Pre-training on High Definition X-ray Images: An Experimental Study

Xiao Wang, Yuehang Li, Wentao Wu, Jiandong Jin, Yao Rong, Bo Jiang, Chuanfu Li, Jin Tang

***Abstract*— Existing X-ray based pre-trained vision models are usually conducted on a relatively small-scale dataset (less than 500k samples) with limited resolution (e.g., 224 $\times$ 224). However, the key to the success of self-supervised pre-training large models lies in massive training data, and maintaining high resolution in the field of X-ray images is the guarantee of effective solutions to difficult miscellaneous diseases. In this paper, we address these issues by proposing the first high-definition (1280 $\times$ 1280) X-ray based pre-trained foundation vision model on our newly collected large-scale dataset which contains more than 1 million X-ray images. Our model follows the masked auto-encoder framework which takes the tokens after mask processing (with a high rate) is used as input, and the masked image patches are reconstructed by the Transformer encoder-decoder network. More importantly, we introduce a novel context-aware masking strategy that utilizes the chest contour as a boundary for adaptive masking operations. We validate the effectiveness of our model on two downstream tasks, including X-ray report generation and disease recognition. Extensive experiments demonstrate that our pre-trained medical foundation vision model achieves comparable or even new state-of-the-art performance on downstream benchmark datasets. The source code and pre-trained models of this paper will be released on https://github.com/Event-AHU/Medical_Image_Analysis.**

***Index Terms*— High Definition X-ray Image, Pre-trained Big Models, Masked Auto Encoder, Medical Report Generation**

I. Introduction

MEDICAL image analysis based on X-ray is one of the most important research directions in smart healthcare. The common applications of X-ray include lesion segmentation [1], detection [2], disease prediction [3], and medical report generation [4]. Previous works usually focus on a single task based on pre-trained backbone networks on ImageNet dataset [5] which are deep convolutional neural networks like VGG [6], ResNet [7], etc. Early deep learning methods greatly accelerated the development of medical image analysis but gradually reached their performance bottleneck. Potential reasons include privacy issues with medical data leading to scarcity, experienced doctors being unable to provide high-quality annotated data due to various reasons, limitations of the receptive field of CNN models, etc.

Inspired by the success of self-attention based Transformer [8] and self-supervised learning [9] in the natural language processing community, the researchers also designed new architectures for the perceptron of image/video data. Specifically, the ViT [10] and Swin-Transformer network [11] stand out as top contenders, the self-supervised learning strategies such as reconstruction-based masked auto-encoder [12], and contrastive learning [13], all sparking a new wave of research in the academic community. Naturally, these methods and techniques have also been introduced into the field of X-ray image analysis, and some progress and achievements have been made. To be specific, Wu et al. propose MedKLIP [14] which uses the paired image-text reports (about 227k studies from the MIMIC-CXR v2 dataset) for domain-specific knowledge extraction to enhance the medical image-language pre-training. Chen et al. propose to bridge the fusion-encoder and dual-encoder type for medical vision-text pre-training via PTUnifier [15]. Multi-Modal Masked auto-encoder (M^3AE) is proposed by Chen et al. [16] which attempt to learn cross-modal domain knowledge in a self-supervised learning manner by reconstructing missing pixels and tokens from randomly masked images and texts. Xiao et al. [17] verified that the pre-training of ViT on 266,340 chest X-rays using MAE can achieve better results than CNN model DenseNet-121 using the MoCo v2 framework. However, it is easy to find that the challenging X-ray image based tasks are still far from being solved well.

According to our observation, reflection, and discussions and consultations with senior experts in the medical field, we believe that these models may still be limited by the following factors:

- **The conflict between the high resolution of X-ray images and the standard resolution of pre-trained models used in natural images**: Specifically, existing models are typically trained on standard image resolutions, such as 224×224, while actual X-ray image resolutions may reach levels as high as 2000×3000. This discrepancy can lead to a significant loss of image information originally present in high-resolution data when down-sampling occurs.
- **Existing works rarely take into account the contextual prior information of chest X-ray images**: Standard

• Xiao Wang, Yuehang Li, Bo Jiang, and Jin Tang are with Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui University, Hefei 230601, China; Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University, Hefei 230601, China; School of Computer Science and Technology, Anhui University, Hefei 230601, China. (email: {xiaowang, tangjin}@ahu.edu.cn)
• Wentao Wu, Jiandong Jin are with the School of Artificial Intelligence, Anhui University, Hefei 230601, China.
• Chuanfu Li is with the First Affiliated Hospital of Anhui University of Chinese Medicine, Hefei 230022, China (email: licf@ahtcm.edu.cn).
• Corresponding author: Bo Jiang (email: jiangbo@ahu.edu.cn)

MAE pre-trained models utilize random sampling strategies for token masking, with some works [12] considering the impact of different mask rates on the results, while overlooking the importance of differences between the inside and outside of the chest contour.

- **Current X-ray based large-scale models have only been pre-trained on small-scale datasets.** Existing X-ray image based pre-trained models are usually pre-trained on public datasets, such as MIMIC [18] which contains about 300K images only. This level of data might not be abundant for large-scale model pre-training.

Considering the aforementioned issues, it is natural to raise the following questions: "*Can we further expand the scale of X-ray images for the pre-training task? Is it possible to conduct this pre-training task at higher resolutions? Could more prior information be leveraged to improve the effectiveness of pre-training?*"

In this paper, we answer these questions by proposing a novel MAE framework that adopts the context-aware masking strategy pre-trained on massive (about 1 million) high-definition X-ray images. As shown in Fig. 1, we resize the input X-ray image into 1280×1280 and believe the high-definition images can better preserve the detailed information of the original data. Then, we partition it into non-overlapping image patches and project the patches into token representations. The input tokens are dropped out with a high ratio (larger than 70%) by following the MAE [12], however, we propose a novel context-aware masking strategy instead of random masking used in MAE. Because the chest X-ray images contain prominent contour line information, and typically, doctors are more concerned with information about lesions in the chest area. The visible tokens are added with position encodings, then, we feed them into a Transformer encoder. The masked tokens are randomly initialized and concatenated with the output of the Transformer encoder and fed into the Transformer decoder network for masked region reconstruction. After the pre-training phase is finished, we extract the Transformer encoder as the backbone network for downstream tasks to validate its effectiveness in these tasks, including Chinese/English report generation and disease prediction.

To sum up, the contributions of this paper can be summarized as the following three aspects:

- We propose the first pre-trained foundation model using high-definition X-ray images (1280×1280), based on the masked auto-encoder framework.
- We exploit a new context-aware masking strategy for the X-ray image based masked auto-encoder framework.
- We conduct extensive experiments on multiple downstream tasks, including Chinese/English medical report generation and disease classification.

The rest of this paper is organized as follows: we first introduce the related works in section II by reviewing pre-training on the medical images, and downstream tasks. In section III, we focus on describing our proposed framework, including an overview, pre-training stage, and downstream tasks. In section IV, we conduct extensive experiments to validate the effectiveness of our model from both qualitative and quantitative views. Finally, we conclude this paper and propose future works in section V.

II. Related Works

In this section, we will introduce the algorithms mostly related to ours, including Pre-training on Medical Images and Downstream Tasks. More works can be found in the following surveys [19]–[23].

A. Pre-training on Medical Image

The pre-training techniques proposed for the medical image analysis can be categorized into two main streams, i.e., the *contrastive learning* based [14], [24]–[27] and *masked token based reconstruction* schemes [16], [28], [29]. A brief summary of these models is provided in Table I. For the masked token based reconstruction frameworks, the inputs are usually masked with a high ratio and attempt to reconstruct them using an encoder-decoder network. Specifically, Chen et al. [16] propose the M^3AE which takes the masked medical image and language as the input and conducts feature-level fusion using cross-attention. They propose two independent decoders for the reconstruction of vision and language modality. Zhou et al. [29] introduce the MRM (masked record modeling) framework which reconstructs masked image patches and masked report tokens following a multi-task scheme. Xiao et al. [17] pre-train a foundation model based on masked autoencoder [12] and conduct extensive ablation studies on the advantages between ViT and CNN.

