Region-Aware Hilbert-Ordered Vision Mamba for Radiology Report Generation

Anonymous ACL submission

Abstract

Automated radiology report generation must produce clinically accurate descriptions that precisely attribute findings to anatomical regions, a prerequisite for trust in clinical deployment. Yet existing methods predominantly exploit global image–text correlations, thereby overlooking localized pathologies and producing factually inconsistent findings even when surface-level fluency is high. We argue that faithful report generation requires explicitly disentangling long-range anatomical context from localized lesion evidence, a decomposition that prior unified encoders cannot achieve. To this end, we propose RA-HVM, a Region-Aware Hilbert Vision Mamba framework for generating structured radiology reports. RA-HVM encodes chest X-ray patches, reordered along a Hilbert space-filling curve, via a Vision Mamba backbone, capturing efficient long-range spatial dependencies while preserving anatomical locality. In parallel, a Faster R-CNN detector identifies candidate lesion regions, which a Region Enhancement refines to model inter-region dependencies. A Q-Former then compresses the fused global–local representation into visual prefix embeddings for language-conditioned decoding, enabling the generator to attend simultaneously to holistic anatomy and fine-grained regional evidence. Experiments on Medical Report Generation benchmarks, including IU X-Ray, MIMIC-CXR, and CheXpert Plus, demonstrate that RA-HVM consistently outperforms strong baselines, achieving gains of +0.023 BLEU-4 and +0.043 CIDEr on IU X-Ray, +0.044 BLEU-1 and +0.243 CIDEr on MIMIC-CXR, and +0.006 BLEU-4 with competitive ROUGE-L performance on CheXpert Plus.

1 Introduction

Radiology report generation aims to convert medical images into clinically meaningful free-text reports describing anatomical structures, abnormal findings, and diagnostic impressions. This task is clinically important because radiologists must examine large volumes of imaging studies and produce accurate reports under substantial workload pressure (Johnson et al., 2019; Rimmer, 2017; Rosenkrantz et al., 2016). Although recent medical vision-language models have improved report fluency, generated reports often remain incomplete or factually inconsistent, particularly when subtle localized abnormalities are missed or when visual evidence is weakly aligned with the generated text (Miura et al., 2021; Chen et al., 2020).

Most existing methods follow an encoder–decoder paradigm, where CNN-, Transformer-, or pretrained vision-based encoders extract visual features and language decoders such as LSTM, GRU, Transformer, or large language models generate the report (Chen et al., 2020, 2021). However, this design remains limited in three aspects. First, global image-level features may suppress small but clinically important findings, such as focal opacities, nodules, pleural abnormalities, or device-related changes. Second, region-guided methods improve local grounding but often require expensive fine-grained annotations or fail to model dependencies among detected regions (Tanida et al., 2023). Third, Transformer-based visual encoders scale quadratically with the number of tokens, making dense radiology image modelling computationally expensive. These limitations indicate that effective report generation requires visual representations that are global, local, spatially structured, and language-aligned.

To address these challenges, we propose RA-HVM, a multi-stage global-local framework for radiology report generation. First, we pre-train a Vision Mamba Model (HO-VMamba) using Hilbert-ordered autoregressive visual modeling. Each image is divided into non-overlapping patches, reordered by a Hilbert space-filling curve, and processed as a locality-preserving visual sequence.

This enables efficient long-range anatomical modelling with approximately linear sequence complexity while preserving spatial neighbourhood structure (Gu and Dao, 2023; Zhu et al., 2024b). Second, we introduce a detector-guided regional pathway to capture lesion-sensitive local evidence. A ResNet backbone and Faster R-CNN detector extract candidate anatomical and abnormal regions, which are refined by a Region Enhancement to model inter-region relationships and strengthen discriminative local cues (He et al., 2016; Ren et al., 2015). Third, we fuse the global HO-VMamba features and enhanced regional features, and align them with the report text using a Q-Former inspired by parameter-efficient vision-language modelling (Li et al., 2023b). The Q-Former compresses the fused visual evidence into language-compatible query embeddings, which are projected into the GPT-2 embedding space and used as visual prefix tokens for report generation (Radford et al., 2019).

Our contributions are summarized as follows. First, we introduce HO-VMamba pre-training for radiology images, enabling efficient global visual modeling with preserved spatial locality. Second, we design a detector-guided regional enhancement pathway to capture localized anatomical and abnormal findings. Third, we propose a Q-Former-based alignment module that converts fused global-local visual features into GPT-2-compatible prefix embeddings. Finally, we optimize RA-HVM with image-text contrastive learning, image-text matching, and autoregressive report-generation objectives, and evaluate it on IU X-Ray, MIMIC-CXR, and CheXpert Plus.

2 Proposed Methodology

2.1 Overview

As illustrated in Fig. 1, the proposed RA-HVM is a multi-stage radiology report generation framework that integrates HO-VMamba, detector-guided regional representation learning, and Q-Former-based vision–language alignment. For each chest X-ray, image patches are reordered using a Hilbert curve and encoded by a Vision Mamba backbone to capture global anatomical structure and long-range spatial dependencies. In parallel, a frozen VinDr-CXR-pretrained Faster R-CNN with a ResNet-34 backbone extracts anatomical and abnormality-focused regions for local evidence modeling. Global and regional visual representations are then fused and compressed by learnable Q-Former queries into GPT-2-compatible prefix embeddings for autoregressive report generation. This design enables RA-HVM to align global context with localized pathological cues, improving the clinical coherence and evidential grounding of generated reports. The details of each stage are described in the following subsections.

