A Disease-Aware Dual-Stage Framework for Chest X-ray Report Generation

Puzhen Wu[1], **Hexin Dong**[1], **Yi Lin**[1], **Yihao Ding**[*2], **Yifan Peng**[*1]

[1]Population Health Sciences, Weill Cornell Medicine, New York, NY, USA
[2]School of Physics, Mathematics and Computing, University of Western Australia, Crawley, Australia
{puw4002, hed4006, yil4033, yip4002}@med.cornell.edu, yihao.ding@uwa.edu.au

Abstract

Radiology report generation from chest X-rays is an important task in artificial intelligence with the potential to greatly reduce radiologists' workload and shorten patient wait times. Despite recent advances, existing approaches often lack sufficient disease-awareness in visual representations and adequate vision-language alignment to meet the specialized requirements of medical image analysis. As a result, these models usually overlook critical pathological features on chest X-rays and struggle to generate clinically accurate reports. To address these limitations, we propose a novel dual-stage disease-aware framework for chest X-ray report generation. In Stage 1, our model learns Disease-Aware Semantic Tokens (DASTs) corresponding to specific pathology categories through cross-attention mechanisms and multi-label classification, while simultaneously aligning vision and language representations via contrastive learning. In Stage 2, we introduce a Disease-Visual Attention Fusion (DVAF) module to integrate disease-aware representations with visual features, along with a Dual-Modal Similarity Retrieval (DMSR) mechanism that combines visual and disease-specific similarities to retrieve relevant exemplars, providing contextual guidance during report generation. Extensive experiments on benchmark datasets (i.e., CheXpert Plus, IU X-ray, and MIMIC-CXR) demonstrate that our disease-aware framework achieves state-of-the-art performance in chest X-ray report generation, with significant improvements in clinical accuracy and linguistic quality.

Code — https://github.com/bionlplab/2026_AAAI_CXR_report_generation

Introduction

Generating chest X-ray reports is a critical yet highly complex challenge in medical AI. The goal is to automatically generate comprehensive diagnostic reports from chest radiographs, thereby assisting radiologists by potentially reducing their workload and mitigating workforce shortages. Achieving this goal can also accelerate patient assessment and improve workflow efficiency.

Most state-of-the-art chest X-ray report generation models follow an encoder–decoder architecture. In this framework, a vision encoder processes the chest X-ray image. At the same time, a language decoder generates the corresponding report. Notably, the R2Gen model introduced a memory-driven Transformer decoder to better capture and retain visual features during the report generation process (Chen et al. 2020). Building on this foundation, subsequent works have sought to incorporate domain knowledge and enhance feature representations. For example, the DCL approach by (Li et al. 2023) integrated a dynamic graph of medical entities coupled with contrastive learning to emphasize abnormal findings, thereby demonstrating the benefit of combining structured knowledge with visual encoding. Meanwhile, the adoption of specialized pre-trained language models like Bio-ClinicalBERT (Alsentzer et al. 2019) has enriched the inclusion of medical context during text generation.

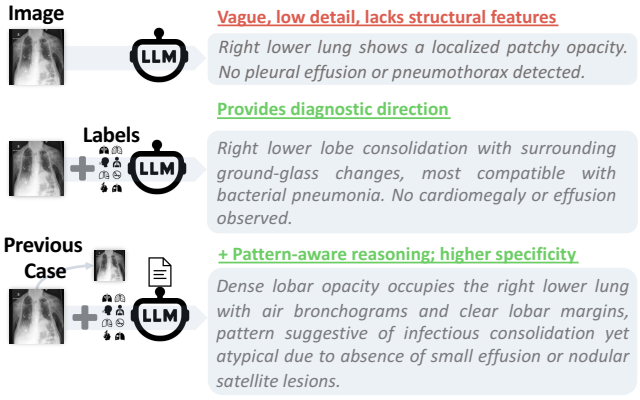

Figure 1: Disease-Aware Semantic Tokens and retrieved examples make image-only LLM reports highly specific.

The rapid advancement of large language models (LLMs) has further propelled progress in this field. For example, R2GenGPT (Wang et al. 2023b) aligned extracted visual features with the embedding space of a frozen LLM, effectively leveraging a pre-trained GPT-style model to generate more fluent and clinically accurate reports. Advancements in vision encoding have also played a significant role. Transformer-based encoders, such as ViT (Dosovitskiy et al. 2021), have been employed to capture global image features, while self-supervised learning methods, such as masked autoencoders (MAE) (He et al. 2022), have improved image representations by pre-training on large collections of unla-

[*]These authors contributed equally.

beled chest X-rays (Wang et al. 2025b). Moreover, multimodal representation learning approaches (e.g., CLIP (Radford et al. 2021), which couples images and text with natural language supervision) have inspired techniques to better align visual and textual features. Collectively, these developments have substantially advanced the state of the art in automatic chest X-ray report generation (Xu et al. 2025).

Despite substantial progress, several gaps remain unaddressed. Current models often struggle to capture fine-grained, disease-specific details that are essential in clinical practice. Radiology reports are expected to comprehensively describe all critical findings (e.g., precise locations and severities of lesions). However, standard encoder–decoder models often overlook less prominent findings and tend to generate only high-level impressions. Many existing approaches also lack explicit mechanisms to ensure that each clinically important disease present in the image is both recognized and appropriately described. This can lead to the omission of key conditions or the inclusion of irrelevant observations. Moreover, most contemporary frameworks generate the report solely based on features extracted from the input image and learned language patterns, without utilizing information from similar prior cases that might provide valuable contextual cues. This limitation reduces the model's ability to reinforce its predictions with evidence, particularly for rare conditions or atypical presentations.

In this paper, we propose a novel Disease-Aware Dual-Stage Framework for chest X-ray report generation that directly addresses the issues above. The key innovation lies in incorporating an intermediate semantic understanding stage centered on disease-specific analysis before producing the report. In Stage 1, our model extracts high-level Disease-Aware Semantic Tokens (DASTs) from the image. These DASTs consist of a set of discrete tokens that encode the presence and characteristics of diseases or abnormal findings, thereby creating an explicit semantic bridge between visual features and medical terminology.