For the contrastive learning based models, the relations between the medical images and reports are mainly considered for pre-training. Specifically, GLoRIA [24] proposed by Huang et al. which is an attention-based framework for learning global and local representations by contrasting image sub-regions and words in the paired report. G2D [25] (Global to Dense level representation learning) proposed by Liu et al. is a medical vision-language pre-training framework that improves the granularity and more accurate grounding for the learned features. Zhan et al. propose the UniDCP [26] which is a unified model and supports multiple medical fine-tuning tasks. They design cross-modal prompts to harmonize heterogeneous inputs from multiple pre-training tasks. Wang et al. propose the PhenotypeCLIP [27] which also follows the contrastive learning framework and learns more fine-grained phenotype-based representations to bridge the gap between vision and language efficiently. CXR-CLIP [30] first generates image-text pairs from image-label datasets via prompt engineering and conducts pre-training using three kinds of contrastive losses. Different from these works which conduct pre-training on low-resolution X-ray images or contrastive learning between X-ray and English text, in this work, we propose a pre-trained foundation model on high-definition X-ray images and support medical report generation in both English and Chinese.

B. Downstream Tasks

In this paper, we focus on handling two representative tasks, including *medical report generation* [41]–[48] and *disease*

No.	Name	Publication	Pre-train Data	Backbone	Modality	Pre-train paradigm	Downstream Tasks	URL
01	ARL [31]	ACMMM-2022	ROCO, MedICaT, MIMIC-CXR 771k(224×224)	CLIP-ViT-B + RoBERTa-base	Image-Text	ARL	VQA, DP, retrieval	GitHub
02	M^3AE [16]	MICCAI-2022	ROCO, MedICaT 298k(256×256)	CLIP-ViT-B + RoBERTa-base [	Image-Text	M^3AE	VQA, DP, retrieval	GitHub
03	MedKLIP [14]	ICCV-2023	MIMIC-CXR 377k(224×224)	ResNet-50 + ClinicalBERT	Image-Text	KLIP	DP, Seg	GitHub
04	Medical_MAE [17]	WACV-2023	ChestX-Ray14, CheXpert, MIMIC-CXR (256×256)	ViT-S	Image	MAE	DP	GitHub
05	ECAMP [32]	arXiv-2023	MIMIC-CXR 377k(448×448)	ViT	Image-Text	MAE	DP	GitHub
06	MRM [33]	ICLR-2023	MIMIC-CXR 377k(224×224)	ViT	Image-Text	MAE	DP	GitHub
07	G2D [25]	arXiv-2023	MIMIC-CXR 213k(256×256)	ResNet-50 + ClinicalBERT	Image-Text	G2D	DP, Seg, Detect	-
08	UniDCP [26]	arXiv-2023	ROCO, MIMIC-CXR 458k(224×224)	CLIPViT	Image-Text	UniDCP	VQA, RG, DP, Seg, retrieval	-
09	T3D [34]	arXiv-2023	BIMCV 8k($96 \times 96 \times 96$)	SwinUNTER + RadBERT	3D Volume-Text	T3D	DP, Seg	-
10	PhenotypeCLIP [27]	ACL-2023	CheXpert	ResNet-50 + BERT	Image-Text	PhenotypeCLIP	RG	-
11	MaCo [35]	arXiv-2023	MIMIC-CXR	ViT-B + BERT	Image-Text	MAE	DP, Seg	-
12	CXR-CLIP [30]	MICCAI-2023	MIMIC-CXR, CheXpert, ChestX-ray14 528k(224×224)	ResNet-50	Image-Text	CLIP	DP, retrieval	GitHub
13	PTUnifier [36]	ICCV-2023	ROCO, MedICaT, MIMIC-CXR 437k(288×288)	CLIP-ViT-B + RoBERTa-base	Image-Text	PTUnifier	VQA, RG, DP, retrieval	GitHub
14	MPMA [28]	TMM-2023	ROCO, MIMIC-CXR 458k(224×224)	ViT + BERT	Image-Text	MPMA	DP, RG, VQA	-
15	IMITATE [37]	arXiv-2023	MIMIC-CXR 377k(224×224)	ResNet-50 + Bio-ClinicalBERT	Image-Text	IMITATE	DP, Seg, Detect	-
16	MeDSLIP [38]	arXiv-2024	MIMIC-CXR 377k(224×224)	ResNet-50 + Bio-ClinicalBERT	Image-Text	ProtoCL	DP, Seg,	-
17	ASG [39]	arXiv-2024	MIMIC-CXR 377k(224×224)	ResNet50 / ViT-B + BioClinicalBERT	Image-Text	ASG	DP, Seg	-
18	MLIP [40]	arXiv-2024	MIMIC-CXR 377k(224×224)	ViT-B + BioClinicalBERT	Image-Text	MLIP	DP, Seg,	-
	Ours	-	**1M** (1280×1280)	ViT	Image	MAE	RG, DP	GitHub

TABLE I: Comparison between our model and existing X-ray based pre-trained foundation models. RG and DP are short for Report Generation and Disease Prediction, respectively.

classification [49]–[54]. For the report generation, Stephanie et al. propose the MAIRA-1 [41] which combines CXR-specific image encoder and fine-tuned LLM based on Vicuna-7B [42]. Li et al. [43] propose a Knowledge-driven Encode, Retrieve, Paraphrase (KERP) method. At its core is the proposed universal implementation unit-Graph Transformer (GTR), which dynamically transforms high-level semantics across various domains of graph-structured data, including knowledge graphs, images and sequences. ORGAN [44] proposed by Hou et al. is an observation-guided radiology report generation framework. It first generates an observation plan, feeds both the plan and images into the report generation process, and uses an observation graph and a tree-based reasoning mechanism to accurately enrich the information of each observation by capturing multiple formats. Liu et al. [45] proposed an unsupervised model called Knowledge Graph Auto-Encoder (KGAE), which comprises a pre-constructed knowledge graph, a knowledge-driven encoder and a knowledge-driven decoder. In the absence of paired image-report training data, unsupervised KGAE can generate desirable medical reports. Wang et al. [46] explore cross-modal feature interactions and propose a Cross-modal PROtotype driven NETwork (XPRONET) to promote cross-modal pattern learning and leverage it to enhance the task of radiology report generation. Zhang et al. [47] utilizes the Graph-guided Hybrid Feature Encoding (GHFE) module to encode the intrinsic relationships between pathological changes into graph embeddings using prior disease knowledge graphs. The GHFE combines graph embeddings, semantic embeddings, and visual features to form hybrid features, which are then fed into the decoder of a Transformer to generate reports. PromptMRG [48] proposed by Jin et al. which converts the diagnostic results of disease classification branches into prompts to guide report generation. By utilizing cross-modal retrieval and dynamic feature aggregation, it further enhances diagnostic accuracy. Different from existing medical report generation works, we propose a novel context-aware masking strategy for Chinese/English medical report generation.

For the disease classification, Li et al. [49] propose the Unify, Align, and Refine (UAR) method, which introduces the Latent Space Unifier, Cross-modal Representation Aligner and Text-to-Image Refiner to learn multi-level cross-modal alignments. The authors of [50] present a prototype representation learning framework that integrates global and local alignment between medical images and reports. By constructing a sentence-wise prototype memory bank, the network can focus on low-level local visual features and high-level clinical language features. BoMD [51] proposed by Chen et al. is a method designed for learning noisy multi-label CXR by detecting and re-labeling noisy samples from the dataset in a smooth manner. It optimizes a set of multi-label descriptors to promote their similarity with the semantic descriptors generated by a multi-label image annotation language model. Tanida et al. [52] propose a method that focuses explicitly on highlighted anatomical regions through object detection and generates descriptions for specific regions. Its interactive functionality allows radiologists to directly participate in the decision-making process. Kiut [53] proposed by Huang et al. is designed to learn multi-level visual representations. It incorporates a u-connection schema to simulate interactions between different modalities and has developed a symptom

graph and an injectable knowledge distiller to assist in report generation. Wang et al. [54] introduced multiple learnable *expert* tokens to encourage these expert tokens to capture complementary information through orthogonal loss. Each participating expert token guides the cross-modal attention between input words and visual tokens, enhancing the quality of generated reports. In contrast to existing disease prediction approaches, we employ a foundation model pre-trained on high-resolution X-ray images and have developed a novel context-aware masking strategy based on an X-ray image masked auto-encoder framework.