2.1.1 Stage #1: Autoregressive Pre-training of HO-VMamba Model

To obtain a strong radiology-specific visual backbone, we pre-train the HO-VMamba using a self-supervised autoregressive objective. Inspired by autoregressive modeling in large generative models such as ChatGPT (OpenAI, 2023) and GPT-4 (Achiam et al., 2023), the model learns to predict future visual tokens conditioned on past context. Unlike standard raster scanning, we apply a Hilbert curve to preserve spatial locality before sequence modeling.

Let the input X-ray image be $I \in R^{224\times224\times3}$. We divide it into non-overlapping patches $p_i \in R^{16\times16\times3}$, where $i = 1, 2, \ldots, N$ and $N = 196$. Each patch is mapped into a visual token $t_i = \phi(p_i) \in R^{1024}$ using a convolutional embedding layer.

To preserve spatial structure, we reorder tokens using a Hilbert curve π_H, producing the sequence $T_H = \{t_{\pi_H(1)}, t_{\pi_H(2)}, \ldots, t_{\pi_H(N)}\}$. The ordered tokens are then processed by the HO-VMamba, which efficiently models long-range dependencies with linear complexity $O(N)$ via selective state-space modeling and gated state updates.

The autoregressive pre-training objective predicts the next continuous visual token from the preceding Hilbert-ordered context. Unlike language modeling, where the target token is discrete and optimized with a negative log-likelihood loss, our visual tokens are continuous patch embeddings. Therefore, HO-VMamba is trained using next-token embedding regression.

Given the Hilbert-ordered token sequence $T_H = \{t_{\pi_H(1)}, t_{\pi_H(2)}, \ldots, t_{\pi_H(N)}\}$, the prediction at step i is defined as

$$\hat{t}_{i+1} = g_\Theta(t_{\leq i}), \tag{1}$$

where $g_\Theta(\cdot)$ denotes the HO-VMamba prediction head. The autoregressive pre-training loss is

$$\mathcal{L}_{\text{AR}} = \frac{1}{N-1} \sum_{i=1}^{N-1} \left\| \hat{t}_{i+1} - t_{i+1} \right\|_2^2 . \tag{2}$$

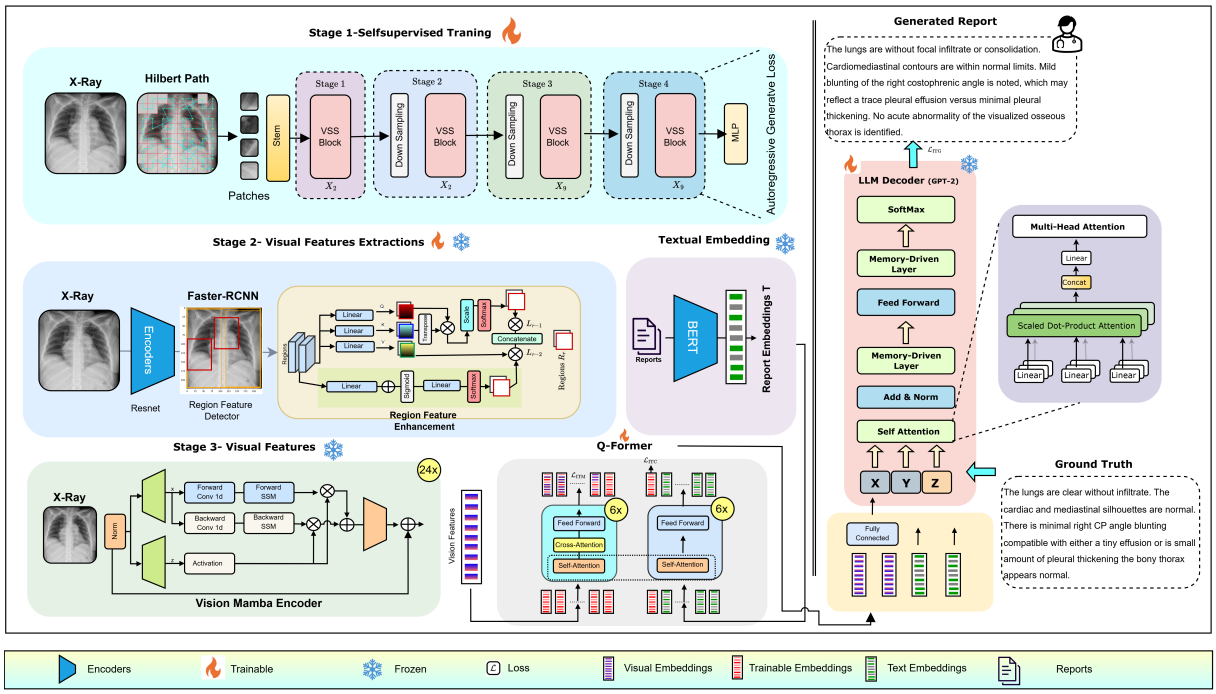

Figure 1: Proposed RA-HVM framework: (1) HO-VMamba global encoding with Hilbert-ordered image patches, (2) detector-guided regional feature extraction for local pathological evidence, and (3) Q-Former-based visual–language alignment for autoregressive report generation.

where $g_\Theta(\cdot)$ is the prediction head that maps contextualized token representations to the next-token prediction space. Random token masking is applied during training to prevent trivial identity learning and to encourage robust contextual representation learning.

