Token-Mixer: Bind Image and Text in One Embedding Space for Medical Image Reporting

Yan Yang, Jun Yu, *Member, IEEE*, Zhenqi Fu, Ke Zhang, Ting Yu, Xianyun Wang, Hanliang Jiang, Junhui Lv, Qingming Huang, *Fellow, IEEE*, and Weidong Han

***Abstract*— Medical image reporting focused on automatically generating the diagnostic reports from medical images has garnered growing research attention. In this task, learning cross-modal alignment between images and reports is crucial. However, the exposure bias problem in autoregressive text generation poses a notable challenge, as the model is optimized by a word-level loss function using the teacher-forcing strategy. To this end, we propose a novel Token-Mixer framework that learns to bind image and text in one embedding space for medical image reporting. Concretely, Token-Mixer enhances the cross-modal alignment by matching image-to-text generation with text-to-text generation that suffers less from exposure bias. The framework contains an image encoder, a text encoder and a text decoder. In training, images and paired reports are first encoded into image tokens and text tokens, and these tokens are randomly mixed to form the mixed tokens. Then, the text decoder accepts image tokens, text tokens or mixed tokens as prompt tokens and conducts text generation for network optimization. Furthermore, we introduce a tailored text decoder and an alternative training strategy that well integrate with our Token-Mixer framework. Extensive experiments across three publicly available datasets demonstrate Token-Mixer successfully enhances the image-text alignment and thereby attains a state-of-the-art performance. Related codes are available at https://github.com/yangyan22/Token-Mixer.**

***Index Terms*— Medical image report generation, vision and language, image-text alignment, alternative training, deep learning.**

Yan Yang, Jun Yu, Ke Zhang and Xianyun Wang are with the Key Laboratory of Complex Systems Modeling and Simulation, School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: yangyan@hdu.edu.cn; yujun@hdu.edu.cn; ke.zhang@hdu.edu.cn; wangxianyun@hdu.edu.cn). (Corresponding author: Weidong Han and Jun Yu)

Zhenqi Fu is with the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: fuzhenqi@mail.tsinghua.edu.cn).

Ting Yu is with the School of Information Science and Technology, Hangzhou Normal University, Hangzhou 311121, China (e-mail: yut@hznu.edu.cn).

Hanliang Jiang is with the Regional Medical Center for National Institute of Respiratory Diseases, Sir Run Run Shaw Hospital, College of Medicine, Zhejiang University, Hangzhou 310016, China (e-mail: aock@zju.edu.cn).

Junhui Lv are with the Department of Neurosurgery, Sir Run Run Shaw Hospital, College of Medicine, Zhejiang University, Hangzhou 310016, China (email: 3415030@zju.edu.cn).

Qingming Huang is with the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, 101408, China (email: qmhuang@ucas.ac.cn).

Weidong Han is with the Department of Colorectal Medical Oncology, Zhejiang Cancer Hospital, Hangzhou 310022, China, and also with the College of Mathematical Medicine, Zhejiang Normal University, Jinhua 321017, China (e-mail: hanwd@zju.edu.cn).

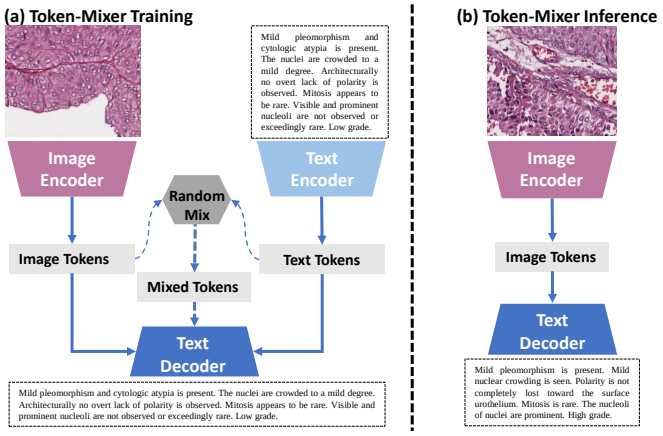


Fig. 1. Overview of our Token-Mixer. (a) In training, the process of image-to-text generation is aligned with text-to-text generation. Image tokens, text tokens or mixed tokens are input to a shared text decoder to conduct report generation for network optimization. (b) During inference, only image tokens are fed to the decoder for report generation.

I. Introduction

INTERPRETING medical images and writing free-text diagnostic reports can be knowledge-demanding and time-consuming in clinical processes. Due to the heavy workload in clinics, even experienced doctors are prone to misdiagnosis and make errors. Toward this end, researchers are motivated to study automated medical image report generation, which has shown great potential to improve the routine diagnosis experience by offering second opinions in computer-aided diagnosis systems [1].

In recent years, various methods have been proposed for medical image report generation, including template-based, retrieval-based and free-text generation-based models. Template-based models [2]–[4] attach either fully-structured templates or semi-structured tags to the given image, and retrieval-based approaches retrieve the target report or sentences for the query image from a predefined corpus [5]. However, both template-based and retrieval-based methods cannot generate diverse and flexible reports with rich contextual information [6].

Currently, free-text generation models under the encoder-decoder framework are prevailing in medical image report generation. In this paradigm, the encoder extracts visual features from the input image and the decoder generates the report word by word conditioned on visual features using the maximum likelihood-based autoregressive model [7]. Concretely,

the model is trained to predict the succeeding word given the previous ground-truth sequence and the image, and the network is optimized by the loss function at the word level. However, at testing, the succeeding word is generated conditioned on the previously generated sequence and the image. Here, we argue such discrepancy between training and testing (i.e., exposure bias) together with the long report length would inevitably result in the poor image-text alignment. Toward this end, prior arts try to build the cross-modal alignment with the assistance of memory networks, knowledge graphs or cross-modality pre-training. Prior work [8] proposed reinforcement learning over a cross-modality memory to align visual and textual features for radiology report generation. KAGE [9] projected image and report to a shared latent space with a knowledge graph and a knowledge-driven encoder. Prior art [10] proposed the vision-language pre-training (i.e., ClinicalBERT) to learn the cross-modality alignment.