III. Our Proposed Approach

In this section, we will first give an overview to help better understand our pre-training framework. Then, we dive into the details of pre-training on high-definition X-ray images. After that, we will introduce the downstream tasks used to validate the effectiveness of our pre-trained model, including Chinese/English report generation and disease prediction.

A. Overview

As illustrated in Fig. 1, our pre-training scheme follows the masked auto-encoder framework [12] that attempts to reconstruct the highly masked input patches. Note that, we take the high-definition X-ray images as the input to better retain its raw detailed information. The image is partitioned into non-overlapping regions and transformed into token representations using a convolutional layer. Inspired by the fact that the cue in the chest part may be more important than other regions, therefore, we introduce a simple but effective context-aware masking strategy by masking more patches inside the chest regions. We believe this will help the model to focus on these regions in the pre-training phase. The visible tokens are fed into the ViT encoder and the outputs are concatenated with the randomly initialized masked tokens. Then, a Transformer decoder network is adopted to reconstruct the input image. Once the pre-training is finished, we fine-tuning the Transformer encoder for the downstream task to achieve a higher performance. More details will be introduced in subsequent sub-sections, respectively.

B. Pre-training Stage

In this section, we will focus on the details of our pre-training from the perspective of Input Processing, Context-Aware Masking, Transformer Encoder and Decoder, and Loss Function.

Input Processing. Given the raw X-ray images, we first resize them into a fixed-resolution $\mathcal{I} \in \mathbb{R}^{H\times W\times 1}$, here, we set both H and W as 1280. Following the Transformer encoder used in MAE [12], we partition the whole image into N non-overlapping regions $\{P_1, P_2, ..., P_N\}$, and the resolution of each region is $64\times 64\times 1$. Then, we adopt a convolution layer (kernel size 64×64) to transform the image patches into the token representations $\{X_1, X_2, ..., X_N\}$ whose dimension is 1024. After we get these tokens, the MAE framework masks them with a high ratio, and the rest of them are treated as visible tokens. For example, He et al. [12] mask 70% of them for the natural image, and Xiao et al. [17] claim that the best downstream performance can be achieved when removing 90% of them in the pre-training stage of their MAE-based X-ray model. However, seldom of they consider the context information of chest X-ray images for the masking operation.

Context-Aware Masking. In this work, we propose a novel context-aware masking strategy to process the transformed tokens. The key insight is that more useful cues can be mined in the chest region of X-ray image. Specifically, we manually define a boundary line of the chest, as shown in Fig. 1, and mask the tokens inside of the chest line with a higher probability. Once we remove the masked tokens, the visible tokens $X_i, i \in \{1, 2, ..., M\}$ are added with position encodings $E_i, i \in \{1, 2, ..., M\}$, therefore, we feed the $\mathcal{X}_i = X_i + E_i, i \in \{1, 2, ..., M\}$ into the Transformer encoder network.

Transformer Encoder and Decoder. In our practical implementation, we adopt the ViT-L [10] (16 heads and 1024 embedding dimension, 304M trainable parameters) as the Transformer encoder network which contains 24 Transformer blocks. The key operator in the Transformer is multi-head self-attention and the detailed computing process of self-attention can be formulated as:

$$SelfAttenion = Softmax(\frac{QK^T}{\sqrt{c}})V \tag{1}$$

where Q, K and V are processed input tokens $\mathcal{X}_i$, c is the dimension of input tokens. $Softmax(\cdot)$ denotes the Softmax layer.

Given the output of Transformer encoder, we integrate them with the masked tokens, whose parameters are randomly initialized, and feed into the Transformer decoder network. For the detailed network architecture of the Transformer decoder, we directly borrow from the vanilla MAE framework for the masked X-ray image reconstruction. It contains 8 Transformer blocks.

Loss Function. After we obtain the reconstructed image patches, we compute its distance with the ground truth image patch using the L_2 loss function. Note that, existing work [17] demonstrates that other loss functions like L_1, *smooth-*L_1, *SSIM*, and *adversarial loss* do not improve the MAE framework.

C. Downstream Tasks

In this work, we validate our proposed framework by introducing the pre-trained Transformer encoder (i.e., ViT-L) into two downstream tasks, including X-ray based report generation and disease prediction, as illustrated in Fig. 1 (b) and (c).

X-ray based Report Generation. Given the X-ray image, the task of report generation targets describing the disease information using natural language. Usually, this task is formulated as an English sentence generation. In addition, we also build a new X-ray dataset for Chinese report generation. More details about this dataset will be introduced in section IV-A.2. In our implementation, we build our report generator based on the R2Gen[1] toolkit proposed in [55].

[1] https://github.com/zhjohnchan/R2Gen

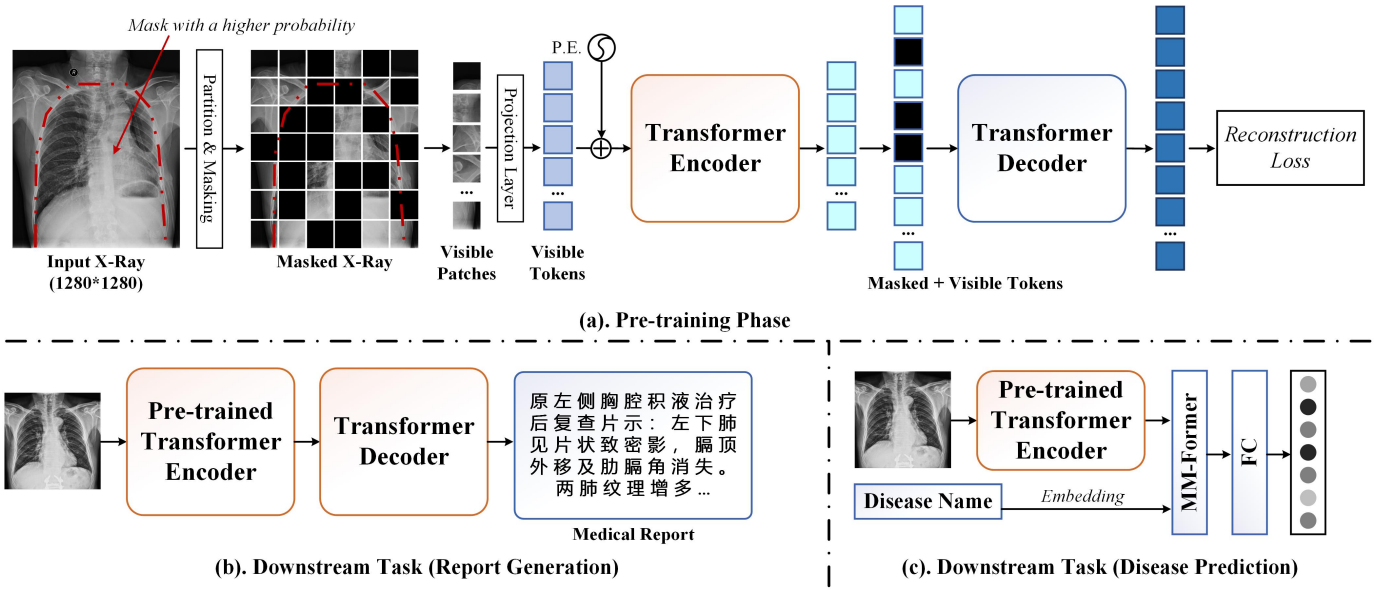


Fig. 1: (a) An illustration of our proposed high-definition X-ray image based pre-training framework using masked auto-encoder. (b, c) are two downstream tasks used for the validation of our pre-training framework.

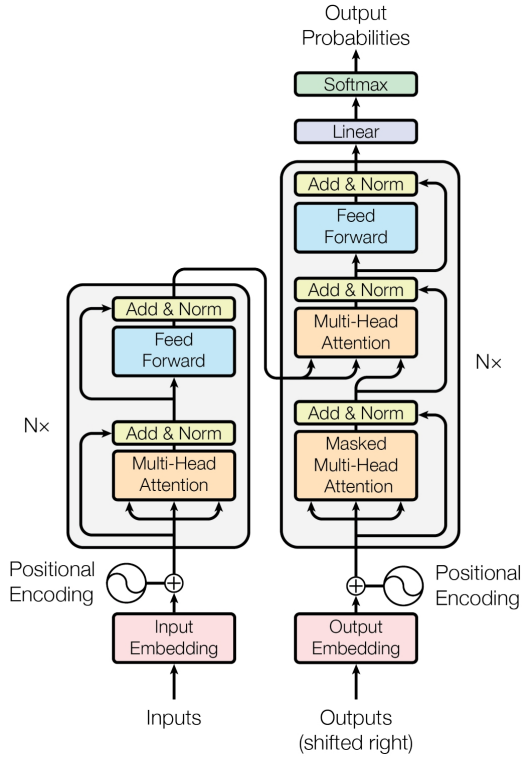


Fig. 2: The detailed architectures of Transformer from [8].