2.1.2 Stage 2: Faster R-CNN-Based Regional Feature Extraction and Enhancement

Although the HO-VMamba captures global anatomical structure, radiology report generation also requires localized clinically relevant visual evidence from chest X-ray images. To address this, we introduce a detector-guided regional branch based on Faster R-CNN (Ren et al., 2015).

We first train the Faster R-CNN detector on the VinDr-CXR dataset (Nguyen et al., 2020), which provides bounding-box annotations for thoracic abnormalities. The detector uses a ResNet-34 (He et al., 2016) backbone for feature extraction. After pretraining, all components of the detector (backbone, RPN, and ROI head) are frozen and used as a fixed region proposal module during report generation training.

Given an input image I, the detector extracts backbone feature maps:

$$F_{\text{det}} = f_{\text{ResNet34}}(I), \tag{3}$$

which are processed by the Region Proposal Network (RPN) to generate candidate regions:

$$B = f_{\text{RPN}}(F_{\text{det}}), \tag{4}$$

where $B = \{(b_i, s_i)\}_{i=1}^{N_b}$ denotes region proposals, $b_i = (x_i, y_i, w_i, h_i)$ is the bounding box, and s_i is the confidence score. The top-m proposals are selected after non-maximum suppression, where $m \in [10, 100]$.

For each selected region, RoI pooling produces regional descriptors:

$$r_i = \text{RoIPool}(F_{\text{det}}, b_i), \quad r_i \in R^D. \tag{5}$$

The regional feature set is defined as:

$$R = \{r_1, r_2, \ldots, r_m\} \in R^{m \times D}. \tag{6}$$

To refine region-level representations, we introduce a **Region Enhancement** that models inter-region dependencies using multi-head self-attention. For the j-th attention head, the query, key, and value matrices are computed as $Q_j = RW_j^Q$, $K_j = RW_j^K$, and $V_j = RW_j^V$, where W_j^Q, W_j^K, and W_j^V denote learnable projection matrices.

$$H_j(R) = \text{Softmax}\left(\frac{Q_j K_j^\top}{\sqrt{d_k}}\right) V_j, \tag{7}$$

and aggregated as:

$$R_{\text{rel}} = \text{Concat}(H_1(R), \ldots, H_h(R))W^O. \quad (8)$$

In parallel, a feature refinement branch applies a residual MLP transformation:

$$R_{\text{trans}} = \text{MLP}(R) + R. \quad (9)$$

Finally, the enhanced regional representation is obtained by concatenation:

$$Z_r = R_{\text{rel}} \oplus R_{\text{trans}}, \quad (10)$$

2.1.3 Stage 3: Frozen HO-VMamba Feature Extraction

After Stage 1, the pretrained HO-VMamba Model is reused as a frozen global visual feature extractor for downstream report generation. For each input chest X-ray image, we apply the same patch partitioning and Hilbert ordering strategy used during autoregressive pretraining to maintain spatial and sequential consistency. The resulting Hilbert-ordered token sequence T_H is passed through the frozen encoder to obtain the global visual representation:

$$Z_m = f^{\text{frozen}}_{\text{HO-VMamba}}(T_H), \quad (11)$$

where Z_m denotes the global anatomical representation extracted from the input image. Unlike Stage 1, no token masking or the autoregressive reconstruction objective is applied during this stage, and the full token sequence is used for feature extraction.

The enhanced regional representation Z_r is concatenated with the global HO-VMamba representation Z_m. The concatenated feature is then projected through a linear transformation to obtain a 512-dimensional fused visual representation Z_v, which is used for downstream report generation.

2.2 Q-Former-Based Visual-Language Alignment

To align radiology image representations with clinical language semantics, we employ a Q-Former module as a lightweight visual-language bridge between the frozen visual encoders and the GPT-2 decoder, following the parameter-efficient design of BLIP-2 (Li et al., 2023b). The Q-Former consists of a stack of Transformer blocks with shared self-attention layers, enabling interaction between visual query embeddings and textual representations. A set of learnable query embeddings is used to extract clinically relevant information from the fused visual representation through cross-attention operations.

The fused representation Z_v is then processed by the Q-Former using learnable query embeddings:

$$Q_v = \text{QFormer}(Q, Z_v), \quad (12)$$

where Q denotes the learnable query tokens and Q_v represents the visual-language aligned query features.

To reduce memory consumption and preserve pretrained visual knowledge, only the Q-Former parameters are optimized during representation learning, while the visual encoders and Clinical-BERT (Alsentzer et al., 2019) text encoder remain frozen. The Q-Former parameters are initialized from BERT-base-uncased, while the cross-attention layers and query embeddings are randomly initialized.

Following BLIP-2, the Q-Former is optimized using three complementary objectives: image-text contrastive learning, image-text matching, and image-grounded text generation. The image-text contrastive objective maximizes the similarity between matched image-report pairs while minimizing mismatched pairs:

$$\mathcal{L}_{\text{ITC}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\text{sim}(q_i, t_i)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}(q_i, t_j)/\tau\right)}, \quad (13)$$

where q_i and t_i denote the aligned visual and textual embeddings, N is the batch size, and τ is a learnable temperature parameter.

The image-text matching objective is formulated as a binary classification problem:

$$\mathcal{L}_{\text{ITM}} = -\frac{1}{N}\sum_{i=1}^{N} \left[y_i \log p_i + (1 - y_i)\log(1 - p_i)\right], \quad (14)$$

where p_i is the predicted matching probability and $y_i \in \{0, 1\}$ indicates whether the image-report pair is matched.

