructed token $\hat{\mathbf{H}}_{I,i}^{t,n}$ at spatial position n in sample i, its original counterpart $\tilde{\mathbf{H}}_{I,i}^{t,n}$ is treated as the positive, while all other original tokens serve as negatives. The reconstruction loss is defined as:

$$\mathcal{L}_{\text{recon}} = -\frac{1}{BN_I} \sum_{t\in\{pri,cur\}} \sum_{i=1}^{B} \sum_{n=1}^{N_I} \log \frac{\exp(\cos(\hat{\mathbf{H}}_{I,i}^{t,n}, \tilde{\mathbf{H}}_{I,i}^{t,n})/\tau)}{\sum_{j=1}^{B}\sum_{m=1}^{N_I} \exp(\cos(\hat{\mathbf{H}}_{I,i}^{t,n}, \tilde{\mathbf{H}}_{I,j}^{t,m})/\tau)}, \tag{4}$$

where $\cos(\cdot,\cdot)$ denotes cosine similarity.

Overall Objective. The overall optimization objective of ProTrans is $\mathcal{L} = \mathcal{L}_{\text{state}}^{\text{MoD}} + \mathcal{L}_{\text{trans}}^{\text{MoD}} + \lambda_r \mathcal{L}_{\text{recon}}$, where λ_r adjusts the constraint from reconstruction.

3 Experiments and Results

Pretraining Dataset. We constructed the pretraining dataset using MIMIC-CXR-JPG [10] and Chest ImaGenome dataset [21], collecting longitudinal CXR

images paired with corresponding radiology reports from MIMIC-CXR-JPG and disease progression annotations between consecutive studies from Chest ImaGenome. In total, we utilized 98,940 bi-temporal image–report pairs for VLP, excluding samples that overlapped with CXR-MS-T dataset [3, 4].

Implementation Details We adopted ViT-B/16 as the visual encoder and BioClinicalBERT [1] as the text encoder [23]. AdamW was used with an initial learning rate 5×10^{-5} and a weight decay 0.05. The temperature τ was set to 0.07, the momentum 0.995, trade-off coefficients α 0.8, and λ_r 0.5. The progression encoder f_P consists of 3 blocks, each including both spatial and temporal attention layers with 12 attention heads and an embedding dimension of 768.

3.1 Downstream Adaptation Setup

To evaluate the effectiveness of the learned progression representations, we performed experiments with the following two downstream tasks.

Task1: Disease Progression Classification. We conducted this task on the MS-CXR-T dataset, which comprises 1,326 paired CXR images covering five findings—Consolidation, Edema, Pleural Effusion, Pneumonia, and Pneumothorax—each annotated with a three-way progression label (i.e., improving, stable or worsening). Following [23], we used the visual backbone to extract progression representations $\tilde{\mathbf{P}}_I^{\rightarrow}$ and trained a SVM [6] classifier for prediction under a 10-fold cross-validation protocol. This evaluated the discriminative ability of the learned representations without additional backbone fine-tuning.

To further assess cross-modal semantic alignment, we adopted a classification strategy with text prototyping. Specifically, each progression category is formulated as a text prototype in the form of "[finding] [progression]", which is encoded into text embeddings. Classification was performed by computing similarities between the progression representation and prototype embeddings, assigning the label with the highest similarity score.

Task2: Progression Captioning. This task aims to generate natural language descriptions that characterize changes across CXR pairs. We evaluated it on the ICG dataset [15], which contains 11,439 paired CXRs with the corresponding radiological descriptions of visual differences. For each CXR pair, the learned progression representation was mapped by a trainable projector into the embedding space of a frozen medical LLM, BioMistral-7B [11], and used as visual-prefix tokens for generating progression descriptions. During fine-tuning, only the projector was optimized. We utilized the official splits with 10,679 samples for fine-tuning and 760 for testing. Performance was evaluated using both lexical (ROUGE-L and BERTScore) and clinical metrics (RadGraph-F1, CheXbert, GREEN and Temporal-F1) [22].

3.2 Results

Results on disease progression classification. As shown in Table 1, our method consistently outperforms prior feature fusion-based methods (BioViL-T, MedST, TempA-VLP, Coca-CXR) and the discrepancy-based method (Diff-

Table 1. Disease progression classification results (accuracy, %) on the MS-CXR-T dataset. Gray indicates results cited from the original papers as the implementations are not publicly available. We report mean±std. over 3 random seeds. The best results are in bold. †Coca-CXR used next-token prediction for classification.