In this paper, to alleviate exposure bias and enhance cross-modal image-text alignment, we propose a novel Token-Mixer framework that binds image and text in one latent space for medical image report generation. As depicted in Fig. 1, beyond the target image-to-text generation, we introduce the extra text-to-text generation to assist cross-modal alignment. The image-to-text generation is matched with the text-to-text generation that we found suffers less from the exposure bias in our preparatory experiments (refer to "Preparatory Experiments" in "Exposure Bias Experiments" section where we compare the exposure bias between image-to-text generation and text-to-text generation). Token-Mixer includes three cooperative modules: an image encoder, a text encoder and a text decoder. First, medical images and paired reports are encoded to image tokens and text tokens respectively, which are then randomly mixed to form mixed tokens. Afterwards, the text decoder accepts these token sequences as prompt and conducts report generation to optimize the modules jointly. In this way, the image and text could be well matched, even though the model is optimized by the word-level loss function under the teacher-forcing strategy. Additionally, we introduce the alternative training strategy and a novel tailored text decoder to fit with the Token-Mixer. In the inference stage, only image tokens are fed to the text decoder for report generation. Overall, the main contributions of our paper are summarized as:

- We propose a novel Token-Mixer framework to bind image and text in one embedding space for medical image report generation, where image-to-text generation is aligned with text-to-text generation to alleviate exposure bias and enhance cross-modal alignment.
- We propose a tailored text decoder and an alternative training strategy that seamlessly integrate with our Token-Mixer framework. The text decoder receives different prompt tokens (i.e., image tokens, text tokens, or mixed tokens) and conducts the report generation alternatively for network parameter optimization.
- Extensive experiments on three publicly available datasets demonstrate our Token-Mixer successfully learns the cross-modal image-text alignment and achieves the state-of-the-art performance.

II. Related Works

In this section, we review related works concerning image-text alignment and medical image report generation.

A. Image and Text Alignment

Image and text alignment has attracted remarkable attention along with the rapid evolution of multi-modal learning techniques. Recently, cross-modality pre-training models that align the image and text in a shared latent space have exhibited impressive performance for down-stream tasks. For instance, CLIP [11] and ALIGN [12] collect large collections of image-text pairs and train models using contrastive learning by predicting whether the images and texts are matched. Likewise, IMAGEBIND [13] learns to align six different modalities into a single embedding space by employing contrastive learning across multi-modal image-paired data. BLIP [14] pre-trains a multi-modal mixture of encoder-decoder model using a dataset bootstrapped from large-scale noisy image-text pairs, which aligns image and text modalities via an image-text contrastive loss and a image-text matching loss. BLIP-2 [15] pre-trains a lightweight Querying Transformer in two stages to bridge the vision-language gap. The first stage bootstraps vision-language representation learning from a frozen pre-trained image encoder. The second stage bootstraps vision-to-language generation from a frozen large language model.

In medical vision-language learning domain, MGCA [16] focuses on the global and local alignment between medical images and reports at the pathological region level, instance level and disease level. ConVIRT [17] learns medical visual representations from medical image-report pairs through contrastive learning. REFERS [18] generalizes radiograph representation learning via cross-supervision between images and free-text radiology reports. MedCLIP [19] decouples images and texts for contrastive learning and scales up the usable training data at low cost. PLIP [20] proposes pathology language–image pretraining by contrastive learning. MPMA [21] proposes a multi-task paired masking with alignment modelling for medical vision-language pre-training. It integrates two pre-training tasks, i.e., cross-modal alignment tasks and joint image-text reconstruction tasks to achieve comprehensive cross-modal interaction. PRIOR [22] proposes to learn fine-grained semantic representation by cross-modal alignment and cross-modal conditional reconstruction. The cross-modal alignment aligns both global and local information via contrastive learning. The cross-modal conditional reconstruction reconstructs the masked image based on the report and generates sentence prototypes based on the image. MRM [23] proposes to learn transferable knowledge-enhanced radiograph representations via reconstructing masked records, i.e., masked radiograph patches and masked report tokens.

B. Medical Image Report Generation

As aforementioned, structured template-based models [2]–[4] attach either fully-structured templates or semi-structured tags to given medical images, and retrieval-based approaches retrieve the target report or sentences for the query medical

image from a predefined corpus [5]. However, both template-based and retrieval-based methods cannot generate diverse and flexible reports with rich contextual information [6]. Therefore, free-text generation methods are now dominating medical image report generation. The free-text generation models mostly follow the encoder-decoder pipeline, where the encoder extracts vision embeddings and the decoder accepts vision embeddings to conduct report generation. R2Gen [24] extracts convolutional visual features from the image and feeds these features to a memory-driven Transformer [25] for report generation. RATCHET [26] achieves report generation by a convolutional encoder and a Transformer decoder. METransformer [27] proposes learnable expert tokens for Transformer-based encoder and decoder. KiUT [28] proposes a U-connection between Transformer encoder and decoder, and employs a symptom graph and a knowledge distiller to assist generation. BioViL-T [29] proposes to incorporate a series of images and correlate them to reports, enabling the alignment between text and multiple temporal images. RAMT [30] proposes a relation-aware mean teacher framework for semi-supervised report generation. RGRG [31] first detects anatomical regions and then describes individual salient regions to form the final report. DCL [32] adopts a dynamic graph to enhance contrastive learning for report generation. Prior work [33] predefines a knowledge graph to assist disease classification and report generation. ATAG [34] proposes an attributed abnormality graph for report generation, which consists of interconnected abnormality nodes and attribute nodes for better capturing more fine-grained abnormality details. AlignTransformer [35] proposes to predict the disease tags from the input image and then align these tags with associated visual regions to promote semantic alignment. CMN [36] employs the cross-modal memory to learn alignment for report generation. PPKED [37] explores posterior and prior knowledge for report generation. Wang *et al*. [38] automates radiographic report generation purely using Transformers. They adopt a vision transformer encoder and a memory augmented text transformer decoder as backbones, and they propose a term weighting loss, an image-text matching loss with temporal-weighting and a multi-label classification loss to regularise the network. ITA [39] proposes an inclusive task-aware framework for report generation, where each Transformer head generates descriptions for a specific radiology structure. Besides, an auto-balance mask loss is proposed to relieve the imbalance regarding different abnormalities. Related work [40] pretrains textbook reconstruction first and then transfers the knowledge learned from textbooks to assist report generation. The medical terminologies are used to associate textbook reconstruction with report generation.

Recently, there has been a prevailing trend towards employing large language models [41], [42] for multi-modal tasks. Notably, GPT-4V [43] has shown remarkable capabilities for general visual-language tasks, e.g., image captioning. However, it encounters significant challenges in disease diagnosis and generating reports [44]. In these tasks, GPT-4V can hardly make accurate diagnoses. Consequently, various domain-specific large multimodal models tailored for the medical field are proposed. ChatCAD [45] uses large language models (LLMs) for report generation. Through a series of network processes, input images yield diverse outputs that are then translated into text prompts for LLMs to generate reports. LLaVA-Med [46] proposes a novel curriculum learning method for adapting LLaVA [41] to the biomedical domain. Med-PaLM M [47] proposes a single multitask, multimodal biomedical AI system capable of performing multiple tasks with remarkable efficacy. CheXagent [48] proposes an instruction-tuned vision-language foundation model for automated chest X-ray interpretation.