In Stage 2, we leverage a Disease-Visual Attention Fusion (DVAF) module to generate the textual report. This module fuses image features with disease-specific semantic tokens via an attention mechanism. The DVAF module ensures that the decoder attends to the pertinent visual regions and their corresponding disease semantics when constructing each sentence of the report. This approach substantially improves the completeness and accuracy of disease descriptions. Additionally, our framework incorporates a Dual-Modal Similarity Retrieval (DMSR) mechanism that retrieves relevant previous cases by jointly considering image and semantic similarity derived from reports. By referencing image-report pairs similar to the current case, the model gains contextual guidance, helping it avoid common errors and include clinically relevant details that appear in analogous cases. As illustrated in Figure 1, augmenting an image-only LLM with DASTs and retrieved examples progressively refines the generated report.

We developed and validated our models using three official splits of CheXpert Plus (Chambon et al. 2024), IU X-Ray (Demner-Fushman et al. 2016), and MIMIC-CXR (Johnson et al. 2019). Extensive experiments demonstrated that our model achieves state-of-the-art performance across these benchmarks.

To sum up, our main contributions are as follows: (1) We introduce Disease-Aware Semantic Tokens (DASTs) that encode disease-specific features, guiding the model toward clinically important abnormalities; (2) We design a Disease-Visual Attention Fusion (DVAF) module that aligns disease tokens with visual regions, enabling a precise description of findings; (3) We develop a Dual-Modal Similarity Retrieval (DMSR) mechanism that retrieves relevant image–report exemplars using visual and semantic similarity, enhancing completeness and factual accuracy; and (4) We conduct comprehensive evaluations on CheXpert Plus, IU X-Ray, and MIMIC-CXR, achieving state-of-the-art performance across both textual and clinical metrics.

Related Work

We briefly review three areas of related work: (1) chest X-ray report generation, (2) large-scale pre-trained models for medical vision–language tasks, and (3) state-space models as efficient backbones.

X-ray Medical Report Generation. Early research on automatic report generation for chest X-rays introduced various strategies to improve model performance and coherence. DCL proposed a vision-based dynamic graph mechanism, leveraging external knowledge to strengthen image representations (Li et al. 2023). RGRG employed an object detector to localize salient lesion regions and generated textual descriptions for each area that were then combined into a complete report (Tanida et al. 2023). HERGen treated previous reports of the same patient as a temporally ordered sequence, capturing longitudinal information across visits (Xue et al. 2024a). More recently, R2GenGPT replaced the traditional decoder with an LLM to improve the output quality (Wang et al. 2023b). These innovations address domain knowledge, temporal context, and the capabilities of modern LLMs. CXPMRG-Bench (Wang et al. 2025b) provides the latest controlled experimental results currently through systematic pre-training and benchmark evaluation.

Pre-trained Large Models. With the success of large-scale pre-training, several works have adopted similar ideas for report generation. Wang et al. introduced a context-aware masked autoencoder for high-resolution chest X-rays, enabling richer visual features to be extracted from unlabeled images before report generation (Wang et al. 2025a). CXR-CLIP enlarged the training corpus by synthesizing additional image–text pairs and applied a contrastive loss to align images and reports in a shared space (You et al. 2023). PTUnifier unifies modalities through learnable visual and textual prompt pools, integrating fusion and dual-encoder paradigms (Chen et al. 2023). Although promising, these methods typically employ Transformer-based vision encoders (Wu et al. 2023) with quadratic complexity (Dosovitskiy et al. 2021) and single-stage pre-training, limiting the use of abundant unpaired X-ray data.

State-Space Models as Efficient Backbones. State-space models have emerged as efficient alternatives to Transform-

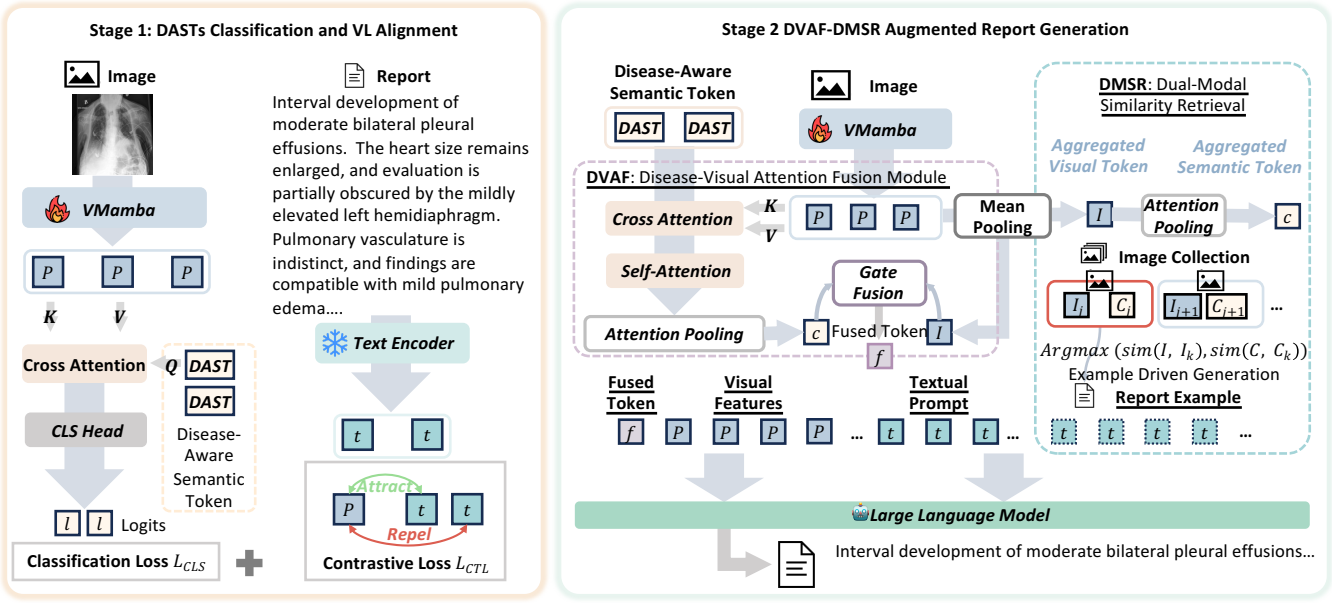

Figure 2: Overview of the proposed two-stage framework. Stage 1 jointly trains a VMamba image encoder and a text encoder to learn disease-aware semantic tokens (DASTs), aligning visual and textual features through classification and contrastive losses. Stage 2 fuses the learned DASTs with visual features, retrieves a similar study via dual-modal similarity retrieval, and feeds these cues into a large language model to generate the final radiology report.