To further encourage clinically grounded report generation, an image-conditioned autoregressive text generation objective is employed using teacher forcing (Williams and Zipser, 1989). The loss is optimized by minimizing the cross-entropy between the predicted token distribution and the ground-truth report tokens:

$$\mathcal{L}_{\text{ITG}} = -\frac{1}{T}\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, Q_v), \quad (15)$$

where y_t denotes the target token at timestep t, $y_{<t}$ represents all preceding report tokens, Q_v denotes the aligned visual query features produced by the Q-Former, and T is the report length. A multimodal causal attention mask is applied during training to preserve autoregressive decoding behavior.

The final optimization objective combines the image-text contrastive, image-text matching, and image-grounded text generation losses:

$$\mathcal{L} = \mathcal{L}_{\text{ITC}} + \mathcal{L}_{\text{ITM}} + \mathcal{L}_{\text{ITG}}. \tag{16}$$

During report generation, the aligned query features are projected into the GPT-2 embedding space through a linear projection layer before autoregressive decoding.

3 Experiments

Experiments are conducted on the IU X-Ray (Demner-Fushman et al., 2016), MIMIC-CXR (Johnson et al., 2019), and CheXpert Plus (Chambon et al., 2024) benchmark datasets. Following prior work (Chen et al., 2021; Yang et al., 2023), we adopt a 7:1:2 train/validation/test split for IU X-Ray and use the official data splits for MIMIC-CXR and CheXpert Plus. We first introduce the datasets, implementation details, and evaluation metrics. We then compare RA-HVM with representative state-of-the-art methods using standard natural language generation metrics and qualitative report analysis. Finally, ablation studies are performed to evaluate the contribution of each major component.

3.1 Datasets

We evaluate RA-HVM on three benchmark chest X-ray report generation datasets: **IU X-Ray** (Demner-Fushman et al., 2016), **MIMIC-CXR** (Johnson et al., 2019), and **CheXpert Plus** (Chambon et al., 2024). IU X-Ray contains 7,470 images paired with 3,955 radiology reports, while MIMIC-CXR contains 377,110 images and 227,827 reports. CheXpert Plus provides 223,462 paired chest X-ray images and radiology reports collected from 187,711 studies across 64,725 patients. These datasets contain paired chest radiographs and free-text reports describing thoracic findings.

3.2 Implementation Details

The HO-VMamba Model is pretrained with the Stage-1 autoregressive objective using chest X-ray images resized to 224×224. For autoregressive pretraining, all images from the training split are used, ensuring that the encoder learns global anatomical structure from the full available training distribution. HO-VMamba-Large is trained for 150 epochs with AdamW (Loshchilov and Hutter, 2017), a base learning rate of 1×10^{-5}, weight decay of 0.05, batch size of 128, and cosine learning-rate decay with 5 warmup epochs.

For the detector-guided regional branch, a Faster R-CNN (Ren et al., 2015) detector with a ResNet-34 backbone is pretrained on VinDr-CXR (Nguyen et al., 2020) for 50 epochs using AdamW, a learning rate of 1×10^{-4}, and batch size 16. The pretrained detector is then frozen to generate region proposals and RoI-pooled regional features for downstream report generation.

For visual-language alignment, the Q-Former is trained using 16 learnable query embeddings and optimized with the combined $\mathcal{L}_{\text{ITC}}$, $\mathcal{L}_{\text{ITM}}$, and $\mathcal{L}_{\text{ITG}}$ objectives against a frozen ClinicalBERT (Alsentzer et al., 2019) text encoder. Training is performed for 25 epochs using AdamW with a learning rate of 1×10^{-4}, weight decay of 0.01, $\beta_1 = 0.9$, $\beta_2 = 0.999$, cosine learning-rate scheduling, and 1,000 warmup steps. During the report generation stage, the second phase of the BLIP-2 training procedure increases the number of learnable query embeddings to 64, where the Q-Former is connected to GPT-2 for image-grounded report generation. A linear projection layer maps the image-informed query embeddings from the Q-Former into the GPT-2 embedding space. During training, the pretrained Q-Former, the projection layer, and the final GPT-2 language modeling layer are optimized using the image-grounded text generation loss. The model is trained for 20 epochs using AdamW with a learning rate of 1×10^{-3}, batch size 36, cosine scheduling, and 1,000 warmup steps. Dataset-specific maximum decoding lengths are set to 80 for IU X-Ray, 76 for MIMIC-CXR, and 90 for CheXpert Plus. The final model is selected based on the lowest validation loss. The model was trained using four NVIDIA RTX A5000 GPUs, each with 25 GB of memory.

3.3 Evaluation Metrics

We evaluate the quality of generated radiology reports from three complementary perspectives: textual quality, semantic consistency, and clinical correctness. For textual evaluation, we report standard natural language generation met-

Table 1: Comparison with state-of-the-art medical report generation methods on three benchmark datasets **IU X-Ray, MIMIC-CXR, CheXpert Plus**. Best results are highlighted in **bold**, and the strongest competing baseline results are underlined. All baseline scores are reported directly from the original papers.