In this paper, we focus on small-scale free-text generation models for medical image report generation. We propose a Token-Mixer framework that learns to align image and text in one embedding space via the token mixing and alternative training strategies. We introduce the extra text-to-text generation and align it with the target image-to-text generation to alleviate exposure bias and enhance implicit cross-modal alignment. Moreover, our research makes a pioneering endeavour in quantifying exposure bias for medical image report generation. Comprehensive experiments on three public datasets have shown that our method successfully alleviates the exposure bias and promotes the cross-modal alignment.

III. Method

In this section, we present technical details of the proposed Token-Mixer. As illustrated in Fig. 2, Token-Mixer aligns the image-to-text generation with text-to-text generation via a token mixing strategy to promote cross-modal interaction and alignment. It includes three jointly-trained modules: an image encoder that extracts image tokens from input images, a text encoder that encodes text tokens from paired reports, and a tailored text decoder that decodes the prompt tokens to generate reports. In training, the decoder accepts image tokens, text tokens or the image-text mixed tokens as prompt tokens and generates report texts. The alternative token selection (ATS) module can select different prompt tokens alternatively. At inference, image tokens are used for report generation. Besides, an alternative training strategy is introduced for parameter optimization.

A. Image Encoder

The image encoder extracts image tokens from the input medical image. In this paper, we first extract visual features from the input image using a convolutional neural network or a vision Transformer. Then, we flatten and embed the visual features with a multi-layer perceptron network (MLP) to yield image tokens. Subsequently, the image tokens will be fed to the text decoder to conduct report generation. Precisely, image tokens are defined as $V = \{v_1, v_2, \ldots, v_N\} \in \mathbb{R}^{N \times d_{\text{model}}}$, N is the sequential length of image tokens and d_{model} is the model dimension of the proposed Token-Mixer network.

B. Structured Text Encoder

The text encoder yields text tokens from the paired report. We employ a structured text encoder to generate text tokens as shown in Fig. 3. We first embed the report into textual

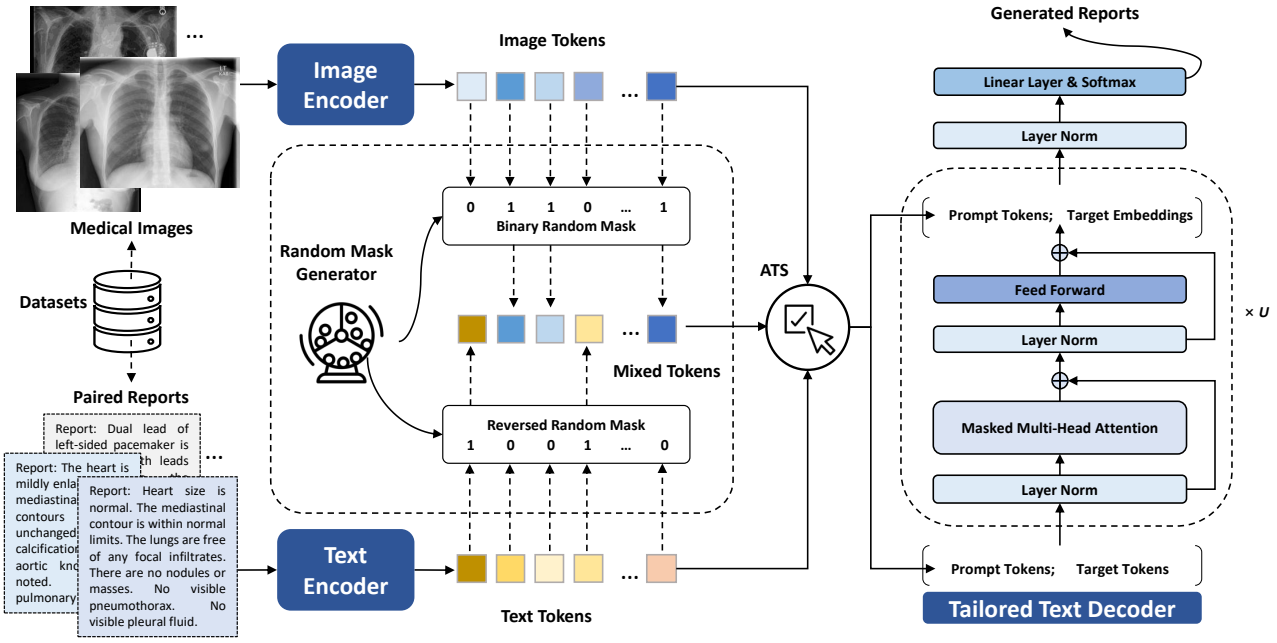


Fig. 2. The workflow of Token-Mixer. It includes three collaborative jointly-trained modules, i.e., an image encoder, a text encoder and a tailored text decoder. In training, the image and paired report are encoded to image tokens and text tokens, which are randomly mixed as mixed tokens. Then, the decoder takes different tokens as prompt tokens to conduct report generation for network optimization. ATS means the alternative token selection. [;] is the concatenation operation. At inference, we use image tokens as the prompt tokens to generate reports.

embeddings with a vanilla Transformer encoder and then feed these embeddings to a structured embedding layer to acquire text tokens. Concretely, text tokens are encoded with the same sequential length as image tokens to facilitate alignment between text and image tokens. We define the report as $R = \{r_1, r_2, ..., r_L\}$, L is the report length, r_i is the word embedding of the i_{th} word. The text encoder first embeds the report with a vanilla Transformer encoder composed of the multi-head attention layers and feed-forward layers, yielding the textual embeddings $S = \{s_1, s_2, \ldots, s_L\} \in \mathbb{R}^{L \times d_{\text{model}}}$.

Subsequently, a structured embedding layer [49] is employed to further encode textual embeddings S into structured text tokens. As shown in Fig. 3 (b), this process is characterized by the following equations (1) and (2):

$$A = \text{softmax}\left(W_2 \tanh\left(W_1 S^{\text{T}}\right)\right) \tag{1}$$

where $A \in \mathbb{R}^{N \times L}$ is an attention matrix. N is the sequential length of text tokens, which shares the same length as image tokens. $W_1 \in \mathbb{R}^{d_{\text{middle}} \times d_{\text{model}}}$ and $W_2 \in \mathbb{R}^{N \times d_{\text{middle}}}$ are linear matrices. d_{model} is the model dimension of our Token-Mixer and d_{middle} is the intermediate transformation dimension. The text tokens can be obtained by:

$$T = \{t_1, t_2, \ldots, t_N\} = (A \times S)\, W \tag{2}$$

where $W \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ is a linear projection matrix. $T = \{t_1, t_2, \ldots, t_N\} \in \mathbb{R}^{N \times d_{\text{model}}}$ denotes the text tokens.