ers for sequence modeling because they scale linearly with sequence length. Mamba introduced selective state transitions and a parallelizable recurrence, achieving strong performance with lower computational cost (Gu and Dao 2024). Vision Mamba (Vim) (Zhu et al. 2024) adapted this design to images and provided a global receptive field with linear complexity (Liu et al. 2024b). VMamba extended Vim and has been incorporated into the R2GenCSR generator as a more efficient backbone (Wang et al. 2024b). Inspired by these successes (Wu et al. 2025), we employ a Mamba-based encoder in our dual-stage disease-aware framework, benefiting from faster computation and effective long-range dependency modeling on high-resolution medical images.

Method

Overview

Our framework integrates three key components (Figure 2): a lightweight VMamba encoder for image interpretation (Liu et al. 2024b), a pretrained Bio-ClinicalBERT for medical language modeling (Alsentzer et al. 2019), and an LLM for radiology report generation. This architecture is specifically designed to combine efficient visual representation learning with rich medical language understanding and generative capabilities. To fully exploit the proposed architecture, we implement a two-stage training strategy. In Stage 1, we introduce 14 DASTs, each corresponding to a specific pathology category. Through cross-attention interactions with visual patch tokens and subsequent multi-label classification tasks, each DAST is trained to encode disease-specific semantic information relevant to its corresponding pathology. We also employ a contrastive loss to more closely align representations learned from visual and textual modalities. In Stage 2, we freeze the LLM's parameters and employ a DVAF module to connect the learned disease-aware representations with visual features. Additionally, we introduce a novel DMSR mechanism that jointly considers both visual similarity and disease-specific similarity to retrieve the most relevant exemplar from the training set. This retrieved exemplar provides precise contextual guidance to the report generation process. The final fused visual-disease representations, along with the retrieved exemplar, are then fed into the LLM to generate detailed radiology reports.

Stage 1: Disease-Aware Semantic Tokens (DASTs) Classification and Vision Language Alignment

Visual Encoder. To efficiently extract both local anatomical features and global contextual information from chest X-ray images, while maintaining linear computational complexity, we adopt the VMamba architecture (Liu et al. 2024b). VMamba leverages selective state-space models to capture long-range dependencies with $\mathcal{O}(N)$ complexity, offering an advantage over the quadratic $\mathcal{O}(N^2)$ computational cost of conventional Transformers. Given an input image $I \in \mathbb{R}^{H \times W}$, we first divide it into N non-overlapping patches and obtain the initial patch embeddings via

$$\mathbf{x} = \text{PatchEmbed}(I), \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^{N \times C}$ denotes the sequence of raw patch tokens with C feature dimensions. These tokens are then fed into L VMamba blocks to model long-range interactions:

$$\mathbf{z} = \text{VMamba}(\mathbf{x}), \tag{2}$$

where $\mathbf{z} \in \mathbb{R}^{N \times C}$ represents the refined visual patch tokens enriched with global contextual information.

DAST Classification. To complement the visual representation, we incorporate $D = 14$ learnable DASTs, denoted as $\{\mathbf{t}_d\}_{d=1}^{D}$, each corresponding to a pathology category defined in the CheXpert ontology (Chambon et al. 2024). Each DAST $\mathbf{t}_d \in \mathbb{R}^C$ is randomly initialized and subsequently optimized during training to encode disease-specific semantics that inform downstream report generation. To fuse visual and semantic information, we apply a cross-attention mechanism in which the DASTs serve as queries and visual patch tokens act as keys and values. This design allows each DAST to selectively attend to relevant visual regions associated with its pathology while learning disease-specific semantic representations. The resulting refined disease features are then passed through independent classification heads to enable multi-label pathology prediction within the image. The learned DASTs capture both visual evidence and medical semantics. The multi-label classification objective serves as an auxiliary task that encourages DASTs to encode meaningful pathological information while simultaneously guiding the visual encoder to learn discriminative and disease-relevant visual features. This approach ultimately improves the semantic richness and medical accuracy of generated reports in subsequent stages of the framework.

Training Objective. To optimize both disease classification and multimodal alignment, we design a dual loss function that combines a multi-label classification loss with a contrastive alignment loss. This joint training strategy ensures accurate pathology detection and robust vision-language correspondences, which are essential for high-quality report generation in Stage 2.

The classification loss $\mathcal{L}_{\text{CLS}}$ is computed using binary cross-entropy applied to the outputs of the disease-specific classification heads. This auxiliary task encourages the DASTs to encode clinically relevant semantic information, thereby enhancing the medical coherence and accuracy of the generated reports.

Furthermore, we implement contrastive learning to align visual and textual modalities. This alignment is crucial for bridging the semantic gap between visual pathological findings and their textual descriptions, thereby enabling more effective cross-modal understanding in downstream report generation. Specifically, visual features are extracted from the VMamba encoder via global average pooling over the patch tokens, while textual features are obtained from a frozen Bio-ClinicalBERT encoder. Denote I_i as the i-th image and R_j as j-th report; the contrastive loss is defined as

$$\mathcal{L}_{\text{CTL}} = \text{Similarity}\left(\text{Visual}(I_i), \text{Textual}(R_j)\right), \quad (3)$$

where the similarity is measured by cosine similarity between paired and unpaired samples. The final training objective is formulated as a combination of two loss components

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CLS}} + \mathcal{L}_{\text{CTL}}. \quad (4)$$

This joint optimization ensures that the learned visual representations are both diagnostically informative and semantically aligned with medical language, providing a strong foundation for subsequent report generation.