Disease Prediction. This task can be treated as a standard multi-category classification problem by mapping the input X-ray image into a distribution of the response score of each category. However, this may ignore the semantic information of category names which is also useful for high-performance recognition. In this work, we follow the VTB[2] [56], which formulates the multi-label classification task as a vision-text fusion problem, for the disease prediction. As shown in Fig. 1 (c), we adopt the pre-trained ViT encoder to extract the features of the input X-ray image and utilize the CLIP text encoder to embed the given disease name. Then, we fuse the two modalities using a multi-modal Transformer network and predict the disease using a fully connected (FC) layer.

[2]https://github.com/cxh0519/VTB

IV. Experiments

In this section, we will first introduce the datasets, evaluation metrics, and implementation details in sub-section IV-A, IV-B, IV-C, respectively. Then, we will focus on reporting and analyzing the results of the medical report generation and disease prediction, in sub-section IV-D, IV-E. After that, we will give extensive ablation studies of our model in sub-section IV-F and visualize the reconstruction on the masked tokens, similar matric, generated medical reports, and disease predictions in sub-section IV-G. Also, we describe the limitations of this work in sub-section IV-H.

A. Datasets

1) Pre-training Dataset: To pre-train a high-performance X-ray foundation model, the first thing we need to do is the collection of large-scale X-ray images. Therefore, a large-scale and high-resolution dataset that contains $1,053,791$ X-ray medical images is collected for the pre-training. Some representative samples are visualized in Fig. 3.

2) Downstream Datasets: We conduct extensive experiments on medical report generation task and disease prediction task, and the involved datasets including **IU-Xray** [57], our Private Chinese Chest X-ray image based report generation dataset (termed **PCC-Xray** in this paper), and **RSNA-Pneumonia** [58] dataset. A brief introduction to these datasets is given below.

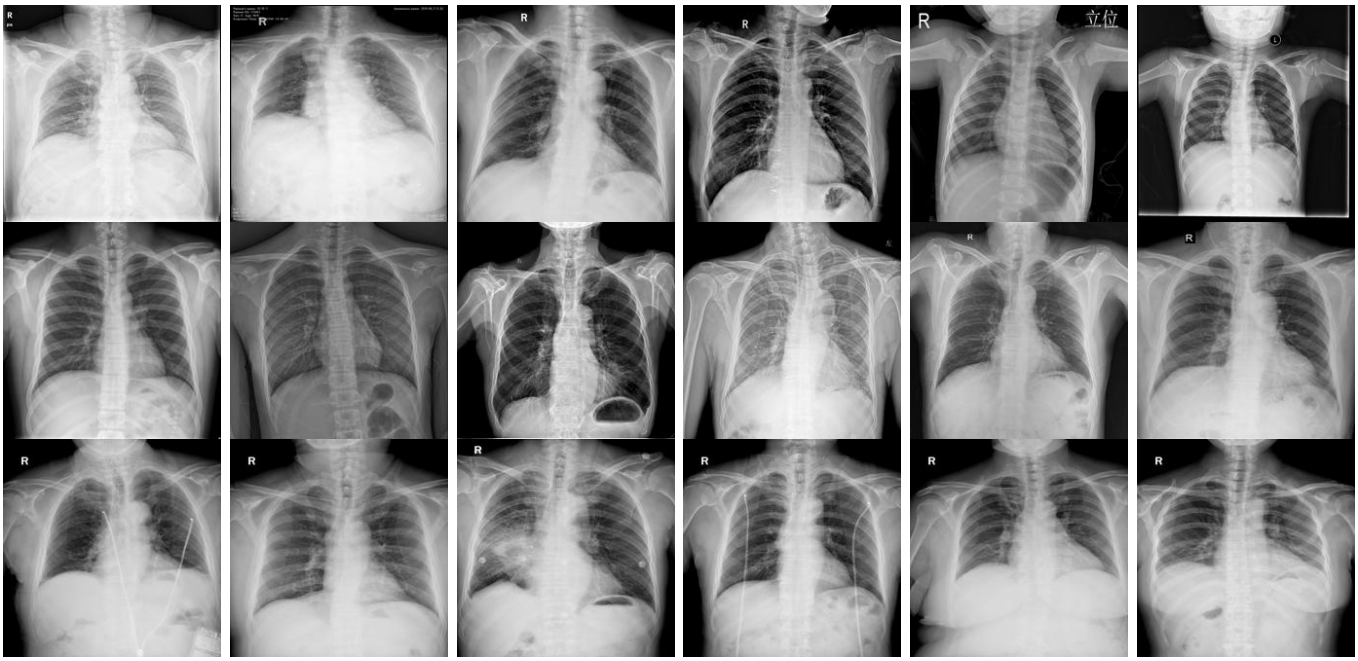


Fig. 3: Some representative samples of our collected PCC-Xray dataset.

Fig. 4: The word cloud of our newly collected PCC-Xray dataset for Chinese medical report generation.

• [*English Report Generation*] IU-Xray dataset [57] is a widely used dataset for the evaluation of radiology reporting systems. It contains *7,470* chest X-ray images and corresponding *3,955* reports. Following previous works [57], in our experiments, we utilize the processed dataset obtained by excluding samples with one image only. Specifically, our training, validation, and testing subset contains *2069/296/590* samples, respectively.

• [*Chinese Report Generation*] PCC-Xray dataset: It was built by the First Affiliated Hospital of Anhui University of Chinese Medicine and Anhui University which contains $200,172$ high-resolution chest X-ray images. Each X-ray image is meticulously annotated with a Chinese medical report, with an average of *71* Chinese characters per sentence. Regarding the Chinese medical reports of interest, we provide a word cloud as shown in Fig. 4. It can be observed that our reports cover descriptions of many common diseases and some difficult and miscellaneous conditions. We split it into the training, validation, and testing subset which contains *140120/20018/40034* X-ray image and report pairs, respectively.

• [*Disease Prediction*] RSNA-Pneumonia dataset [58] comprises *30k* frontal view chest radiographs, each accompanied by bounding boxes indicating pneumonia opacities if present. We follow the official data split, which includes training/validation/testing sets consisting of *25184/1500/3000* samples, respectively.

B. Evaluation Metrics

For the *X-ray report generation task*, we adopt the widely used four metrics for the evaluation, including **CIDEr** (Captions Generated by Diverse Experts) [59], **BLEU-4** (Bilingual Evaluation Understudy-4) [60], **ROUGE-L** (Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence) [61], and **METEOR** (Metric for Evaluation of Translation with Explicit Ordering) [62]. Specifically, CIDEr measures the consensus between the generated captions and multiple reference captions. It evaluates the quality of image captioning by computing the cosine similarity between n-grams in the generated caption and those in the reference captions. BLEU-4 evaluates the quality of machine-generated translations or text summaries by comparing them against reference translations or summaries. It measures the precision of n-grams (usually up to 4-grams) in the generated text compared to the reference texts. ROUGE-L assesses the quality of

text summaries or translations by comparing them to reference texts. It focuses on the longest common subsequences between the generated and reference texts, emphasizing recall. METEOR evaluates machine-generated translations or summaries by considering both unigram precision and recall, as well as the alignment between the generated and reference texts. It also incorporates stemming and synonymy matching.

For the *disease prediction task*, in this work, we adopt the **AUROC** (Area Under the Receiver Operating Characteristic Curve), **F1**, and **Accuracy** metrics to compare the performance of different models. To be specific, AUROC is a metric commonly used to evaluate binary classification models. It measures the ability of the model to distinguish between positive and negative samples across different decision thresholds. The F1 score is the harmonic mean of precision and recall. It's commonly used in binary classification tasks to provide a single metric that balances both precision and recall. Accuracy simply measures the proportion of correctly classified instances out of the total instances evaluated.