Furthermore, we randomly mix image tokens and text tokens with a random binary mask to form the mixed tokens. Details of the token mixing process will be introduced in the subsection of alternative training.

C. Tailored Text Decoder

A simple tailored text decoder is proposed for report generation conditioned on the prompt tokens, i.e., image tokens $V = \{v_1, v_2, \ldots, v_N\}$, text tokens $T = \{t_1, t_2, \ldots, t_N\}$ or the mixed tokens. Details of the decoder are presented in the right of Fig. 2, where we employ the GPT network structure [50]. The main novelty is that we feed the prompt tokens to all sub-layers and concatenate them with the target embeddings for masked attention, which effectively facilitate the interaction between the input prompt tokens and the target tokens. Concretely, the generation of word r_t at time step t is conditioned on prompt tokens and the previous report sequence $R_{t-1} = \{r_1, r_2, ..., r_{t-1}\}$. There are multiple sub-layers in the decoder. Each sub-layer containing a masked multi-head attention (MHA) layer and a feed forward layer (FFN) can be defined as:

$$X^i = \left[P; E^i\right] \tag{3}$$

$$O^i = \text{FFN}\left(\text{MHA}(X^i, X^i, X^i)\right) \tag{4}$$

$$E^{i+1} = O^i[N+1 : N+t-1] \tag{5}$$

where P denotes the prompt tokens with a length of N, which will be fed to all sub-layers for masked multi-head attention. E^i is the embeddings corresponding to the position of the previous report sequence R_{t-1} in the i_{th} sub-layer with a length of $t-1$. X^i and O^i are the input and output of the i_{th} sub-layer, of which the length is $N+t-1$. $[;]$ denotes the concatenation operation. $[start : end]$ is the sequence slicing operation, where $start$ and end are the starting and ending indices of the slice. Notably, the model size and computational

FLOPs of the tailored decoder remain the same as the decoder where prompt tokens are only fed to the first sub-layer. Given a query matrix Q, a key matrix K, and a value matrix V, the MHA including h scaled dot-product attention heads can be defined as:

$$\mathrm{MHA}(Q,K,V) = [\mathrm{head}_1;\ldots;\mathrm{head}_h]\,W^O \tag{6}$$

$$\mathrm{head}_n = \mathrm{Attention}(QW_n^Q, KW_n^K, VW_n^V) \tag{7}$$

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{8}$$

where $W_n^Q \in \mathbb{R}^{d_{\mathrm{model}}\times d_k}$, $W_n^K \in \mathbb{R}^{d_{\mathrm{model}}\times d_k}$, $W_n^V \in \mathbb{R}^{d_{\mathrm{model}}\times d_v}$ and $W^O \in \mathbb{R}^{nd_v\times d_{\mathrm{model}}}$ are projection matrices. d_{model} is the model dimension of Token-Mixer. d_k is the dimension of keys and queries. d_v is the dimension of values. T is the transpose operation. Both FFN and MHA are followed by the residual addition and layer normalization. FFN composed of two linear transformation functions with a ReLU activation function in between is defined as:

$$\mathrm{FFN}(x) = \max(0, xW_{f1} + b_1)W_{f2} + b_2 \tag{9}$$

where W_{f1} and W_{f2} are the projection matrices for the feed forward network. b_1 and b_2 denote the bias values.

Finally, we extract the target embeddings denoted as $E^U = \{e_1^U, e_2^U, \ldots, e_{t-1}^U\}$ from the output of the last sub-layer for report generation. U is the total count of sub-layers. The word r_t at time step t will be generated conditioned on e_{t-1}^U through a linear transformation and a softmax activation:

$$r_t \sim p_t = \mathrm{softmax}\left(e_{t-1}^U W_p + b_p\right) \tag{10}$$

where $W_p \in \mathbb{R}^{d_{\mathrm{model}}\times d_{\mathrm{vocab}}}$ and b_p are learnable parameters, d_{vocab} is the vocabulary size for report generation. p_t is the probability distribution of the current word over the vocabulary. Note that we only include words occurring before the special token END within the pre-defined maximum length in the generated report.

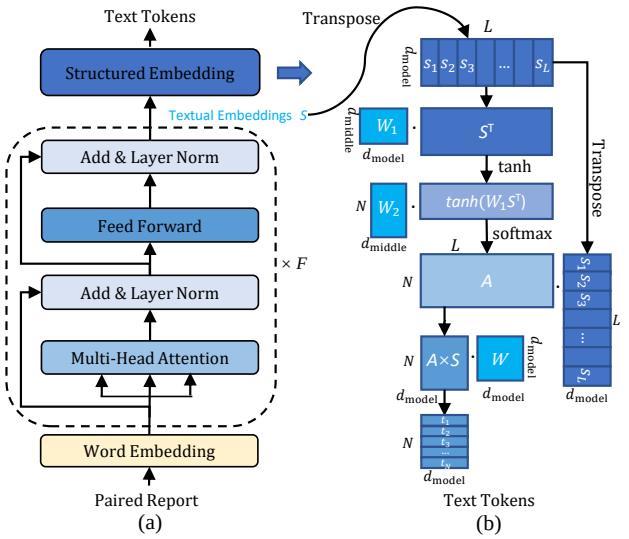


Fig. 3. Illustration of the structured text encoder: (a) The overall workflow of text encoding, (b) Details of the structured embedding layer.

D. Alternative Training Strategy

In this paper, Token-Mixer is integrated with an alternative training strategy, which turns out important in real practice. The image encoder, text encoder and text decoder are jointly trained by this strategy. Specifically, we supply the tailored text decoder with image tokens, text tokens or mixed tokens alternatively using the alternative token selection (ATS) module, thereby enabling an effective joint optimization for all modules in Token-Mixer. The network optimization is carried out by the following cross-entropy loss function:

$$\mathcal{L}(P) = -\sum_{t=1}^{L}\log\left(p\left(r_t \mid (R_{t-1}, P)\right)\right) \tag{11}$$

where $\mathcal{L}(P)$ is the loss function of the proposed Token-Mixer. The prompt tokens P can be image tokens V, text tokens T or the mixed tokens. p is the probability of generating word r_t at time step t conditioned on the previous report sequence R_{t-1} and the prompt tokens P.