Classifier	Method	Consolidation	Edema	Pl. Effusion	Pneumonia	Pneumothorax	Avg.
SVM	BioViL-T [4]	56.93±1.77	61.55±0.95	53.94±0.89	67.24±0.20	**55.46±0.01**	59.02±0.34
	MedST [23]	60.57±1.18	67.35±0.32	58.47±1.50	65.00±0.34	54.18±0.81	61.12±0.34
	Diff-RRG [29]	47.94±2.71	50.50±2.16	48.66±0.00	67.10±0.00	**55.46±0.01**	53.93±0.77
	Ours	**61.54±0.23**	**69.53±0.64**	**61.81±0.39**	**69.35±0.84**	**55.46±0.01**	**63.54±0.29**
Text prototype	BioViL-T [4]	51.99±1.21	57.39±0.77	50.24±0.00	66.39±0.34	55.50±0.22	56.30±0.43
	MedST [23]	61.06±0.24	65.29±0.35	63.03±0.30	67.79±0.52	56.14±1.33	62.66±0.42
	Diff-RRG [29]	55.15±0.74	51.12±0.00	48.79±0.73	61.04±0.21	50.71±0.71	53.36±0.39
	TempA-VLP [27]	65.2	77.1	59.4	73.4	43.1	64.8
	Coca-CXR† [8]	69.6	71.8	68.1	56.4	59.3	65.0
	Ours	**63.86±0.70**	**69.68±0.77**	**66.42±0.11**	**69.05±0.79**	**60.06±0.80**	**65.81±0.38**

Table 2. Progression captioning results on the ICG dataset.

Method	ROUGE-L	BERTScore	CheXbert	Green	Radgraph-F1	Temporal-F1
LLava-Med [12]	0.167	0.381	0.166	0.031	0.053	0.119
LLava-Rad [7]	0.184	0.442	0.383	0.125	0.087	0.136
Maira [2]	0.167	0.415	0.075	0.077	0.062	0.118
Libra [30]	0.190	0.428	0.366	0.141	0.081	0.145
Diff-RRG [29]	0.153	0.407	0.221	0.101	0.059	0.110
Ours	**0.259**	**0.535**	**0.386**	**0.230**	**0.126**	**0.238**

RRG). Using the SVM classifier, our method obtains the highest average accuracy of 63.54%, indicating that the learned progression representations exhibit stronger linear separability. This improvement can be attributed to the explicit modeling of directional progression in the semantic space, where tighter alignment is enforced between visual progression features and direction-aware progression descriptions. As a result, the advantage of our method becomes even more pronounced in the text prototype setting. To gain deeper insights into spatial semantic correspondence, we visualize similarity maps between the learned progression representations and their associated text prototypes. As shown in Fig. 2(a), similarity responses are mainly concentrated in lung regions. When projected back onto the prior and current CXR images, these highlighted areas correspond to clinically relevant locations associated with edema, indicating that the model effectively localizes progression-related regions.

Results on Progression captioning. Table 2 presents progression captioning results on the ICG dataset, compared with several competing medical LLMs. Our method achieves the best performance across both lexical and clinical evaluation metrics, with large margins, about 39% in Green and Temporal-F1, demonstrating stronger clinical fidelity and temporal awareness. Although LLaVA-Med, LLaVA-Rad, and Maira can process multi-image inputs, they are not explicitly designed to model longitudinal relationships, thus struggling to capture subtle bi-temporal differences. Libra introduces a dedicated module to fuse multi-scale contextual information across images while Diff-RRG focuses on patch-level difference modeling. However, both methods mainly capture low-

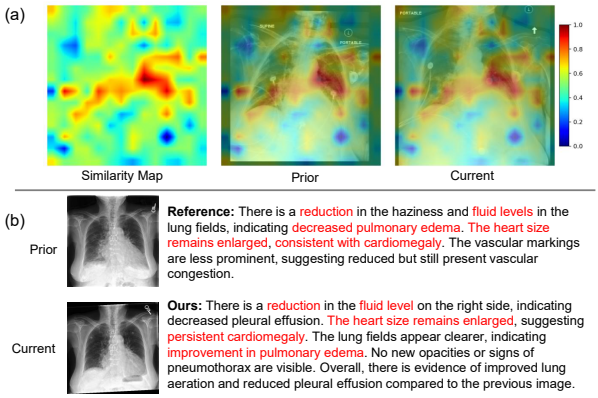

Fig. 2. (a) Similarity map between the learned progression features and the text prototype "edema improved", which is projected back onto the prior and current CXR images, and (b) qualitative comparison of generated progression findings.

level visual variations and fail to establish semantic grounding with progression. In contrast, our method learns progression representations that encode semantic transitions between clinical states, thus providing more informative temporal cues to the language model. Fig. 2(b) illustrates that our method accurately captures and describes clinically meaningful differences between bi-temporal images.