C. Implementation Details

• **Pre-training Stage.** In the pre-training phase, we resize the X-ray image into a fixed resolution, i.e., 1280×1280. The learning rate is set as 0.00025, and the weight decay is 0.04. The batch size is 1024 and training for a total of 83 epochs on our dataset. The AdamW [63] is adopted as the optimizer. The pre-training is conducted on a server with eight NVIDIA A800 GPUs (80GB) and about 660 hours are needed for our pre-training phase.

• **Downstream Tasks.** For the generation of Chinese medical reports, we fine-tune the model using our PCC-Xray dataset, which comprises a total of 200,000 X-ray images. The training configuration is as follows: we resize the images to 224×224, set the batch size to 16, employ RoBERTa [64] as the tokenizer, and conduct training over 60 epochs. In the case of English medical report generation, we fine-tune the model on the IU-Xray dataset. Before inputting the images into the model, we resize them to 384×384. We set the batch size to 16, establish a maximum sequence length of 30, and keep other parameters identical to R2Gen.

For the disease prediction task, we performed experiments on the RSNA-Pneumonia dataset using code derived from the Visual-Textual Baseline (VTB). We configured the batch size to 200 and set the input image size to 224×224, while leaving the remaining configurations unchanged.

D. Results of Medical Report Generation

In this sub-section, we conduct extensive experiments on the Chinese/English report generation and compare it with current state-of-the-art report generators.

1) Chinese Report Generation: As shown in Table II, the baseline R2Gen+MAE achieves $0.660, 0.588, 0.536, 0.498, 0.594$ on our newly collected PCC-Xray dataset on the BLEU-1, BLEU-2, BLEU-3, BLEU-4, and ROUGE-L metrics, meanwhile, the results can be improved to $0.679, 0.609, 0.560, 0.523, 0.611$ when using the ViT backbone network pre-trained on our X-Ray dataset using the masked auto-encoder framework. This comparison demonstrates that pre-training on large-scale X-ray images indeed helps feature representation learning more than on natural images. When the context-aware masking strategy is adopted in the pre-training, the results can be further improved to $0.719, 0.656, 0.611, 0.577, 0.651$. It is easy to find that the context-aware masking works for our X-ray based pre-training foundation model.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE_L
R2Gen+MAE	0.660	0.588	0.536	0.498	0.594
Ours	0.697	0.631	0.583	0.548	0.635
Ours (w/o CaM)	0.694	0.627	0.579	0.543	0.628

TABLE II: Experimental results on the PCC-Xray dataset. w/o denotes without the following item.

Methods	CIDEr	BLEU-4	ROUGE-L	METEOR
R2Gen [55]	0.398	0.165	0.371	0.187
KERP [43]	0.280	0.162	0.339	-
HRGP [66]	0.343	0.151	0.322	-
MKG [67]	0.304	0.147	0.367	-
PPKED [68]	0.351	0.168	0.376	0.190
MGSK [65]	0.382	0.178	0.381	-
CA [69]	-	0.169	0.381	0.193
CMCL [70]	-	0.162	0.378	0.186
DCL [57]	0.586	0.163	0.383	0.193
Ours	**0.677**	**0.185**	**0.395**	**0.197**
Ours (w/o CaM)	0.611	0.171	0.375	0.195

TABLE III: The performances of our proposed model compared with other state-of-the-art systems on IU-Xray dataset. The best results in each column are highlighted in bold. w/o denotes without the following item.

2) English Report Generation: As shown in Table III, we also adapt our framework to handle the English medical report generation task. In this section, we report the experimental results on the IU-Xray dataset which is 0.677, 0.185, 0.395, 0.197 on the CIDEr, BLEU-4, ROUGE-L, METEOR metric, respectively. Compared with our baseline method R2Gen [55] which obtains 0.398, 0.165, 0.371, 0.187 on these metrics, our model achieves a significant improvement. Compared with other models, such as DCL [57] and MGSK [65], our results are also better than theirs. These results fully validated the effectiveness of our proposed foundation model for the perception of X-ray images.

E. Results of Diseases Recognition

As shown in Table IV, we report our experimental results on the disease prediction task and compare them with recent state-of-the-art recognition models. Obviously, our recognition results are comparable to these models but still inferior to them. We think this may be caused by the fact that our model is pre-trained on high-resolution X-ray images, but the images in RSNA-Pneumonia dataset [58] are the standard resolution. In our future works, we will consider pre-training on multi-scale X-ray images to further improve the generation and robustness of our model. Another possible reason is that the VTB is proposed for pedestrian attribute recognition, the parameter configurations may be different from the disease recognition task.

Backbone	Method	RSNA-Pneumonia (AUC)		
		1%	10%	100%
CNN	GLoRIA [24]	86.1	88.0	88.6
	PRIOR [50]	85.7	87.1	89.2
	MedKLIP# [14]	87.3	88.0	89.3
	KAD [71]	89.8	91.8	92.5
Transformer	MAE [12]	84.2	89.6	91.3
	REFERS [72]	89.4	91.6	92.7
	MGCA# [73]	90.7	92.6	93.4
	MRM [33]	91.3	92.7	93.3
	ECAMP [32]	91.5	92.9	93.8
	Ours	83.4	86.3	88.2
	Ours (w/o CaM)	46.3	83.7	86.9

TABLE IV: Results of disease recognition on RSNA-Pneumonia dataset. Methods with # leverage disease-level annotations. w/o denotes without the following item.

Resolution	BLEU-4	METEOR	ROUGE-L	CIDEr
384×384	**0.185**	**0.197**	**0.395**	**0.677**
448×448	0.162	0.197	0.369	0.625
512×512	0.163	0.192	0.361	0.596

TABLE V: Performance on the IU-Xray test dataset using different input sizes.

F. Ablation Study

In this sub-section, we conduct extensive experiments to further help the readers better understand our framework.

1) Random Masking vs Context-aware Masking: To verify the effectiveness of our proposed Context-aware Masking (CaM, for short) for the X-ray based masked auto-encoder, we compare the performance of report generation with and without (w/o) the CaM, as shown in Table IV. With the help of CaM, we achieve 83.4, 86.3, 88.2 when using 1%, 10%, and 100% of the training data of RSNA dataset, but we only get 46.3, 83.7, and 86.9 on these settings when removing this module, i.e., Ours (w/o CaM). This comparison fully validated the effectiveness and importance of the context-aware masking strategy. Similar conclusions can also be drawn from Table II and Table III.

2) Does High Definition X-ray Image Works for Report Generation?: As shown in Table V, we set different resolutions of X-ray images to test its influence on the final results, including 384×384, 448×448, and 512×512. We can find that better results can be obtained when the resolution is set as 384×384, i.e., 0.185, 0.197, 0.395, 0.677 on the BLEU-4, METEOR, ROUGE-L, CIDEr metric. This result is consistent with current vision models which can achieve higher performance when slightly increasing the resolution from 224×224. However, the performance drops when further increasing the resolution, as reported in [74], [75].

3) The Curve of Relationship between Epoch and Accuracy: As shown in Table VI and Fig. 5, we report the corresponding results on the IU-Xray testing subset in the pre-training phase, i.e., 20^{th}, 40^{th}, 60^{th}, 80^{th}, and 83^{th} epoch. Generally speaking, better results can be obtained in the late stage of our pre-training.

4) Influence on the Maximum Length of Medical Report Predicted by the Report Generator: As shown in Fig. 6, we give a visualization of the distribution of the number of words

Method	BLEU-4	METEOR	ROUGE-L	CIDEr
Epoch-20	**0.175**	0.206	0.373	0.662
Epoch-40	0.169	0.197	0.368	0.658
Epoch-60	0.169	0.220	0.382	**0.708**
Epoch-80	0.173	0.209	0.376	0.674
Epoch-83	0.168	**0.220**	**0.382**	0.706

TABLE VI: The detailed accuracy on the IU-Xray testing dataset in the training phase.