For alternative training, we train Token-Mixer in multiple rounds and each round contains multiple epochs as shown in Algorithm 1. Specifically, in the beginning epochs, we optimize the network by $\mathcal{L}(V) + \mathcal{L}(T)$. In the second-to-last epoch, we optimize the network by $\mathcal{L}(Y_t)$, where the mixed tokens Y_t can be obtained by mixing image and text tokens with a random binary mask as shown in Fig. 2:

$$Y_t = T \odot M + V \odot (1 - M)\,. \tag{12}$$

In the last epoch of the round, we optimize the network by $\mathcal{L}(Y_v)$, where the mixed tokens Y_v can be obtained by:

$$Y_v = V \odot M + T \odot (1 - M) \tag{13}$$

where M is the random binary mask with a mixing ratio λ. $\odot$ denotes the element-wise multiplication. $Y_t \in \mathbb{R}^{N\times d_{\mathrm{model}}}$ and $Y_v \in \mathbb{R}^{N\times d_{\mathrm{model}}}$ are two types of mixed tokens in a round. As the round increases, the mixing ratio will increase accordingly.

Algorithm 1 Alternate Training Strategy

Input: Image tokens V, text tokens T, random masks M.
Hyper-Parameter: Max-Epoch, Round-Gap.
Output: Optimized network.

Initialize the Token-Mixer network parameters.
Let epoch = 0.
while epoch $\leq$ Max-Epoch **do**
  epoch = epoch + 1
  if epoch mod Round-Gap = 0 **then**
    Get mixed tokens by $Y_v = V \odot M + T \odot (1 - M)$
    Optimize the network by $\mathcal{L}(Y_v)$
    Increase the mask ratio λ of the random mask M.
  else if epoch mod Round-Gap = Round-Gap - 1 **then**
    Get mixed tokens by $Y_t = T \odot M + V \odot (1 - M)$
    Optimize the network by $\mathcal{L}(Y_t)$
  else
    Optimize the network by $\mathcal{L}(V) + \mathcal{L}(T)$
  end if
end while
return optimized Token-Mixer parameters.

TABLE I
QUANTITATIVE PERFORMANCE COMPARISONS ON MIMIC-CXR DATASET. THE BEST RESULTS ARE MARKED IN BOLD.

Method	Year	Image Encoder	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	Parameters
R2Gen [24]	2020	ResNet-101	0.353	0.218	0.145	0.103	0.142	0.277	-	90.8M
CMN [36]	2021	ResNet-101	0.353	0.218	0.148	0.106	0.142	0.278	-	64.8M
PPKED [37]	2021	ResNet-152	0.360	0.224	0.149	0.106	0.149	0.284	-	-
AlignTransformer [35]	2021	ResNet-50	0.378	0.235	0.156	0.112	**0.158**	0.283	-	-
KGAE [9]	2021	ResNet-50	0.369	0.231	0.156	0.118	0.153	**0.295**	-	-
RATCHET [26]	2021	DenseNet-121	0.326	0.205	0.139	0.099	0.136	0.280	0.140	51M
CMM+RL [8]	2022	ResNet-101	0.381	0.232	0.155	0.109	0.151	0.287	-	-
Clinical-BERT [10]	2022	DenseNet-121	0.383	0.230	0.151	0.106	0.144	0.275	0.167	102M
ITA [39]	2022	ResNet-101	0.395	0.253	0.170	0.121	0.147	0.284	-	-
Wang *et al*. [38]	2022	ViT-3-layers	0.351	0.223	0.157	0.118	-	0.287	0.281	-
DCL [32]	2023	ViT-Base	-	-	-	0.109	0.150	0.284	0.281	-
METransformer [27]	2023	ViT-Base	0.386	0.250	0.169	**0.124**	0.152	0.291	**0.362**	152M
RAMT [30]	2023	DenseNet-121	0.362	0.229	0.157	0.113	0.153	0.284	-	-
Our Token-Mixer	2023	ResNet-50	**0.409**	**0.257**	**0.175**	**0.124**	**0.158**	0.288	0.163	125.18M

TABLE II
QUANTITATIVE PERFORMANCE COMPARISONS ON IU X-RAY DATASET. THE BEST RESULTS ARE MARKED IN BOLD.

Method	Year	Image Encoder	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	Parameters
R2Gen [24]	2020	ResNet-101	0.470	0.304	0.219	0.165	0.187	0.371	-	90.8M
CMN [36]	2021	ResNet-101	0.475	0.309	0.222	0.170	0.191	0.375	-	64.8M
PPKED [37]	2021	ResNet-152	0.483	0.315	0.224	0.168	0.190	0.376	0.351	-
AlignTransformer [35]	2021	ResNet-50	0.484	0.313	0.225	0.173	0.204	0.379	-	-
KGAE [9]	2021	ResNet-50	**0.512**	0.327	0.240	0.179	0.195	0.383	-	-
RATCHET [26]	2021	DenseNet-121	0.452	0.292	0.211	0.163	0.183	0.356	**0.603**	51M
CMM+RL [8]	2022	ResNet-101	0.494	0.321	0.235	0.181	0.201	0.384	-	-
Clinical-BERT [10]	2022	DenseNet-121	0.495	0.330	0.231	0.170	**0.209**	0.376	0.432	102M
Wang *et al*. [38]	2022	ViT-3-layers	0.496	0.319	0.241	0.175	-	0.377	0.449	-
DCL [32]	2023	ViT-Base	-	-	-	0.163	0.193	0.383	0.586	-
METransformer [27]	2023	ViT-Base	0.483	0.332	0.228	0.172	0.192	0.380	0.435	152M
RAMT [30]	2023	DenseNet-121	0.482	0.310	0.221	0.165	0.195	0.377	-	-
Our Token-Mixer	2023	ResNet-50	0.483	**0.338**	**0.250**	**0.190**	0.208	**0.402**	0.482	125.18M

IV. EXPERIMENTS

To test the performance of Token-Mixer, we conduct comprehensive experiments (including performance comparisons, ablation studies, clinical accuracy evaluations, case studies, exposure bias experiments and human evaluations) on public datasets, namely MIMIC-CXR [51], IU X-Ray [52], and Bladder Pathology [1].

A. Datasets and Evaluation Metrics

We evaluate the performance of our Token-Mixer on three publicly available datasets: (1) IU X-Ray [52] contains 7,470 images and 3,955 reports. Each individual report is associated with either a single image or multiple images. Following previous works [8], [24], [36], we exclude samples without "findings" and employ the widely-used data split with 70% for training, 10% for validation and 20% for testing. (2) MIMIC-CXR [51] is the largest public dataset for chest X-ray report generation, which contains 377,110 chest X-ray images and 227,835 reports. We adopt the official data split and collect 368,960, 2,991 and 5,159 samples for training, validation and testing, respectively. (3) Bladder Pathology [1] contains 4,253 bladder pathology images collected from 221 whole slide images (WSIs), which includes non-invasive high grade and low grade papillary urothelial carcinoma. We follow the official data split and exclude samples with the label "insufficient information", collecting 2,076 samples for training and 1,734 samples for testing.