Table 3. Ablation study on different components of our proposed method.

Classifier	Method	Consolidation	Edema	Pl. Effusion	Pneumonia	Pneumothorax	Avg.
SVM	w/o $\mathcal{L}_{\text{state}}^{\text{MoD}}$	56.41±0.21	62.56±0.36	61.15±0.09	69.21±0.89	53.87±0.21	60.64±0.00
	w/o $\mathcal{L}_{\text{trans}}^{\text{MoD}}$	44.13±2.04	52.89±0.65	48.66±0.00	67.10±0.00	**55.46±0.01**	53.65±0.54
	w/o $\mathcal{L}_{\text{recon}}$	59.43±1.27	64.80±2.02	**63.09±0.30**	66.82±0.51	54.36±0.21	61.70±0.46
	w/o bidirection	**62.69±0.71**	66.15±0.27	59.20±0.45	66.41±0.98	**55.46±0.01**	61.98±0.15
	Ours	61.54±0.23	**69.53±0.64**	61.81±0.39	**69.35±0.84**	**55.46±0.01**	**63.54±0.29**
Text prototype	w/o pretraining	43.91±0.25	69.62±1.00	64.40±0.64	68.02±1.52	55.19±0.00	60.23±0.32
	w/ pretraining	**63.86±0.70**	**69.68±0.77**	**66.42±0.11**	**69.05±0.79**	**60.06±0.80**	**65.81±0.38**

Ablation Study. Table 3 presents the ablation results of our framework. Removing any key module signficantly compromises the overall model performance, demonstrating their complementary roles in learning effective progression representations. In particular, eliminating the transition alignment loss $\lambda_{trans}^{\text{MoD}}$ leads to the most significant drop in performance, suggesting that providing explicit semantic supervision for visual progression feature learning is critical for accurately modeling longitudinal changes. Furthermore, when the progression pretraining stage is removed and the downstream text prototype model is trained directly, performance consistently deteriorates. This degradation underscores the importance of the proposed pretraining stage in providing a strong temporal semantic foundation for longitudinal downstream tasks.

4 Conclusion

In this paper, we have proposed ProTrans, a novel VLP framework for progression representation learning that formulates disease progression as a directional semantic transition between bi-temporal clinical states. By leveraging radiology reports to guide state grounding and transition alignment, together with reversed temporal learning to enforce trajectory consistency, ProTrans effectively captures direction-aware and semantically grounded progression patterns. Extensive experiments on progression classification and captioning tasks demonstrate the potential of our method in learning robust progression semantics for CXRs.
Limitations. One limitation of ProTrans is that reversed temporal supervision is mainly suitable for clinically reversible progression patterns. For irreversible or non-regressing findings, such as calcification, fibrosis, and scarring, flipped semantic labels may be clinically invalid and thus require careful phrase-level filtering. Future work will explore more clinically grounded transition modeling for such findings.

Acknowledgments. This work was partly supported by the National Science Foundation through the Faculty Early Career Development Program (CAREER) under Award 2046708.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alsentzer, E., Murphy, J., Boag, W., Weng, W.H., Jindi, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. In: Proceedings of the 2nd clinical natural language processing workshop. pp. 72–78 (2019)
2. Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., Ilse, M., Pérez-García, F., Salvatelli, V., Sharma, H., et al.: Maira-2: Grounded radiology report generation. arXiv preprint arXiv:2406.04449 (2024)
3. Bannur, S., Hyland, S., Liu, Q., Pérez-García, F., Ilse, M., de Castro, D.C., Boecking, B., Sharma, H., Bouzid, K., Schwaighofer, A., et al.: MSCXR-T: Learning to exploit temporal structure for biomedical vision-language processing (2023)
4. Bannur, S., Hyland, S., Liu, Q., Perez-Garcia, F., Ilse, M., Castro, D.C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., et al.: Learning to exploit temporal structure for biomedical vision-language processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15016–15027 (2023)
5. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., 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)
6. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory. pp. 144–152 (1992)