Stage 2: Augmented Report Generation with Disease-Visual Attention Fusion (DVAF) and Dual-Modal Similarity Retrieval (DMSR)

In Stage 2, we aim to generate detailed radiology reports by leveraging the rich visual features and DASTs learned in Stage 1. We freeze the LLM's parameters and employ DVAF to fuse disease-aware tokens with image features. Simultaneously, DMSR is used to retrieve the most similar case from the training set. The fused features and the retrieved exemplar, are then fed into the LLM to generate the report.

DVAF. We begin with the visual patch tokens $\mathbf{z} \in \mathbb{R}^{N \times C}$ and the set of refined DASTs $\{\mathbf{t}_d\}_{d=1}^{14}$, both obtained from Stage 1. To transform this disease-specific information into a representation suitable for report generation, we propose the DVAF module to aggregate the DASTs through a cascade of operations: (1) cross-attention with visual patch tokens to inject spatial context, (2) self-attention to model inter-disease relationships, and (3) attention pooling to yield a single, unified disease representation

$$\mathbf{p} = \text{AttnPool}\left(\text{SelfAttn}\left(\text{CrossAttn}\left(\{\mathbf{t}_d\}, \mathbf{z}, \mathbf{z}\right)\right)\right). \quad (5)$$

To incorporate global image context, we average the patch tokens $\bar{\mathbf{z}}$ and fuse it with $\mathbf{p}$ using a gating mechanism

$$\mathbf{f} = \mathbf{W}_{\text{gate}}[\mathbf{p}; \bar{\mathbf{z}}]. \quad (6)$$

The resulting fusion token $\mathbf{f}$ encapsulates both visual and pathological information. This token is then appended to the original patch sequence to construct the final LLM input:

$$\mathbf{V} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N, \mathbf{f}] \in \mathbb{R}^{(N+1) \times C}. \quad (7)$$

We then apply a trainable linear projection followed by layer normalization to align with the LLM's latent space.

DMSR. To improve fluency and completeness of the generated reports, we introduce DMSR. For each input, we construct a composite query vector by combining the mean visual features $\bar{\mathbf{z}}$ and the disease classification logits $\mathbf{l}$. We then compute similarity scores between this query vector and entries stored in the database:

$$s_k = \text{Similarity}(\bar{\mathbf{z}}, \bar{\mathbf{z}}_k) + \lambda \cdot \text{Similarity}(\mathbf{l}, \mathbf{l}_k), \quad (8)$$

where λ balances visual and disease similarity. By leveraging both modalities, the DMSR mechanism identifies the most relevant exemplar from the database. The corresponding report R_{ret} is then retrieved and inserted as an in-context example in the prompt provided to the LLM.

Report Generation and Optimization. The final input to the LLM consists of the retrieved report prompt R_{ret}, the projected visual representation $\tilde{\mathbf{V}}$, and the target report tokens $\mathbf{y}$. The model is trained using a standard language modeling loss

$$\mathcal{L}_{\text{LM}} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, \tilde{\mathbf{V}}, R_{\text{ret}}), \quad (9)$$

where $y_{<t}$ is the previously generated tokens, and T is the report length. Notably, during Stage 2 training, we freeze all components from Stage 1 and only optimize the projection layers and normalization parameters. This approach ensures efficient adaptation to the report generation task while preserving the learned representations from Stage 1.

Algorithm	Venue	Encoder	Decoder	BLEU4	ROUGE-L	CIDEr	Precision	Recall	F1
R2GenRL (Qin and Song 2022)	ACL'22	Transformer	Transformer	0.035	0.186	0.012	0.193	0.229	0.196
XProNet (Wang, Bhalerao, and He 2022)	ECCV'22	Transformer	Transformer	0.100	0.265	0.121	0.314	0.247	0.259
MSAT (Wang et al. 2022)	MICCAI'22	ViT-B/16	Transformer	0.036	0.156	0.018	0.044	0.142	0.057
ORGan (Hou et al. 2023)	ACL'23	CNN	Transformer	0.086	0.261	0.107	0.288	0.287	0.277
M2KT (Yang et al. 2023)	MIA'21	CNN	Transformer	0.078	0.247	0.077	0.044	0.142	0.058
TIMER (Wu, Huang, and Huang 2023)	CHIL'23	Transformer	Transformer	0.083	0.254	0.104	0.345	0.238	0.234
R2Gen (Chen et al. 2020)	EMNLP'20	Transformer	Transformer	0.081	0.246	0.077	0.318	0.200	0.181
R2GenCMN (Chen et al. 2021)	ACL'21	Transformer	Transformer	0.087	0.256	0.102	0.329	0.241	0.231
Zhu et al. (Zhu et al. 2023)	MICCAI'23	Transformer	Transformer	0.074	0.235	0.078	0.217	0.308	0.205
CAMANet (Wang et al. 2024a)	JBHI'23	Swin-Former	Transformer	0.083	0.249	0.090	0.328	0.224	0.216
ASGMD (Xue et al. 2024b)	ESWA'24	ResNet-101	Transformer	0.063	0.220	0.044	0.146	0.108	0.055
Token-Mixer (Yang et al. 2024)	TMI'23	ResNet-50	Transformer	0.091	0.261	0.098	0.309	0.270	0.288
PromptMRG (Jin et al. 2024)	AAAI'24	ResNet-101	BERT	0.095	0.222	0.044	0.258	0.265	0.281
R2GenGPT (Wang et al. 2023b)	Meta-Rad'23	Swin-Former	Llama2	0.101	0.266	0.123	0.315	0.244	0.260
WCL (Yan et al. 2021)	EMNLP'21	Transformer	Transformer	0.084	0.253	0.103	0.335	0.259	0.256
VLCI (Chen et al. 2025)	TIP'25	Transformer	Transformer	0.080	0.247	0.072	0.341	0.175	0.163
MambaXray (Wang et al. 2025b)	CVPR'25	Vim	Llama2	0.112	0.276	0.139	0.377	0.319	0.335
Ours	–	VMamba	Phi-4	**0.133**	**0.291**	**0.227**	**0.394**	**0.356**	**0.361**

Table 1: Comparison with state-of-the-art methods on the CheXpert Plus dataset.