To assess the generation performance, we employ a set of machine translation metrics (i.e., Bilingual evaluation understudy (BLEU) [53], consensus-based image description evaluation (CIDEr) [54], recall-oriented understudy for gisting evaluation (ROUGE) [55] and METEOR [56]) as well as the clinical accuracy metrics (i.e., Accuracy, F1, Recall, Precision and AUC). Furthermore, we conduct human evaluations on generated reports for performance assessment.

B. Experimental Configurations

Token-Mixer is trained in an end-to-end manner with PyTorch on four NVIDIA GeForce RTX 2080Ti GPUs. Images are resized and randomly cropped into 224×224 in training, and images are resized to 224×224 for testing. We replace words occurring less than 3 times in IU X-Ray and words occurring less than 10 times in MIMIC-CXR with UNK. We set letters to lower-case and remove non-alpha tokens from reports. The maximum length of the report is set to 60 in IU X-Ray and Bladder Pathology. The maximum length of the report is set to 100 in MIMIC-CXR. We extract convolutional feature maps with a dimension of $3 \times 3 \times 1024$ from images by the pretrained ResNet-50 [57] in MedCLIP [19] and a convolutional layer with a 5×5 kernel. Then, we flatten feature maps and use a two-layer linear network to yield image tokens with a dimension of 9×512. Accordingly, we encode the paired report into text tokens with a dimension of 9×512. The number of sub-layers in the text encoder and text decoder is set to 3 and 12. The number of heads in MHA is set to 8.

The model dimension d_{model} is set to 512. The batch size is set to 64. AdamW optimizer [58] is used to optimize Token-Mixer with a learning rate of 2e-5 for the convolution neural networks and 1e-4 for the rest parameters. Furthermore, we use drop-out, weight decay and early-stopping strategies to prevent over-fitting. The Max-Epoch and Round-Gap are set to 60 and 4, respectively.

C. Performance Comparisons

In this section, we conduct performance comparisons with state-of-the-art methods in recent years, i.e., R2Gen [24], CMN [36], PPKED [37], AlignTransformer [35], KGAE [9], RATCHET [26], CMM+RL [8], Clinical-BERT [10], ITA [39], Wang *et al*. [38], DCL [32], METransformer [27], and RAMT [30]. All comparison methods take the medical image as input at testing stage (i.e., image-to-text generation) and employ Transformer or its' variants as the backbone decoder. For the image encoder, R2Gen, CMN, CMM+RL and ITA utilize ResNet-101. RATCHET, Clinical-BERT and RAMT use DenseNet-121 [59]. PPKED uses ResNet-152. AlignTransformer and KGAE employ ResNet-50 as the vision backbone. DCL, METransformer and Wang *et al*. take Vision Transformer (ViT) [60] as the vision encoder. More information of the methods can be found in "Related Works". Our Token-Mixer employs the pretrained ResNet-50 [19] as the vision backbone and takes the Transformer encoder combined with a structured embedding layer as the backbone of the text encoder. And we design a tailored Transformer-based decoder as the text decoder. The decoder's network structure is similar to GPT [50]. The text encoder and text decoder are not pretrained. Quantitative performance comparisons on MIMIC-CXR, IU X-Ray and Bladder Pathology are shown in Table I, Table II and Table III respectively, where the best results are highlighted in bold. As can be observed, our Token-Mixer demonstrates remarkable performance across most of the metrics within three datasets. Notably, our method attains the best BLEU-1, BLEU-2, BLEU-3, BLEU-4 and METEOR scores in MIMIC-CXR. Moreover, we achieve a performance enhancement in BLEU-2, BLEU-3, BLEU-4 and ROUGE-L scores in IU X-Ray. Furthermore, our method shows better performance across metrics in Bladder Pathology. It is noted that we reproduce R2Gen, CMN, CMM+RL and RATCHET for performance comparison in the Bladder Pathology dataset. Compared with existing methods, our Token-Mixer introduces the extra text-to-text generation for network training to alleviate exposure bias and effectively learns the cross-modal alignment. To the best of our knowledge, our paper may be the first work trying to quantify the exposure bias and alleviate the exposure bias in report generation.

D. Ablation Studies

We conduct a series of ablation studies on IU X-Ray, MIMIC-CXR and Bladder Pathology to test efficacy of different components in Token-Mixer. Results are shown in Table IV. "Text-to-text generation" is the baseline text-to-text generation network that employs the structured text encoder as the text encoder and a decoder that takes prompt tokens only to the first layer as the text decoder. The rest ablation studies are all image-to-text generation networks. "Baseline" is the image-to-text generation network that employs ResNet-50 as the image encoder and a decoder that takes prompt tokens only to the first layer as the text decoder. The "Baseline" takes only images as input and is optimized without alternative training. The variant "Baseline w tailored decoder" replaces the text decoder in "Baseline" with the proposed tailored text decoder. The variant "Baseline w ATS" trains the image-to-text generation and the text-to-text generation alternatively as the Token-Mixer framework, where the decoder accepts prompt tokens only to the first layer. The results of "Baseline", "Baseline w tailored decoder", "Baseline w ATS" and Token-Mixer in Table IV are the image-to-text generation during testing. As indicated in Table IV, both "Baseline w tailored decoder" and "Baseline w ATS" outperforms "Baseline", highlighting the efficacy of the proposed tailored text decoder and the alternative training strategy. Furthermore, Token-Mixer achieves the best performance across nearly all metrics, indicating the importance of component cooperation in Token-Mixer. Another phenomenon is that "Text-to-text generation" achieves high BLEU4 scores across three datasets, which means that the network can well reconstruct the input text from text tokens under the auto-regressive text generation pipeline. For all the ablation studies, we report the performance as mean and standard deviation across 3 runs.

TABLE III
QUANTITATIVE PERFORMANCE COMPARISONS ON BLADDER PATHOLOGY DATASET. THE BEST RESULTS ARE MARKED IN BOLD.