Dataset	Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
IU X-Ray	MRAM (Xue et al., 2018)	0.464	0.358	0.270	0.195	0.274	0.366	–
	R2Gen (Chen et al., 2020)	0.470	0.304	0.219	0.165	0.187	0.371	–
	R2GenCMN (Chen et al., 2021)	0.475	0.309	0.222	0.170	0.191	0.375	–
	PPKED (Liu et al., 2021)	0.491	0.321	0.241	0.155	0.149	0.377	0.365
	R2RL (Qin and Song, 2022)	0.494	0.321	0.235	0.181	0.201	0.384	–
	ICT (Zhang et al., 2023)	0.503	0.341	0.246	0.186	0.208	0.390	–
	S3-Net (Pan et al., 2023)	0.499	0.334	0.246	0.172	0.206	0.401	–
	M2KT (Yang et al., 2023)	0.497	0.319	0.230	0.174	–	0.399	0.407
	LLaVA-Med (Li et al., 2023a)	0.528	0.346	0.237	0.186	–	0.422	–
	SSVE (Divya et al., 2024)	0.492	0.321	0.233	0.180	0.203	0.379	–
	MiniGPT-4 (Zhu et al., 2024a)	0.494	0.329	0.220	0.179	–	0.390	–
	BiomedGPT (Zhang et al., 2024)	0.516	0.343	0.233	0.183	–	0.403	–
	SERPENT-VLM (Kapadnis et al., 2024)	0.547	0.356	0.242	0.190	–	0.452	–
	DART (Park et al., 2025)	0.486	0.348	0.265	0.208	0.205	0.411	–
	MediVLM (Goswami et al., 2025)	0.471	0.296	0.231	0.170	0.194	0.375	–
	DMVF (Tang et al., 2025)	0.518	0.349	0.252	0.190	0.215	0.398	0.411
	RA-HVM (Ours)	**0.561**	**0.359**	0.261	**0.231**	0.229	0.451	**0.454**
MIMIC-CXR	MRAM (Xue et al., 2018)	0.310	0.198	0.126	0.094	0.119	0.257	–
	R2Gen (Chen et al., 2020)	0.353	0.218	0.145	0.103	0.142	0.277	–
	R2GenCMN (Chen et al., 2021)	0.353	0.218	0.148	0.106	0.142	0.278	–
	PPKED (Liu et al., 2021)	0.336	0.220	0.133	0.101	0.147	0.246	–
	R2RL (Qin and Song, 2022)	0.381	0.232	0.155	0.109	0.151	0.287	–
	ICT (Zhang et al., 2023)	0.376	0.233	0.157	0.113	0.144	0.276	–
	S3-Net (Pan et al., 2023)	0.358	0.239	0.158	0.125	0.154	0.291	–
	M2KT (Yang et al., 2023)	0.386	0.237	0.157	0.111	–	0.274	–
	SSVE (Divya et al., 2024)	0.373	0.230	0.161	0.112	0.162	0.297	–
	DART (Park et al., 2025)	0.437	0.279	0.191	0.137	0.310	0.175	–
	MediVLM (Goswami et al., 2025)	0.377	0.246	0.158	0.123	0.149	0.293	–
	DMVF (Tang et al., 2025)	0.396	0.245	0.162	0.116	0.154	0.289	0.116
	RA-HVM (Ours)	**0.481**	**0.287**	**0.198**	0.129	0.184	**0.298**	**0.359**
CheXpert Plus	R2GenCMN (Chen et al., 2021)	–	–	–	0.095	0.142	0.262	–
	R2Gen-GPT (Wang et al., 2023)	–	–	–	0.102	0.157	0.267	0.179
	R2GenCSR (Wang et al., 2026)	–	–	–	0.103	0.163	0.272	0.193
	MambaXray-VL (Wang et al., 2025)	–	–	–	0.112	0.157	0.276	0.139
	RA-HVM (Ours)	**0.431**	**0.213**	**0.158**	**0.118**	**0.161**	**0.278**	**0.142**

rics, including BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2011), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015), which measure lexical overlap, sentence-level similarity, and consensus-based generation quality between generated and reference reports. To further assess semantic similarity and clinical factuality, we additionally report BERTScore (Zhang et al., 2020), RadGraph F1 (Delbrouck et al., 2022), and RaTEScore (Zhao et al., 2024). BERTScore evaluates semantic similarity using contextualized language embeddings, while RadGraph F1 measures agreement between generated and reference clinical entities and relations. RaTEScore is an entity-aware radiology report evaluation metric that emphasizes clinically important medical entities, including diagnostic findings and anatomical descriptions, while accounting for medical synonyms and negation expressions.

4 Results and Discussion

We compare RA-HVM with representative baselines on IU X-Ray, MIMIC-CXR, and CheXpert Plus, as reported in Table 1. RA-HVM achieves competitive performance across the evaluated benchmarks and improves several key metrics over existing methods, although it does not obtain the best score on every metric. In particular, the results suggest that combining Hilbert-ordered global representation learning with detector-guided regional feature enhancement provides complementary benefits for report generation. Compared with recent vision–language models such as LLaVA-Med and BiomedGPT, RA-HVM shows stronger performance on several evaluation metrics, indicating improved use of structured visual evidence. Overall, as illustrated in Fig 2, the proposed global–local framework achieves robust performance across datasets with varying scales and