Dataset	Training	Val	Testing
CheXpert Plus	40,463	5,780	11,562
IU X-Ray	2,069	296	590
MIMIC-CXR	270,790	2,130	3,858

Table 2: Dataset statistics.

Experimental Setup

Benchmark Datasets

We evaluated our approach on three widely used open-access chest X-ray datasets: CheXpert Plus (Chambon et al. 2024), IU X-Ray (Demner-Fushman et al. 2016), and MIMIC-CXR (Johnson et al. 2019). Each dataset consists of paired chest X-rays and their corresponding reports, as shown in Table 2. Following established evaluation protocols (Wang et al. 2025b), we adopted the same dataset partitioning strategy as in previous works (Table 2).

CheXpert Plus includes 223,228 chest X-rays, available in both DICOM and PNG formats, paired with 187,711 de-identified radiology reports, each parsed into 11 structured sections. It also encompasses demographic information from 64,725 patients, 14 chest pathology labels, and RadGraph annotations (Jain et al. 2021). The **IU X-Ray** dataset comprises 7,470 chest X-ray images paired with 3,955 radiology reports, with each report corresponding to either a frontal view or a combination of frontal and lateral view examinations. The reports are structured into four sections: Indication, Comparison, Findings, and Impression. Following common practice in prior work (Chen et al. 2020), we concatenated the Findings and Impression sections of each report to form the target text. **MIMIC-CXR** consists of chest X-rays and radiology reports from the Beth Israel Deaconess Medical Center Emergency Department in Boston, Massachusetts. Covering the period from 2011 to 2016, it contains 377,110 images and 227,835 reports from 65,379 patients.

Implementation Details

For Stage 1 training, our training data consists of 480,000 image-text pairs compiled from the training datasets of MIMIC-CXR, CheXpert Plus, and IU X-ray. For all three datasets, the 14-category pathology labels are automatically extracted using the CheXpert labeler (Irvin et al. 2019), providing a unified supervision signal for Stage 1 classification. The VMamba encoder is initialized with weights pretrained on ImageNet. During this stage, only the visual encoder is set as trainable, and input images are standardized to a resolution of 224×224 pixels and converted to greyscale. For Stage 2 fine-tuning, we evaluated model performance across three datasets with dataset-specific configurations. Our framework is implemented using PyTorch (Paszke et al. 2019). We utilize the AdamW optimizer (Loshchilov and Hutter 2019) with a learning rate of 1×10^{-4}. Training incorporates linear warm-up for the first 500 steps, followed by cosine decay scheduling (Loshchilov and Hutter 2017). All experiments were conducted on NVIDIA A100 GPUs, using Python 3.10 and PyTorch 2.0.

Evaluation Metrics

Natural Language Generation Metrics. We evaluated linguistic quality with BLEU-4 (Papineni et al. 2002), ROUGE-L (Lin 2004), and CIDEr (Vedantam, Zitnick, and Parikh 2015). These metrics respectively measure n-gram overlap, sequence-level similarity, and term distinctiveness, collectively reflecting lexical accuracy, structural coherence, and semantic fidelity of the generated reports. Here, semantic fidelity is defined as the extent to which the generated report preserves the original clinical meaning.

Clinical Efficacy Metrics. To assess diagnostic validity, we applied the CheXpert labeler (Irvin et al. 2019) to extract labels for 14 common chest pathologies from both generated

Algorithm	Venue	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE-L	CIDEr
IU X-Ray							
R2Gen (Chen et al. 2020)	EMNLP'20	0.470	0.304	0.219	0.165	0.371	–
R2GenCMN (Chen et al. 2021)	IJCNLP'21	0.475	0.309	0.222	0.170	0.375	–
PPKED (Liu et al. 2021)	CVPR'21	0.483	0.315	0.224	0.168	0.376	0.351
AlignTrans (You et al. 2021)	MICCAI'21	0.484	0.313	0.225	0.173	0.379	–
CMCL (Liu, Ge, and Wu 2021)	ACL'21	0.473	0.305	0.217	0.162	0.378	–
Clinical-BERT (Yan and Pei 2022)	AAAI'22	0.495	**0.330**	0.231	0.170	0.376	0.432
METransformer (Wang et al. 2023a)	CVPR'23	0.483	0.322	0.228	0.172	0.380	0.435
DCL (Li et al. 2023)	CVPR'23	–	–	–	0.163	0.383	0.586
R2GenGPT (Wang et al. 2023b)	Meta-Rad'23	0.465	0.299	0.214	0.161	0.376	0.542
PromptMRG (Jin et al. 2024)	AAAI'24	0.401	–	–	0.098	0.160	–
BootstrappingLLM (Liu et al. 2024a)	AAAI'24	**0.499**	0.323	0.238	0.184	**0.390**	–
MambaXray (Wang et al. 2025b)	CVPR'25	0.491	**0.330**	0.241	0.185	0.371	0.524
Ours	–	0.495	0.328	**0.242**	**0.187**	0.384	**0.634**
MIMIC-CXR							
R2Gen (Chen et al. 2020)	EMNLP'20	0.353	0.218	0.145	0.103	0.277	–
R2GenCMN (Chen et al. 2021)	IJCNLP'21	0.353	0.218	0.148	0.106	0.278	–
PPKED (Liu et al. 2021)	CVPR'21	0.360	0.224	0.149	0.106	0.284	0.237
AlignTrans (You et al. 2021)	MICCAI'21	0.378	0.235	0.156	0.112	0.283	–
CMCL (Liu, Ge, and Wu 2021)	ACL'21	0.344	0.217	0.140	0.097	0.281	–
Clinical-BERT (Yan and Pei 2022)	AAAI'22	0.383	0.230	0.151	0.106	0.275	0.151
METransformer (Wang et al. 2023a)	CVPR'23	0.386	0.250	0.169	0.124	**0.291**	**0.362**
DCL (Li et al. 2023)	CVPR'23	–	–	–	0.109	0.284	0.281
R2GenGPT (Wang et al. 2023b)	Meta-Rad'23	0.408	0.256	0.174	0.125	0.285	0.244
PromptMRG (Jin et al. 2024)	AAAI'24	0.398	–	–	0.112	0.268	–
BootstrappingLLM (Liu et al. 2024a)	AAAI'24	0.402	0.262	0.180	0.128	**0.291**	–
MambaXray (Wang et al. 2025b)	CVPR'25	0.422	0.268	0.184	**0.133**	0.289	0.241
Ours	–	**0.428**	**0.272**	**0.187**	0.131	**0.291**	0.232