Method	BLEU-4	METEOR	ROUGE-L	CIDEr
R2Gen [24]	0.276	0.284	**0.538**	1.191
CMN [36]	0.270	0.279	0.535	1.069
CMM+RL [8]	0.251	0.268	0.517	0.992
RATCHET [26]	0.273	0.290	0.530	1.147
Our Token-Mixer	**0.281**	**0.293**	**0.538**	**1.216**

We also conducted additional experiments to test the performance under different experimental settings concerning different backbones, training strategies and mixing ratios. Results are shown in Table V. All experiments employ the Token-Mixer as the backbone network. "Pretrained text encoder" replaces the text encoder with a pretrained text encoder [19]. "Transformer decoder" replaces the tailored decoder with the original Transformer decoder in Token-Mixer. "$\mathcal{L}(V)$ or $\mathcal{L}(T)$" trains the network by alternatively training the image-to-text generation and text-to-text generation without the token-mixing. "$\mathcal{L}(V) + \mathcal{L}(T)$" trains the network by optimizing the image-to-text generation and text-to-text generation simultaneously. "$\mathcal{L}(Y_v) + \mathcal{L}(Y_t)$" trains the network with only mixed tokens. "$\lambda = 0\%$", "$\lambda = 25\%$", "$\lambda = 50\%$", "$\lambda = 75\%$" and "$\lambda = 100\%$" denote different mixing ratios in our Token-Mixer. As can be observed from Table V, replacing the text encoder or text decoder marginally affects the performance. The performance of "$\lambda = 25\%$", "$\lambda = 50\%$" and "$\lambda = 75\%$" drops slightly, while the performance of "$\lambda = 0\%$" and "$\lambda = 100\%$" drops to the performance similar to baseline models. The phenomenon means that token-mixing strategy is essential for the performance. Only image tokens

TABLE IV
ABLATION STUDIES ON IU X-RAY, MIMIC-CXR AND BLADDER PATHOLOGY DATASETS. THE BEST RESULTS CONCERNING IMAGE-TO-REPORT GENERATION ARE MARKED IN BOLD.

Datasets	Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
MIMIC-CXR	Text-to-text generation	0.970±0.002	0.940±0.001	0.917±0.001	0.898±0.001	0.614±0.001	0.952±0.001	8.612±0.014
	Baseline	0.355±0.001	0.221±0.001	0.150±0.001	0.110±0.001	0.141±0.000	0.279±0.001	0.169±0.008
	+proposed decoder	0.363±0.004	0.226±0.002	0.154±0.001	0.112±0.001	0.145±0.001	0.280±0.002	0.171±0.003
	+alternative training	0.394±0.005	0.242±0.003	0.163±0.001	0.117±0.000	0.152±0.002	0.282±0.001	**0.177**±0.007
	Token-Mixer	**0.407**±0.002	**0.255**±0.001	**0.175**±**0.001**	**0.125**±0.001	**0.156**±0.002	**0.289**±0.001	0.162±0.002
IU X-Ray	Text-to-text generation	0.831±0.002	0.770±0.002	0.728±0.002	0.697±0.002	0.463±0.000	0.824±0.001	6.539±0.010
	Baseline	0.459±0.007	0.290±0.006	0.205±0.004	0.153±0.001	0.184±0.004	0.366±0.009	0.475±0.008
	+proposed decoder	0.478±0.004	0.326±0.005	0.239±0.002	0.166±0.002	0.196±0.002	0.386±0.002	0.393±0.007
	+alternative training	0.470±0.002	0.301±0.003	0.219±0.002	0.170±0.001	0.186±0.002	0.368±0.001	**0.678**±0.019
	Token-Mixer	**0.493**±0.008	**0.340**±0.004	**0.248**±0.003	**0.187**±0.003	**0.211**±0.003	**0.385**±0.008	0.451±0.023
Bladder Pathology	Text-to-text generation	0.996±0.001	0.995±0.001	0.995±0.000	0.994±0.000	0.809±0.002	0.997±0.002	9.897±0.011
	Baseline	0.552±0.002	0.404±0.001	0.314±0.002	0.253±0.002	0.272±0.002	0.517±0.002	1.068±0.026
	+proposed decoder	0.557±0.005	0.412±0.004	0.325±0.003	0.265±0.002	0.278±0.003	0.527±0.002	1.133±0.020
	+alternative training	0.564±0.003	0.418±0.003	0.326±0.004	0.264±0.004	0.279±0.004	0.528±0.004	1.162±0.013
	Token-Mixer	**0.572**±0.003	**0.427**±0.004	**0.339**±0.005	**0.277**±0.005	**0.291**±0.004	**0.536**±0.002	**1.205**±0.017

TABLE V
REPORT GENERATION PERFORMANCE WITH OTHER EXPERIMENTAL SETTINGS ON IU X-RAY, MIMIC-CXR AND BLADDER PATHOLOGY DATASETS.

Datasets	Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
MIMIC-CXR	Pretrained text encoder	0.402	0.247	0.166	0.119	0.156	0.284	0.179
	Transformer decoder	0.401	0.244	0.163	0.118	0.151	0.280	0.157
	$\mathcal{L}(V)$ or $\mathcal{L}(T)$	0.377	0.233	0.157	0.113	0.146	0.281	0.155
	$\mathcal{L}(V) + \mathcal{L}(T)$	0.368	0.229	0.155	0.112	0.147	0.283	0.177
	$\mathcal{L}(Y_v) + \mathcal{L}(Y_t)$	0.369	0.208	0.129	0.085	0.137	0.250	0.104
	$\lambda = 0\%$	0.388	0.235	0.157	0.110	0.145	0.276	0.134
	$\lambda = 25\%$	0.390	0.240	0.163	0.117	0.151	0.284	0.191
	$\lambda = 50\%$	0.398	0.243	0.165	0.119	0.153	0.282	0.165
	$\lambda = 75\%$	0.402	0.244	0.162	0.116	0.155	0.283	0.179
	$\lambda = 100\%$	0.369	0.227	0.152	0.110	0.144	0.283	0.154
IU X-Ray	Pretrained text encoder	0.503	0.343	0.245	0.179	0.213	0.365	0.419
	Transformer decoder	0.499	0.336	0.237	0.170	0.222	0.361	0.279
	$\mathcal{L}(V)$ or $\mathcal{L}(T)$	0.474	0.305	0.221	0.168	0.186	0.383	0.413
	$\mathcal{L}(V) + \mathcal{L}(T)$	0.475	0.308	0.217	0.163	0.198	0.372	0.602
	$\mathcal{L}(Y_v) + \mathcal{L}(Y_t)$	0.417	0.262	0.184	0.137	0.169	0.377	0.356
	$\lambda = 0\%$	0.478	0.305	0.219	0.167	0.185	0.379	0.346
	$\lambda = 25\%$	0.484	0.317	0.230	0.173	0.198	0.373	0.372
	$\lambda = 50\%$	0.471	0.310	0.228	0.178	0.194	0.368	0.711
	$\lambda = 75\%$	0.458	0.301	0.219	0.168	0.193	0.369	0.654
	$\lambda = 100\%$	0.461	0.290	0.210	0.162	0.183	0.352	0.551
Bladder Pathology	Pretrained text encoder	0.557	0.411	0.325	0.265	0.276	0.521	1.137
	Transformer decoder	0.558	0.413	0.325	0.263	0.277	0.530	1.124
	$\mathcal{L}(V)$ or $\mathcal{L}(T)$	0.557	0.408	0.319	0.259	0.276	0.525	1.080
	$\mathcal{L}(V) + \mathcal{L}(T)$	0.547	0.406	0.322	0.261	0.272	0.523	1.078
	$\mathcal{L}(Y_v) + \mathcal{L}(Y_t)$	0.472	0.340	0.256	0.196	0.229	0.475	0.502
	$\lambda = 0\%$	0.561	0.411	0.321	0.260	0.280	0.528	1.115
	$\lambda = 25\%$	0.554	0.408	0.323	0.264	0.278	0.525	1.155
	$\lambda = 50\%$	0.556	0.411	0.324	0.265	0.275	0.527	1.122
	$\lambda = 75\%$	0.557	0.412	0.324	0.264	0.275	0.526	1.126
	$\lambda = 100\%$	0.552	0.407	0.319	0.259	0.273	0.523	1.079