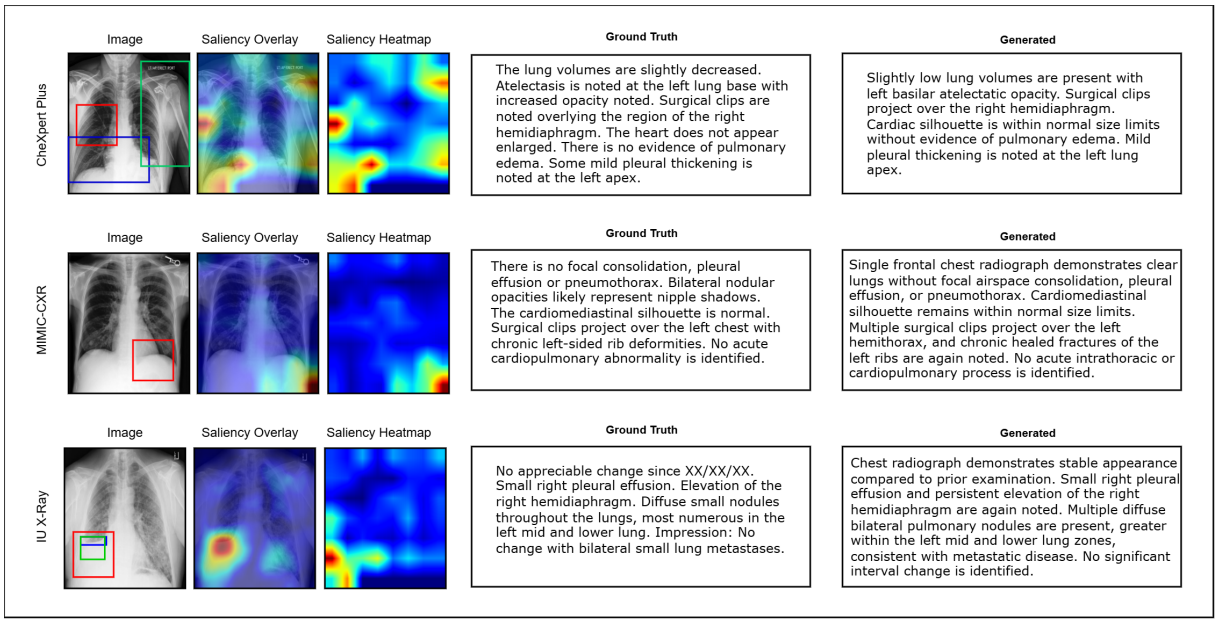

Figure 2: Qualitative chest X-ray report generation results on CheXpert Plus, MIMIC-CXR, and IU X-Ray datasets using the proposed framework with GPT-2 as the report decoder. For each example, the figure presents the input radiograph, saliency overlay, saliency heatmap, corresponding ground-truth report, and the report generated by the model. The saliency visualizations highlight diagnostically relevant regions contributing to report generation, demonstrating the model's ability to capture clinically meaningful abnormalities and generate coherent radiology reports aligned with the reference findings.

reporting complexity.

4.1 Semantic and Clinical Evaluation

Table 3 compares RA-HVM with recent radiology report generation methods using BERTScore, RadGraph-F1 (Jain et al., 2021), and RaTEScore (Zhao et al., 2024) on the IU X-Ray dataset. Baseline results, including MediVLM (Goswami et al., 2025), are directly reported from the corresponding papers. The MediVLM authors (Goswami et al., 2025) reproduced several baseline results using the official implementations, while additional comparisons are drawn from RexRank (Zhang et al., 2025). RA-HVM achieves the best performance in both BERTScore and RadGraph-F1, indicating improved semantic alignment and stronger clinical entity-relation consistency in generated reports. Although MediVLM achieves a slightly higher RaTEScore, our method demonstrates more balanced performance across semantic and clinically grounded evaluation metrics.

4.2 Ablation Study

We perform ablation studies to analyze the contribution of each component in RA-HVM. As shown in Table 2, the full model achieves the best performance across IU X-Ray, MIMIC-CXR, and CheXpert Plus, indicating the effectiveness of the proposed unified global-local design.

Global representation learning. Removing Hilbert ordering or autoregressive pre-training consistently degrades performance across all datasets, suggesting that structured sequence modeling improves global radiological representation learning.

Regional feature modeling. Eliminating the regional branch or the Region Enhancement results in noticeable performance drops, highlighting the importance of incorporating localized anatomical and abnormality-aware cues alongside global features.

Visual-language alignment. The largest performance decrease is observed when the Q-Former is removed, indicating that it plays a key role in bridging visual features with the language space for report generation.

Table 2 further shows that using a finetuned GPT-2 decoder consistently outperforms alternatives such as frozen or replacing the decoder, suggesting that preserving pretrained language priors is beneficial in the limited-data medical setting.

Finally, Table 4 shows that (m = 50) regional proposals provide the best balance between anatomical coverage and noise reduction. Moreover, Hilbert ordering consistently achieves the best performance among all patch-ordering strategies, underscoring

Table 2: Unified ablation study across three datasets. B4, M, R-L, and C denote BLEU-4, METEOR, ROUGE-L, and CIDEr, respectively. Best results are highlighted in **bold**.