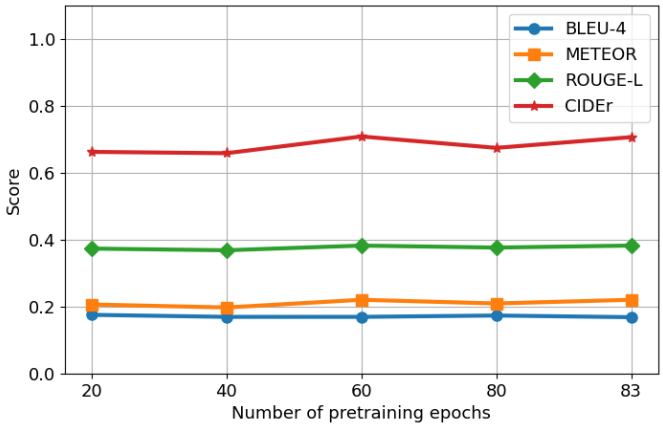


Fig. 5: Variation of the accuracy on the IU-Xray testing dataset in the training phase.

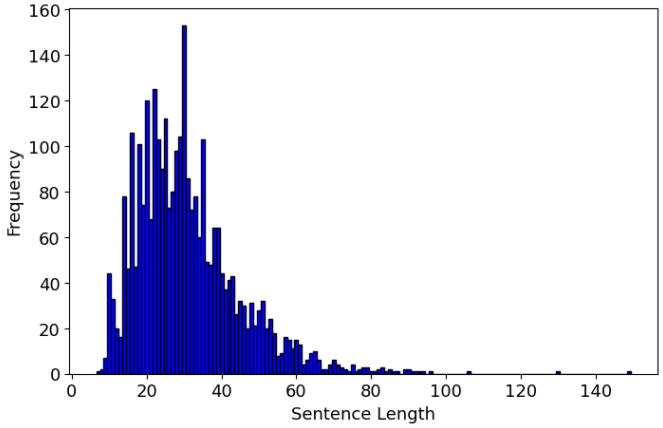


Fig. 6: Distribution of sentence length of IU-Xray dataset.

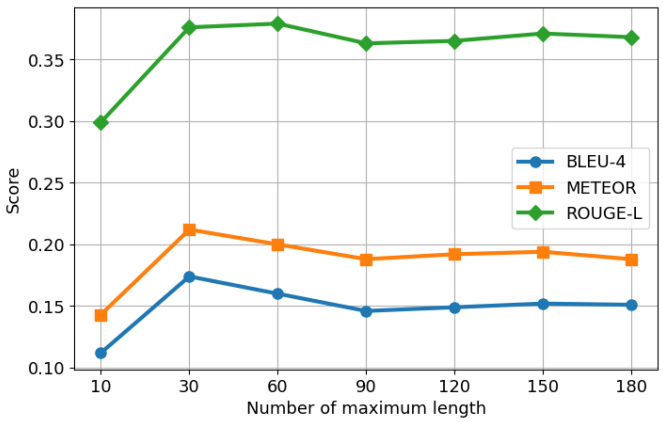


Fig. 7: Variation of different maximum length of report generator on the IU-Xray testing subset.

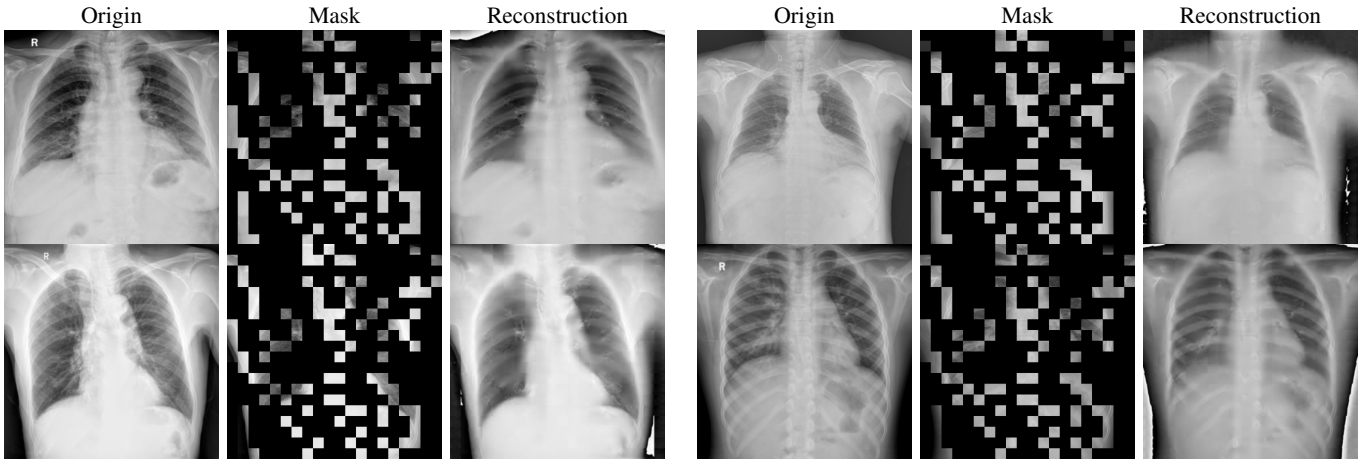


Fig. 8: Visualization of the reconstructed masked tokens on our newly collected PCC-Xray dataset.

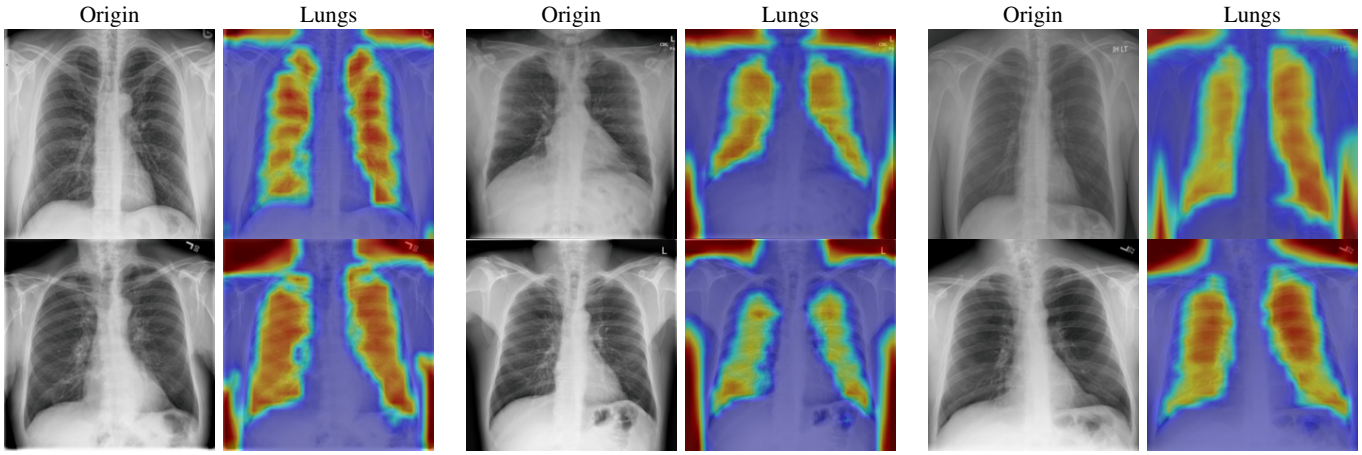


Fig. 9: Visualizations of activation response maps on the IU-Xray dataset.

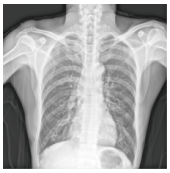
Image	Ground Truth	Ours
	胸廓对称，气管居中；两侧肋骨、肋间隙正常；两下肺纹理增多增粗，其间见斑点、小斑片状密度增高影，两侧肺门和纵隔影未见明显异常；主动脉结突出伴有钙化影，心影横径稍增大；膈肌平滑，双侧肋膈角锐利。	两胸廓对称，肋骨走形自然，未见畸形及骨质破坏。双肺纹理增多、增粗、紊乱。两肺未见明显实质性病变。心影横径增宽。双膈面尚光整，肋膈角尚锐利。
	两侧胸廓对称，两肺未见明显实质性病变，两侧膈面光滑，两侧肋膈角锐利。心影形态、大小未见明显异常。	两侧胸廓对称，两肺未见明显实质性病变，两侧膈面光滑，两侧肋膈角锐利。心影形态、大小未见明显异常。
	两肺纹理稍增多，可见散在分布斑点状高密度影，边界尚清，心影大小形态大致正常范围内，双侧膈肌光滑，肋膈角锐利。	两肺纹理增多，右上肺野见斑点、条状、结节状中等密度影，边界部分清晰。心影大小形态大致正常范围内，双侧膈肌光滑，肋膈角锐利。

Fig. 10: The X-ray image based medical report generation on our PCC-Xray dataset.