7. Chaves, J.M.Z., Huang, S.C., Xu, Y., Xu, H., Usuyama, N., Zhang, S., Wang, F., Xie, Y., Khademi, M., Yang, Z., et al.: Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation. arXiv preprint arXiv:2403.08002 (2024)
8. Chen, Y., Xu, S., Sellergren, A., Matias, Y., Hassidim, A., Shetty, S., Golden, D., Yuille, A.L., Yang, L.: CoCa-CXR: Contrastive captioners learn strong temporal structures for chest x-ray vision-language understanding. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 78–88. Springer (2025)
9. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3942–3951 (2021)
10. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
11. Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.A., Rouvier, M., Dufour, R.: Biomistral: A collection of open-source pretrained large language models for medical domains. arXiv preprint arXiv:2402.10373 (2024)
12. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
13. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021)
14. Liu, K., Ma, Z., Fang, Z., Li, Y., Xie, K., Miao, Q.: PriorRG: Prior-guided contrastive pre-training and coarse-to-fine decoding for chest x-ray report generation. arXiv preprint arXiv:2508.05353 (2025)
15. Ma, C., Ji, Y., Ye, J., Zhang, L., Chen, Y., Li, T., Li, M., He, J., Shan, H.: Towards interpretable counterfactual generation via multimodal autoregression. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 611–620. Springer (2025)
16. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
17. Song, S., Tang, H., Yang, H., Li, X.: DDaTR: Dynamic difference-aware temporal residual network for longitudinal radiology report generation. arXiv preprint arXiv:2505.03401 (2025)
18. Wang, F., Du, S., Yu, L.: Hergen: Elevating radiology report generation with longitudinal data. In: European Conference on Computer Vision. pp. 183–200. Springer (2024)
19. Wang, F., Zhou, Y., Wang, S., Vardhanabhuti, V., Yu, L.: Multi-granularity cross-modal alignment for generalized medical visual representation learning. Advances in neural information processing systems **35**, 33536–33549 (2022)
20. Wang, Y., Duan, J., Li, X., Wang, J.: CheXLearner: Text-guided fine-grained representation learning for progression detection. arXiv preprint arXiv:2505.06903 (2025)
21. Wu, J.T., Agu, N.N., Lourentzou, I., Sharma, A., Paguio, J.A., Yao, J.S., Dee, E.C., Mitchell, W., Kashyap, S., Giovannini, A., et al.: Chest imagenome dataset for clinical reasoning. arXiv preprint arXiv:2108.00316 (2021)

22. Xu, J., Zhang, X., Abderezaei, J., Bauml, J., Boodoo, R., Haghighi, F., Ganjizadeh, A., Brattain, E., Van Veen, D., Meng, Z., et al.: Radeval: A framework for radiology text evaluation, 2025. URL https://arxiv. org/abs/2509.18030
23. Yang, J., Su, B., Zhao, W.X., Wen, J.R.: Unlocking the power of spatial and temporal information in medical multimodal pre-training. arXiv preprint arXiv:2405.19654 (2024)
24. Yang, Z., Wang, G., Hendler, J., Kalra, M.K., Yan, P.: X-WIN: Building chest radiograph world model via predictive sensing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6920–6930 (2026)
25. Yang, Z., Xu, X., Zhang, J., Wang, G., Kalra, M.K., Yan, P.: Chest x-ray foundation model with global and local representations integration. arXiv preprint arXiv:2502.05142 (2025)
26. Yang, Z., Zhang, J., Wang, G., Kalra, M.K., Yan, P.: Cardiovascular disease detection from multi-view chest x-rays with bi-mamba. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 134–144. Springer (2024)
27. Yang, Z., Shen, L.: TempA-VLP: Temporal-aware vision-language pretraining for longitudinal exploration in chest x-ray image. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4625–4634. IEEE (2025)
28. Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C.: Filip: Fine-grained interactive language-image pre-training. arXiv preprint arXiv:2111.07783 (2021)
29. Yun, H., Maeng, J., Kang, E., Suk, H.I.: Diff-RRG: Longitudinal disease-wise patch difference as guidance for llm-based radiology report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 152–161. Springer (2025)
30. Zhang, X., Meng, Z., Lever, J., Ho, E.S.: Libra: Leveraging temporal images for biomedical radiology analysis. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 17275–17303 (2025)
31. Zhou, S., Li, Y., Liu, Y., Liu, L., Wang, L., Zhou, L.: A review of longitudinal radiology report generation: Dataset composition, methods, and performance evaluation. arXiv preprint arXiv:2510.12444 (2025)