Table 3: Comparison with state-of-the-art methods on IU X-Ray and MIMIC-CXR.

and reference reports. Label agreement was then quantified using Precision, Recall, and F1-score (Powers 2011), ensuring the generated text accurately captures clinically findings.

Results and Discussion

Results on Report Generation

On CheXpert Plus, our framework establishes new state-of-the-art performance benchmarks (Table 1). The method significantly outperforms existing approaches on clinical efficacy metrics (Wang et al. 2025b), demonstrating superior diagnostic accuracy and pathology-detection capabilities. The DVAF module effectively integrates visual and disease-specific information, while the DMSR mechanism successfully retrieves relevant exemplars from the large corpus, enabling comprehensive and accurate report generation.

As shown in Table 3, evaluation on IU X-Ray validates the robustness of our approach across different data scales. Despite limited training data, our method maintains competitive performance. The DAST mechanism effectively captures pathological patterns, and DMSR retrieval enhances report completeness and medical terminology usage, resulting in improved quality scores across linguistic evaluations.

Table 3 also demonstrates that our DVAF-DMSR framework achieves superior performance across NLG metrics compared to existing state-of-the-art methods (Chen et al. 2020; Li et al. 2023; Wang et al. 2023b) in MIMIC-CXR. The integration of DASTs and the DMSR mechanism leads to notable improvements in BLEU-4, ROUGE-L, and CIDEr scores. Our approach consistently outperforms traditional encoder–decoder frameworks and recent vision–language models (Wang et al. 2025b), demonstrating the effectiveness of the two-stage training strategy. CIDEr is often lower on MIMIC-CXR because its reports are more free-form and stylistically diverse than those in other datasets with well-structured formats, reducing consistent n-gram overlap despite good BLEU and ROUGE-L scores.

Ablation Study

To validate the effectiveness of each component in our proposed framework, we conduct comprehensive ablation studies on CheXpert Plus and the IU X-Ray datasets. We systematically remove or replace key components to analyze their individual contributions to the overall performance.

Effect of DASTs. We first examine the contribution of DASTs by comparing our full model with a baseline that only uses visual patch tokens (Table 4). The results demonstrate that incorporating DASTs significantly improves linguistic quality, validating the importance of explicit disease-aware representations for medical report generation.

Impact of DVAF. To assess the effectiveness of our proposed DVAF module, we replace it with simple concatenation or mean pooling of visual and disease tokens (Ta-

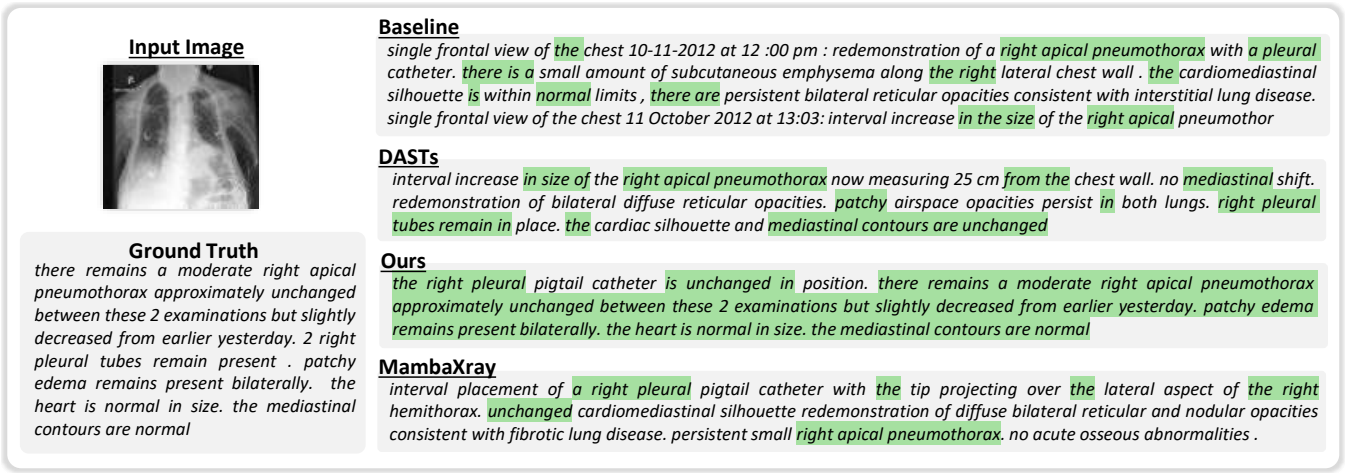

Figure 3: Visualization of generated reports on a sample X-ray.

Configuration	BLEU4	ROUGE-L	CIDEr
CheXpert Plus			
Baseline	0.114	0.283	0.193
+ DASTs + DVAF	0.122	0.288	0.206
+ DASTs + DVAF + DMSR	**0.133**	**0.291**	**0.227**
IU X-Ray			
Baseline	0.161	0.374	0.550
+ DASTs + DVAF	0.175	0.380	0.597
+ DASTs + DVAF + DMSR	**0.187**	**0.384**	**0.634**

Table 4: Component contribution analysis.