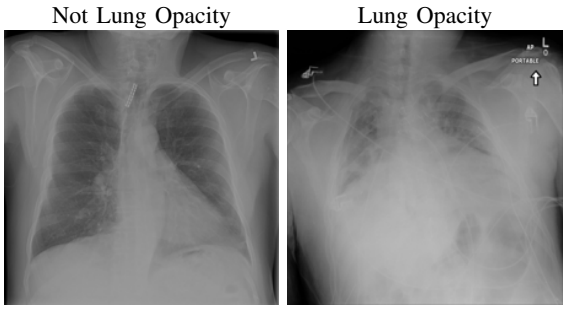


Fig. 11: Visualizations of lung opacity prediction in RSNA-Pneumonia dataset.

in each sentence on the IU-Xray dataset. We can find that most sentences contain about 20-40 words. Interestingly, as the results reported in Fig. 7, the peak results can be achieved when the maximum length of the medical report predicted by the report generator is set as 30. Therefore, we choose this hyper-parameter as 30 in this work for the IU-Xray dataset.

G. Visualization

In this sub-section, we give some visualizations to help the readers better understand the effectiveness of our model, including the reconstructed masked tokens (Section IV-G.1), the activation response maps (Section IV-G.2), generated medical reports (Section IV-G.3), and predicted diseases (Section IV-G.4).

1) Reconstructed Masked Tokens: As shown in Fig. 8, we provide some representative samples predicted by our model. The 1^{th} and 4^{th} column are the raw X-ray images, the 2^{th} and 5^{th} column are masked images, and the 3^{th} and 6^{th} column are the reconstructed images. We can find that our proposed context-aware masking strategy guided MAE framework predict the masked tokens well.

2) Activation Response Maps: As shown in Fig. 9, given the text *lungs*, we can find that the activation maps can accurately highlight the target regions. Therefore, we can achieve a higher performance on the downstream tasks. However, the activation maps are imperfect, as the background regions are also highlighted.

3) Medical Report Generation: In addition to aforementioned visualization on the reconstructed masked tokens and activation response maps, we also show the generated medical reports on the PCC-Xray dataset, as shown in Fig. 10. It is easy to find that our model performs well and accurately predicts the reports.

4) Disease Prediction: As shown in Fig. 11, given the X-ray image from the RSNA-Pneumonia dataset and all the labels (binary classification) we need to recognize, our model can predict the disease accurately.

H. Limitation Analysis

This work attempts to conduct self-supervised pre-training based on MAE (Masked Auto-Encoder) on a high-resolution X-ray dataset and validates it on two mainstream medical downstream tasks. The results indicate that our X-ray based model indeed achieves promising results. Experimental results fully demonstrate that Transformer-based big model frameworks can ensure decent results, but they still cannot achieve the astonishing performance boost seen in large language models. We think the current model can still be improved from the following perspectives: 1). The current framework adopts the Transformer as the core block, bringing a huge computation cost in the pre-training phase. 2). Only X-ray images are used in the pre-training phase which ignores the semantic cues, therefore, the overall performance may still sub-optimal. 3). Current mainstream backbone networks adopts 224×224 as their default resolution of the input image, however, the specific design to address the high-resolution images still further exploring.

V. Conclusion and Future Works

In this work, we summarize the issues of existing X-ray image based pre-training methods and propose to pre-training a high-definition foundation model. Specifically, we follow the self-supervised pre-training framework masked auto-encoder (MAE) and design a new context-aware masking strategy. For the downstream tasks, we test our model on both English/Chinese report generation and disease prediction. The experiments on multiple benchmark datasets fully validated the effectiveness of our model.

In future work, further improvements are still needed to pursue breakthroughs. Specifically speaking, 1). The X-ray based vision foundation model proposed in this paper is based on Transformer but has a complexity of $\mathcal{O}(N^2)$, resulting in high memory consumption and computational costs when handling high-resolution X-ray data. In the future, we will attempt to introduce new lightweight network architectures (such as State Space Model/Mamba [76]) to address its computational complexity issues. 2). Pre-training purely from a visual self-supervised manner can yield decent improvements, but the overall accuracy is still not satisfactory. Subsequently, we will consider multi-modal pre-training approaches, incorporating large language models, knowledge graphs, etc., to further enhance the representation ability of the visual foundation model. 3). In addition to pre-training, to enhance performance on downstream tasks (such as medical reports), we will explore the introduction of knowledge graphs or other useful prompts to improve its performance in text generation.

References

[1] O. Ronneberger, P.Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 9351. Springer, 2015, pp. 234–241.

[2] K. Yan, X. Wang, L. Lu, and R. M. Summers, "DeepLesion: Automated mining of large-scale lesion annotations and universal lesion detection with deep learning," *Journal of medical imaging*, vol. 5, no. 3, pp. 036 501–036 501, 2018.

[3] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," Dec. 2017.

[4] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," *arXiv preprint arXiv:1502.03044*, 2015.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Miami, FL: IEEE, Jun. 2009, pp. 248–255.

[6] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *International Conference on Learning Representations*, 2015.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[9] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *Proceedings of the 37th International Conference on Machine Learning*. PMLR, Nov. 2020, pp. 1597–1607.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ICLR*, 2021.

[11] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada: IEEE, Oct. 2021, pp. 9992–10 002.

[12] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, Jul. 2021, pp. 8748–8763.

[14] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 21 315–21 326.

[15] Z. Chen, S. Diao, B. Wang, G. Li, and X. Wan, "Towards unifying medical vision-and-language pre-training via soft prompts," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, oct 2023, pp. 23 346–23 356.

[16] Z. Chen, Y. Du, J. Hu, Y. Liu, G. Li, X. Wan, and T.-H. Chang, "Multi-modal masked autoencoders for medical vision-and-language pre-training," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 679–689.

[17] J. Xiao, Y. Bai, A. Yuille, and Z. Zhou, "Delving into masked autoencoders for multi-label thorax disease classification," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 3588–3600.

[18] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific data*, vol. 6, no. 1, p. 317, 2019.

[19] X. Wang, G. Chen, G. Qian, P. Gao, X.-Y. Wei, Y. Wang, Y. Tian, and W. Gao, "Large-scale multi-modal pre-trained models: A comprehensive survey," *Machine Intelligence Research*, pp. 1–36, 2023.

[20] P. Shrestha, S. Amgain, B. Khanal, C. A. Linte, and B. Bhattarai, "Medical vision language pretraining: A survey," *arXiv preprint arXiv:2312.06224*, 2023.

[21] B. Azad, R. Azad, S. Eskandari, A. Bozorgpour, A. Kazerouni, I. Rekik, and D. Merhof, "Foundational models in medical imaging: A comprehensive survey and future vision," *arXiv preprint arXiv:2310.18689*, 2023.

[22] Z. Zhao, Y. Liu, H. Wu, Y. Li, S. Wang, L. Teng, D. Liu, X. Li, Z. Cui, Q. Wang *et al.*, "Clip in medical imaging: A comprehensive survey," *arXiv preprint arXiv:2312.07353*, 2023.

[23] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature medicine*, vol. 29, no. 8, pp. 1930–1940, 2023.

[24] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung, "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*, 2021, pp. 3942–3951.

[25] C. Liu, C. Ouyang, S. Cheng, A. Shah, W. Bai, and R. Arcucci, "G2d: From global to dense radiography representation learning via vision-language pre-training," *arXiv preprint arXiv:2312.01522*, 2023.

[26] C. Zhan, Y. Zhang, Y. Lin, G. Wang, and H. Wang, "Unidcp: Unifying multiple medical vision-language tasks via dynamic cross-modal learnable prompts," *arXiv preprint arXiv:2312.11171*, 2023.

[27] S. Wang, B. Peng, Y. Liu, and Q. Peng, "Fine-grained medical vision-language representation learning for radiology report generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 15 949–15 956.