Category	Variant	IU X-Ray				MIMIC-CXR				CheXpert Plus			
		B4	M	R-L	C	B4	M	R-L	C	B4	M	R-L	C
Architecture	Full model	**0.231**	**0.229**	**0.451**	**0.454**	**0.129**	**0.184**	**0.298**	**0.359**	**0.118**	**0.161**	**0.278**	**0.142**
	w/o Hilbert ordering	0.181	0.207	0.389	0.394	0.109	0.148	0.280	0.107	0.104	0.150	0.267	0.128
	w/o autoregressive pre-training	0.176	0.204	0.383	0.382	0.105	0.145	0.276	0.101	0.100	0.146	0.261	0.121
	w/o regional branch	0.171	0.199	0.376	0.361	0.101	0.141	0.270	0.094	0.097	0.141	0.254	0.115
	w/o Region Enhancement	0.180	0.206	0.386	0.389	0.108	0.147	0.278	0.105	0.103	0.149	0.265	0.126
	w/o Q-Former	0.164	0.192	0.364	0.337	0.095	0.135	0.258	0.083	0.090	0.132	0.241	0.101
Decoder	GPT-2 (Fine-tuned)	**0.231**	**0.229**	**0.451**	**0.454**	**0.129**	**0.184**	**0.298**	**0.359**	**0.118**	**0.161**	**0.278**	**0.142**
	GPT-2 (Frozen)	0.178	0.204	0.384	0.379	0.106	0.145	0.276	0.099	0.101	0.146	0.260	0.120
	BioGPT (Fine-tuned)	0.187	0.213	0.395	0.404	0.113	0.153	0.286	0.111	0.109	0.155	0.273	0.134
	BioGPT (Frozen)	0.181	0.207	0.389	0.391	0.108	0.148	0.280	0.104	0.104	0.150	0.266	0.126
	BioMedLM/PubMedGPT (Fine-tuned)	0.184	0.210	0.391	0.397	0.111	0.150	0.283	0.108	0.106	0.152	0.269	0.130
	BioMedLM/PubMedGPT (Frozen)	0.176	0.203	0.382	0.376	0.105	0.144	0.275	0.098	0.100	0.145	0.259	0.119

Table 3: Comparison of semantic and clinical report generation metrics on the IU X-Ray dataset. Best results are highlighted in **bold**, and second-best results are underlined.

Method	BERTScore	RadGraph-F1	RaTEScore
R2Gen (Chen et al., 2020)	0.487	0.251	0.603
X-RGen (Chen et al., 2024)	0.523	0.387	0.686
CMN (Chen et al., 2021)	0.509	0.281	0.603
HistGen (Guo et al., 2024)	0.466	0.205	0.559
RL (Qin and Song, 2022)	0.331	0.228	0.389
RGRG (Tanida et al., 2023)	0.437	0.223	0.620
VLCI (Chen et al., 2023)	0.455	0.288	0.679
RadFM (Wu et al., 2023)	0.459	0.230	0.627
RaDialog (Pellegrini et al., 2023)	0.444	0.205	0.586
CvT2DistilGPT2 (Nicolson et al., 2023)	0.482	0.265	0.620
MediVLM (Goswami et al., 2025)	0.616	0.406	**0.722**
RA-HVM (Ours)	**0.642**	**0.417**	0.719

Table 4: Ablation studies on proposal number and patch ordering.

Study	Setting	B4	M	RL	C
Proposals (m)	10	0.176	0.202	0.381	0.374
	25	0.185	0.211	0.392	0.401
	50	**0.231**	**0.229**	**0.451**	**0.454**
	75	0.188	0.213	0.395	0.406
	100	0.184	0.210	0.391	0.397
Ordering	Random	0.172	0.198	0.374	0.356
	Raster	0.181	0.207	0.388	0.391
	Z-order	0.185	0.211	0.393	0.402
	Hilbert	**0.231**	**0.229**	**0.451**	**0.454**

the importance of preserving spatial continuity during HO-VMamba pre-training.

Overall, the results indicate that each component contributes to the final performance and that their combination yields the strongest results.

5 Conclusion

We proposed RA-HVM, a unified framework for radiology report generation that combines Hilbert-ordered global visual modeling, detector-guided regional feature extraction, and Q-Former-based vision-language alignment. This design enables joint modeling of global anatomical context and fine-grained clinical evidence for more faithful report generation.

Experiments on IU X-Ray, MIMIC-CXR, and CheXpert Plus show consistent improvements over recent state-of-the-art methods across both natural language generation metrics and clinically grounded evaluation measures, including RadGraph-F1.

Ablation studies confirm that each component contributes to performance, with the strongest gains coming from the combination of global structure modeling and regional feature refinement. These results demonstrate the benefit of structured visual encoding coupled with language-aligned decoding for medical report generation.

Limitations

Although RA-HVM achieves competitive performance across multiple radiology report generation benchmarks, several limitations remain. First, the framework relies on detector-guided regional feature extraction; therefore, the effectiveness of the regional branch depends on the quality of the pre-trained detector proposals. Second, combining global HO-VMamba representations with regional enhancement increases computational and memory cost compared with lightweight encoder–decoder baselines. Third, our experiments are mainly conducted on chest radiography datasets, and the generalization of RA-HVM to other imaging modalities, such as CT or MRI, remains unexplored. Finally, although automatic metrics such as BLEU, CIDEr, RadGraph-F1, and RaTEScore provide useful quantitative evaluation, they may not fully capture diagnostic reliability or the quality of clinical decision-making in real-world deployment.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. 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*.

Qi Chen, Yutong Xie, Biao Wu, Xiaomin Chen, James Ang, Minh-Son To, Xiaojun Chang, and Qi Wu. 2024. Act like a radiologist: radiology report generation across anatomical regions. In *Proceedings of the Asian Conference on Computer Vision*, pages 1–17.