Fusion Method	BLEU4	ROUGE-L	CIDEr
CheXpert Plus			
Simple Concatenation	0.110	0.279	0.188
Mean Pooling	0.112	0.285	0.190
DVAF (Ours)	**0.122**	**0.288**	**0.206**
IU X-Ray			
Simple Concatenation	0.156	0.370	0.525
Mean Pooling	0.161	0.375	0.553
DVAF (Ours)	**0.175**	**0.380**	**0.597**

Table 5: Comparison of different fusion methods.

ble 5). The comparison reveals that DVAF applies disease-conditioned spatial attention to highlight lesion regions, then fuses this fine-grained focus with global context to produce a unified feature, leading to substantial improvements in report quality.

Contribution of DMSR. We evaluate the impact of our DMSR mechanism by training models without retrieval augmentation (Table 4). The results show that DMSR consistently enhances performance across all metrics, particularly by improving report completeness and medical terminology accuracy through contextual example guidance. Figure 3 qualitatively confirms these findings: the DMSR-enabled model accurately reinstates key phrases, producing a narrative that mirrors the ground truth more faithfully than Baseline and MambaXray(Wang et al. 2025b) outputs.

Visualization

Figure 3 provides a visualization of four configurations - Baseline (only visual encoder), DASTs (DASTs+DVAF), our full model, and the SOTA MambaXray - against the ground-truth report for a chest X-ray. Exact phrase matches with the reference are highlighted in yellow. The denser yellow regions in our output demonstrate its higher fidelity and clinical completeness compared to competing methods.

Conclusion

In this paper, we present a novel framework for automated chest X-ray report generation that introduces Disease-Aware Semantic Tokens (DASTs), a Disease-Visual Attention Fusion (DVAF) module, and a Dual-Modal Similarity Retrieval (DMSR) mechanism. Our two-stage training strategy effectively combines VMamba's computational efficiency with robust visual-semantic alignments and disease-specific representations. Comprehensive experiments across three benchmark datasets demonstrate state-of-the-art performance in both natural language generation and clinical efficacy metrics. Ablation studies confirm the effectiveness of each proposed component, while visualization results validate the clinical relevance of generated reports. Our framework represents a significant advancement toward practical automated radiology reporting systems. The integration of disease-aware semantic guidance, efficient visual encoding, and retrieval-augmented generation provides a promising foundation for future medical AI applications. Future work may explore extending this approach to other medical imaging modalities and clinical deployment scenarios.

Acknowledgments

This material is based upon work supported by the U.S. National Science Foundation under Award No. CAREER 2145640.

References

Alsentzer, E.; Murphy, J.; Boag, W.; Weng, W.-H.; Jin, D.; Naumann, T.; and McDermott, M. 2019. Publicly available clinical BERT embeddings. In *Clinical Natural Language Processing Workshop*, 72–78.

Chambon, P.; Delbrouck, J.-B.; Sounack, T.; Huang, S.-C.; Chen, Z.; Varma, M.; Truong, S. Q. H.; Chuong, C. T.; and Langlotz, C. P. 2024. CheXpert Plus: Augmenting a Large Chest X-ray Dataset with Text Radiology Reports, Patient Demographics and Additional Image Formats. *arXiv preprint arXiv:2405.19538*.

Chen, W.; Liu, Y.; Wang, C.; Zhu, J.; Li, G.; Liu, C.-L.; and Lin, L. 2025. Cross-Modal Causal Representation Learning for Radiology Report Generation. *Trans. Img. Proc.*, 34: 2970–2985.

Chen, Z.; Diao, S.; Wang, B.; Li, G.; and Wan, X. 2023. PTUnifier: Soft Prompt Unification for Medical Vision–Language Pre-training. In *Iccv*, 23346–23356.

Chen, Z.; Shen, Y.; Song, Y.; and Wan, X. 2021. Cross-modal Memory Networks for Radiology Report Generation. In *Acl-ijcnlp*, 5904–5914.

Chen, Z.; Song, Y.; Chang, T.-H.; and Wan, X. 2020. Generating radiology reports via memory-driven transformer. In *Emnlp*, 1439–1449.

Demner-Fushman, D.; Kohli, M. D.; Rosenman, M. B.; Shooshan, S. E.; Rodriguez, L.; Antani, S.; Thoma, G. R.; and McDonald, C. J. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2): 304–310.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Iclr*.

Gu, A.; and Dao, T. 2024. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv preprint arXiv:2312.00752*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Cvpr*, 15979–15988.

Hou, W.; Xu, K.; Cheng, Y.; Li, W.; and Liu, J. 2023. ORGAN: Observation-Guided Radiology Report Generation via Tree Reasoning. In *Acl*, 8108–8122.

Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In *Aaai*, volume 33, 590–597.

Jain, S.; Agrawal, A.; Saporta, A.; Truong, S. Q.; Duong, D. N.; Bui, T.; Chambon, P.; Zhang, Y.; Lungren, M. P.; Ng, A. Y.; et al. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. In *NeurIPS Track on Datasets and Benchmarks*.

Jin, H.; Che, H.; Lin, Y.; and Chen, H. 2024. PromptMRG: Diagnosis-Driven Prompts for Medical Report Generation. In *Aaai*, 2607–2615.

Johnson, A. E. W.; Pollard, T. J.; Berkowitz, S. J.; Horng, S.; Mark, R. G.; and et al. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1): 317.

Li, M.; Lin, B.; Chen, Z.; Lin, H.; Liang, X.; and Chang, X. 2023. Dynamic graph enhanced contrastive learning for chest X-ray report generation. In *Cvpr*, 3334–3344.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out (ACL Workshop)*, 74–81.

Liu, C.; Tian, Y.; Chen, W.; Song, Y.; and Zhang, Y. 2024a. Bootstrapping Large Language Models for Radiology Report Generation. In *Aaai*, 18635–18643.

Liu, F.; Ge, S.; and Wu, X. 2021. Competence-Based Multimodal Curriculum Learning for Medical Report Generation. In *Acl-ijcnlp*, 3001–3012.

Liu, F.; Wu, X.; Ge, S.; Fan, W.; and Zou, Y. 2021. Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation. In *Cvpr*, 13753–13762.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; and Liu, Y. 2024b. Vmamba: Visual state space model. *NeurIPS*, 37: 103031–103063.