[28] K. Zhang, Y. Yang, J. Yu, H. Jiang, J. Fan, Q. Huang, and W. Han, "Multi-task paired masking with alignment modeling for medical vision-language pre-training," *IEEE Transactions on Multimedia*, 2023.

[29] H.-Y. Zhou, C. Lian, L. Wang, and Y. Yu, "Advancing radiograph representation learning with masked record modeling," in *The Eleventh International Conference on Learning Representations*, 2023.

[30] K. You, J. Gu, J. Ham, B. Park, J. Kim, E. K. Hong, W. Baek, and B. Roh, "Cxr-clip: Toward large scale chest x-ray language-image pre-training," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 101–111.

[31] Z. Chen, G. Li, and X. Wan, "Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 5152–5161.

[32] R. Wang, Q. Yao, H. Lai, Z. He, X. Tao, Z. Jiang, and S. K. Zhou, "Ecamp: Entity-centered context-aware medical vision language pre-training," 2023.

[33] H.-Y. Zhou, C. Lian, L. Wang, and Y. Yu, "Advancing radiograph representation learning with masked record modeling," in *The Eleventh International Conference on Learning Representations*, 2023.

[34] C. Liu, C. Ouyang, Y. Chen, C. C. Quilodrán-Casas, L. Ma, J. Fu, Y. Guo, A. Shah, W. Bai, and R. Arcucci, "T3d: Towards 3d medical image understanding through vision-language pre-training," 2023.

[35] W. Huang, H. Zhou, C. Li, H. Yang, J. Liu, and S. Wang, "Enhancing representation in radiography-reports foundation model: A granular alignment algorithm using masked contrastive learning," *arXiv preprint arXiv:2309.05904*, 2023.

[36] Z. Chen, S. Diao, B. Wang, G. Li, and X. Wan, "Towards unifying medical vision-and-language pre-training via soft prompts," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 23 403–23 413.

[37] C. Liu, S. Cheng, M. Shi, A. Shah, W. Bai, and R. Arcucci, "Imitate: Clinical prior guided hierarchical vision-language pre-training," 2023.

[38] W. Fan, M. N. I. Suvon, S. Zhou, X. Liu, S. Alabed, V. Osmani, A. Swift, C. Chen, and H. Lu, "Medslip: Medical dual-stream language-image pre-training for fine-grained alignment," 2024.

[39] Q. Li, X. Yan, J. Xu, R. Yuan, Y. Zhang, R. Feng, Q. Shen, X. Zhang, and S. Wang, "Anatomical structure-guided medical vision-language pre-training," *arXiv preprint arXiv:2403.09294*, 2024.

[40] J. Liu, H.-Y. Zhou, C. Li, W. Huang, H. Yang, Y. Liang, and S. Wang, "Mlip: Medical language-image pre-training with masked local representation learning," *arXiv preprint arXiv:2401.01591*, 2024.

[41] S. L. Hyland, S. Bannur, K. Bouzid, D. C. Castro, M. Ranjit, A. Schwaighofer, F. Pérez-García, V. Salvatelli, S. Srivastav, A. Thieme *et al.*, "Maira-1: A specialised large multimodal model for radiology report generation," *arXiv preprint arXiv:2311.13668*, 2023.

[42] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez *et al.*, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

[43] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Knowledge-driven encode, retrieve, paraphrase for medical image report generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6666–6673.

[44] W. Hou, K. Xu, Y. Cheng, W. Li, and J. Liu, "Organ: Observation-guided radiology report generation via tree reasoning," *arXiv preprint arXiv:2306.06466*, 2023.

[45] F. Liu, C. You, X. Wu, S. Ge, X. Sun *et al.*, "Auto-encoding knowledge graph for unsupervised medical report generation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 266–16 279, 2021.

[46] J. Wang, A. Bhalerao, and Y. He, "Cross-modal prototype driven network for radiology report generation," in *European Conference on Computer Vision*. Springer, 2022, pp. 563–579.

[47] K. Zhang, H. Jiang, J. Zhang, Q. Huang, J. Fan, J. Yu, and W. Han, "Semi-supervised medical report generation via graph-guided hybrid feature consistency," *IEEE Transactions on Multimedia*, 2023.

[48] H. Jin, H. Che, Y. Lin, and H. Chen, "Promptmrg: Diagnosis-driven prompts for medical report generation," *arXiv preprint arXiv:2308.12604*, 2023.

[49] Y. Li, B. Yang, X. Cheng, Z. Zhu, H. Li, and Y. Zou, "Unify, align and refine: Multi-level semantic alignment for radiology report generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 2863–2874.

[50] P. Cheng, L. Lin, J. Lyu, Y. Huang, W. Luo, and X. Tang, "Prior: Prototype representation joint learning from medical images and reports," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 361–21 371.

[51] Y. Chen, F. Liu, H. Wang, C. Wang, Y. Liu, Y. Tian, and G. Carneiro, "Bomd: bag of multi-label descriptors for noisy chest x-ray classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 284–21 295.

[52] T. Tanida, P. Müller, G. Kaissis, and D. Rueckert, "Interactive and explainable region-guided radiology report generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7433–7442.

[53] Z. Huang, X. Zhang, and S. Zhang, "Kiut: Knowledge-injected u-transformer for radiology report generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 809–19 818.

[54] Z. Wang, L. Liu, L. Wang, and L. Zhou, "Metransformer: Radiology report generation by transformer with multiple learnable expert tokens," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 558–11 567.

[55] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 1439–1449.

[56] X. Cheng, M. Jia, Q. Wang, and J. Zhang, "A simple visual-textual baseline for pedestrian attribute recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6994–7004, 2022.

[57] M. Li, B. Lin, Z. Chen, H. Lin, X. Liang, and X. Chang, "Dynamic graph enhanced contrastive learning for chest x-ray report generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3334–3343.

[58] G. Shih, C. C. Wu, S. S. Halabi, M. D. Kohli, L. M. Prevedello, T. S. Cook, A. Sharma, J. K. Amorosa, V. Arteaga, M. Galperin-Aizenberg *et al.*, "Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia," *Radiology: Artificial Intelligence*, vol. 1, no. 1, p. e180041, 2019.

[59] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[60] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[61] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[62] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[63] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*.

[64] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[65] S. Yang, X. Wu, S. Ge, S. K. Zhou, and L. Xiao, "Knowledge matters: Chest radiology report generation with general and specific knowledge," *Medical image analysis*, vol. 80, p. 102510, 2022.

[66] Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," *Advances in neural information processing systems*, vol. 31, 2018.

[67] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu, "When radiology report generation meets knowledge graph," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12 910–12 917.

[68] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, "Exploring and distilling posterior and prior knowledge for radiology report generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 753–13 762.

[69] F. Liu, C. Yin, X. Wu, S. Ge, P. Zhang, and X. Sun, "Contrastive attention for automatic chest x-ray report generation," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 269–280.

[70] F. Liu, S. Ge, and X. Wu, "Competence-based multimodal curriculum learning for medical 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)*, 2021, pp. 3001–3012.

[71] X. Zhang, C. Wu, Y. Zhang, W. Xie, and Y. Wang, "Knowledge-enhanced visual-language pre-training on chest radiology images," *Nature Communications*, vol. 14, no. 1, p. 4542, 2023.

[72] H.-Y. Zhou, X. Chen, Y. Zhang, R. Luo, L. Wang, and Y. Yu, "Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports," *Nature Machine Intelligence*, vol. 4, no. 1, pp. 32–40, 2022.

[73] F. Wang, Y. Zhou, S. Wang, V. Vardhanabhuti, and L. Yu, "Multi-granularity cross-modal alignment for generalized medical visual representation learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 33 536–33 549, 2022.

[74] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, "Vmamba: Visual state space model," *arXiv preprint arXiv:2401.10166*, 2024.

[75] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," *arXiv preprint arXiv:2401.09417*, 2024.

[76] X. Wang, S. Wang, Y. Ding, Y. Li, W. Wu, Y. Rong, W. Kong, J. Huang, S. Li, H. Yang *et al.*, "State space model for new-generation network alternative to transformers: A survey," *arXiv preprint arXiv:2404.09516*, 2024.