Weixing Chen, Yang Liu, Ce Wang, Guanbin Li, Jiarui Zhu, and Liang Lin. 2023. Visual-linguistic causal intervention for radiology report generation. *arXiv preprint arXiv:2303.09117*, 1(8).

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: long papers)*, pages 5904–5914.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 1439–1449.

Jean-Benoit Delbrouck and 1 others. 2022. Improving factuality in radiology report generation with semantic rewards. In *EMNLP*.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the sixth workshop on statistical machine translation*, pages 85–91.

Peketi Divya, Yenduri Sravani, Chalavadi Vishnu, C Krishna Mohan, and Yen Wei Chen. 2024. Memory guided transformer with spatio-semantic visual extractor for medical report generation. *IEEE Journal of Biomedical and Health Informatics*, 28(5):3079–3089.

Debanjan Goswami, Ronast Subedi, and Shayok Chakraborty. 2025. Medivlm: A vision language model for radiology report generation from medical images. *Findings of the Association for Computational Linguistics: EMNLP*, 2025:10287–10304.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Zhengrui Guo, Jiabo Ma, Yingxue Xu, Yihui Wang, Liansheng Wang, and Hao Chen. 2024. Histgen: Histopathology report generation via local-global feature encoding and cross-modal context interaction. pages 189–199.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

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

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(1):317.

Manav Kapadnis, Sohan Patnaik, Abhilash Nandy, Sourjyadip Ray, Pawan Goyal, and Debdoot Sheet. 2024. Serpent-vlm: self-refining radiology report generation using vision language models. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 283–291.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

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

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13753–13762.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 2021. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304.

Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, and 5 others. 2020. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Preprint*, arXiv:2012.15029.

Aaron Nicolson and 1 others. 2023. Cvt2distilgpt2: Lightweight transformer-based radiology report generation. *arXiv preprint*.

OpenAI. 2023. Chatgpt. Large language model. Accessed: [Insert your access date].

Renjie Pan, Ruisheng Ran, Wei Hu, Wenfeng Zhang, Qibing Qin, and Shaoguo Cui. 2023. S3-net: a self-supervised dual-stream network for radiology report generation. *IEEE Journal of Biomedical and Health Informatics*, 28(3):1448–1459.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Sang-Jun Park, Keun-Soo Heo, Dong-Hee Shin, Young-Han Son, Ji-Hye Oh, and Tae-Eui Kam. 2025. Dart: Disease-aware image-text alignment and self-correcting re-alignment for trustworthy radiology report generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15580–15589.

Lorenzo Pellegrini and 1 others. 2023. Radialog: A large vision-language model for radiology report generation and conversational assistance. *arXiv preprint arXiv:2311.18681*.

Han Qin and Yan Song. 2022. Reinforced cross-modal alignment for radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 448–458.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*.

Abi Rimmer. 2017. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ: British Medical Journal (Online)*, 359.

Andrew B Rosenkrantz, Danny R Hughes, and Richard Duszak Jr. 2016. The us radiologist workforce: an analysis of temporal and geographic variation by using large national datasets. *Radiology*, 279(1):175–184.

Quan Tang, Liming Xu, Yongheng Wang, Bochuan Zheng, Jiancheng Lv, Xianhua Zeng, and Weisheng Li. 2025. Dual-modality visual feature flow for medical report generation. *Medical Image Analysis*, 101:103413.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7433–7442.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Xiao Wang, Yuehang Li, Fuling Wang, Shiao Wang, Chuanfu Li, and Bo Jiang. 2026. R2gencsr: Mining contextual and residual information for llms-based radiology report generation. *IEEE Journal of Biomedical and Health Informatics*.

Xiao Wang, Fuling Wang, Yuehang Li, Qingchuan Ma, Shiao Wang, Bo Jiang, and Jin Tang. 2025. Cxpmrg-bench: Pre-training and benchmarking for x-ray medical report generation on chexpert plus dataset. In *Proceedings of the computer vision and pattern recognition conference*, pages 5123–5133.

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

Ronald J. Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280.

Chaoyi Wu and 1 others. 2023. Radfm: A large-scale vision-language foundation model for radiology. *arXiv preprint arXiv:2308.02463*.

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

and Computer-Assisted Intervention, pages 457–466. Springer.

Shuxin Yang, Xian Wu, Shen Ge, Zhuozhao Zheng, S Kevin Zhou, and Li Xiao. 2023. Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis*, 86:102798.

Junsan Zhang, Xiuxuan Shen, Shaohua Wan, Sotirios K Goudos, Jie Wu, Ming Cheng, and Weishan Zhang. 2023. A novel deep learning model for medical report generation by inter-intra information calibration. *IEEE Journal of Biomedical and Health Informatics*, 27(10):5110–5121.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, and 1 others. 2024. A generalist vision–language foundation model for diverse biomedical tasks. *Nature medicine*, 30(11):3129–3141.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *ICLR*.

Xiaoman Zhang, Hong-Yu Zhou, Xiaoli Yang, Oishi Banerjee, Julián N. Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. 2025. Rexrank: A public leaderboard for ai-powered radiology report generation. In *Proceedings of the First AAAI Bridge Program on AI for Medicine and Healthcare*, volume 281 of *PMLR*, pages 90–99.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Ratescore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 15004–15019. Association for Computational Linguistics.

Deyao Zhu, Xiaoqian Shen, Xiang Li, Mohamed Elhoseiny, and 1 others. 2024a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, volume 2024, pages 18378–18394.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. 2024b. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*.