Loshchilov, I.; and Hutter, F. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. In *Iclr*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *Iclr*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Acl*, 311–318.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; and et al. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS*, 8024–8035.

Powers, D. M. W. 2011. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. *Journal of Machine Learning Technologies*, 2(1): 37–63.

Qin, H.; and Song, Y. 2022. Reinforced Cross-modal Alignment for Radiology Report Generation. In *Findings of ACL*, 448–458.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning transferable visual models from natural language supervision. In *Icml*, volume 139, 8748–8763.

Tanida, T.; Müller, P.; Kaissis, G.; and Rückert, D. 2023. Interactive and Explainable Region-Guided Radiology Report Generation. In *Cvpr*, 7433–7442.

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. CIDEr: Consensus-Based Image Description Evaluation. In *Cvpr*, 4566–4575.

Wang, J.; Bhalerao, A.; and He, Y. 2022. Cross-modal Prototype Driven Network for Radiology Report Generation. In *Eccv*, 563–579.

Wang, J.; Bhalerao, A.; Yin, T.; See, S.; and He, Y. 2024a. CAMANet: Class Activation Map guided Attention Network for Radiology Report Generation. *IEEE J. Biomed. Health Inform.*, Pp.

Wang, X.; Li, Y.; Wang, F.; Wang, S.; Li, C.; and Jiang, B. 2024b. R2GenCSR: Retrieving context samples for large language model based X-ray medical report generation. *arXiv [cs.CV]*.

Wang, X.; Li, Y.; Wu, W.; Jin, J.; Rong, Y.; Jiang, B.; Li, C.; and Tang, J. 2025a. Pre-training on high-resolution X-ray images: an experimental study. *Vis. Intell.*, 3(1).

Wang, X.; Wang, F.; Li, Y.; Ma, Q.; Wang, S.; Jiang, B.; and Tang, J. 2025b. CXPMRG-Bench: Pre-training and Benchmarking for X-ray Medical Report Generation on CheXpert Plus Dataset. In *Cvpr*, 5123–5133.

Wang, Z.; Liu, L.; Wang, L.; and Zhou, L. 2023a. METransformer: Radiology Report Generation by Transformer with Multiple Learnable Expert Tokens. In *Cvpr*, 11558–11567.

Wang, Z.; Liu, L.; Wang, L.; and Zhou, L. 2023b. R2GenGPT: Radiology report generation with frozen LLMs. *Meta-Radiology*, 1(3): 100033.

Wang, Z.; Tang, M.; Wang, L.; Li, X.; and Zhou, L. 2022. A Medical Semantic-Assisted Transformer for Radiographic Report Generation. In *Miccai*, 655–664.

Wu, P.; Lin, M.; Chen, Q.; Chew, E. Y.; Lu, Z.; Peng, Y.; and Dong, H. 2025. AMD-Mamba: A Phenotype-Aware Multi-Modal Framework for Robust AMD Prognosis. *arXiv preprint arXiv:2508.02957*.

Wu, P.; Weng, H.; Luo, W.; Zhan, Y.; Xiong, L.; Zhang, H.; and Yan, H. 2023. An improved Yolov5s based on transformer backbone network for detection and classification of bronchoalveolar lavage cells. *Computational and Structural Biotechnology Journal*, 21: 2985–3001.

Wu, Y.; Huang, I.; and Huang, X. 2023. Token Imbalance Adaptation for Radiology Report Generation. In *ACM CHIL (PMLR Vol.209)*, 72–85.

Xu, Z.; Ma, H.; Ding, Y.; Zhang, G.; Weng, C.; and Peng, Y. 2025. Natural Language Processing in Support of Evidence-based Medicine: A Scoping Review. *Findings of the Association for Computational Linguistics: ACL 2025*, 21421–21443.

Xue, Y.; Tan, Y.; Qin, J.; and Xiang, X. 2024a. History Enhanced Radiology Report Generation with Longitudinal Context. *Medical Image Analysis*, 87: 102799.

Xue, Y.; Tan, Y.; Tan, L.; Qin, J.; and Xiang, X. 2024b. Generating Radiology Reports via Auxiliary Signal Guidance and a Memory-Driven Network. *Expert Systems with Applications*, 237: 121260.

Yan, A.; He, Z.; Lu, X.; Du, J.; Chang, E.; Gentili, A.; McAuley, J.; and Hsu, C.-N. 2021. Weakly Supervised Contrastive Learning for Chest X-ray Report Generation. In *Findings of ACL (EMNLP)*, 4009–4015.

Yan, B.; and Pei, M. 2022. Clinical-BERT: Vision-Language Pre-training for Radiograph Diagnosis and Reports Generation. In *Aaai*, 2982–2990.

Yang, S.; Wu, X.; Ge, S.; Zheng, Z.; Zhou, S.; and Xiao, L. 2023. Radiology Report Generation with a Learned Knowledge Base and Multi-modal Alignment. *Medical Image Analysis*, 86: 102798.

Yang, Y.; Yu, J.; Fu, Z.; Zhang, K.; Yu, T.; Wang, X.; Jiang, H.; Lv, J.; Huang, Q.; and Han, W. 2024. Token-Mixer: Bind Image and Text in One Embedding Space for Medical Image Reporting. *Ieee Tmi*.

You, D.; Liu, F.; Ge, S.; Xie, X.; Zhang, J.; and Wu, X. 2021. AlignTransformer: Hierarchical Alignment of Visual Regions and Disease Tags for Medical Report Generation. In *Miccai*, 72–82.

You, K.; Gu, J.; Ham, J.; Park, B.; Kim, J.; Hong, E. K.; Baek, W.; and Roh, B. 2023. CXR-CLIP: Toward Large-Scale Chest X-ray Language–Image Pre-training. In *Miccai*, 101–111.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Icml*.

Zhu, Q.; Mathai, T. S.; Mukherjee, P.; Peng, Y.; Summers, R. M.; and Lu, Z. 2023. Utilizing Longitudinal Chest X-rays and Reports to Pre-fill Radiology Reports. In *Miccai*, 189–198.