and text tokens ("$\lambda = 0\%$" and "$\lambda = 100\%$") for training may not enhance the alignment. Also, the performance of "$\mathcal{L}(V)$ or $\mathcal{L}(T)$" and "$\mathcal{L}(V)$ + $\mathcal{L}(T)$" declines to the level of baseline models, while the performance drop sharply when only mixed tokens are used for training (i.e., "$\mathcal{L}(Y_v) + \mathcal{L}(Y_t)$").

E. Clinical Accuracy Evaluations

In this section, we further assess the performance of Token-Mixer using clinical accuracy metrics in MIMIC-CXR (containing 14 disease classes) and Bladder Pathology (containing 3 classes, i.e., normal, low grade and high grade). For Bladder Pathology, we extract classification labels directly from the reference and generated reports to calculate accuracy scores. Results are shown in Table VI. For MIMIC-CXR, we first annotate disease labels for the reference and generated reports with a CheXbert labeler [61]. There are four tags in each disease: "-1: uncertain", "0: negative", "1: positive", and "None: not mentioned". We set "None" and "-1" to "0", and calculate accuracy scores by comparing labels of the generated and reference reports. Results are shown in Table VII. As can be observed, Token-Mixer yields remarkable performance in Bladder Pathology, and achieves good performance with respect to some diseases such as "cardiomegaly", "support devices" and "pleural effusion" in MIMIC-CXR.

Furthermore, we test the clinical accuracy performance of the baseline model for comparison. The baseline uses ResNet-

TABLE VI
ACCURACY SCORES ON BLADDER PATHOLOGY DATASET. ("AVE." MEANS "AVERAGE".)

Class	Accuracy	Precision	Recall	F1	AUC
Normal	0.985	0.333	1.000	0.5	0.993
Low grade	0.741	0.600	0.717	0.653	0.735
High grade	0.741	0.844	0.739	0.788	0.742
Micro average	0.822	0.733	0.733	0.733	0.800
Macro average	0.822	0.592	0.819	0.647	0.823
Baseline micro ave.	0.802	0.704	0.704	0.704	0.778
Baseline macro ave.	0.802	0.480	0.491	0.470	0.660

TABLE VII
ACCURACY SCORES ON MIMIC-CXR DATASET. ("CARDIO." MEANS "CARDIOMEDIASTINUM". "AVE." MEANS "AVERAGE".)

Class	Accuracy	Precision	Recall	F1	AUC
Atelectasis	0.718	0.446	0.368	0.403	0.604
Cardiomegaly	0.722	0.636	0.527	0.578	0.679
Consolidation	0.940	0.148	0.042	0.065	0.515
Edema	0.835	0.463	0.313	0.374	0.623
Pleural effusion	0.797	0.711	0.603	0.652	0.745
Pleural other	0.957	0.095	0.014	0.024	0.504
Pneumonia	0.941	0.149	0.036	0.058	0.513
Pneumothorax	0.970	0.250	0.231	0.240	0.608
Enlarged cardio.	0.898	0.135	0.041	0.062	0.508
Lung lesion	0.930	0.357	0.038	0.069	0.517
Lung opacity	0.651	0.601	0.198	0.298	0.560
Fracture	0.948	0.333	0.005	0.010	0.502
Support devices	0.819	0.755	0.721	0.737	0.796
No finding	0.759	0.142	0.585	0.229	0.678
Micro average	0.849	0.538	0.403	0.461	0.669
Macro average	0.849	0.373	0.266	0.271	0.597
Baseline micro ave.	0.839	0.494	0.324	0.392	0.631
Baseline macro ave.	0.839	0.314	0.210	0.215	0.572

50 as the image encoder and a decoder that takes prompt tokens only to the first layer as the text decoder. Results are shown at the bottom of Table VI and Table VII. As can be observed, Token-Mixer show better performance in accuracy.

F. Case Studies

Report generation examples on chest X-ray images and bladder pathology images are presented in Fig. 4. We compare the reports generated by our Token-Mixer, CMN, R2Gen and RATCHET. The results of R2Gen and CMN are from the released models. As can be seen, Token-Mixer generates abnormalities accurately for the given chest X-rays covering almost all diseases. In other models, lots of the abnormality descriptions are missed. For pathology images, Token-Mixer generates impressive reports that are almost the same as the reference reports. For instance, in Fig. 4 (a), our method generate "left-sided dual-chamber pacemaker device", "moderate enlargement of the cardiac silhouette", "the aorta remains tortuous and diffusely calcified" and "lungs are hyperinflated" accurately.

Additionally, as presented in Fig. 5, we leverage the Gradient-weighted Class Activation Mapping (Grad-CAM) [62] to create visualized activation maps that shed light on the regions that the model attends to when generating the complete report. As can be observed, Token-Mixer attends to pertinent abnormality regions accurately. For instance, Token-Mixer precisely attends to low lung volumes and bronchovascular crowding in Fig. 5 (b), and attends to the tubes and pleural effusions as shown in Fig. 5 (d).

TABLE VIII
PREPARATORY EXPERIMENTS OF EXPOSURE BIAS CONCERNING IMAGE-TO-TEXT (I2T) AND TEXT-TO-TEXT (T2T) GENERATION.

Datasets	Inference Strategies	BLEU-4	CIDEr
MIMIC-CXR	I2T teacher-forcing	0.312	1.45
	I2T normal sampling	0.112	0.186
	I2T relative change	0.641	0.872
	T2T teacher-forcing	0.927	9.06
	T2T normal sampling	0.883	8.63
	T2T relative change	0.047	0.047
IU X-Ray	I2T teacher-forcing	0.506	4.058
	I2T normal sampling	0.175	0.393
	I2T relative change	0.654	0.903
	T2T teacher-forcing	0.716	6.72
	T2T normal sampling	0.683	6.44
	T2T relative change	0.046	0.042

G. Exposure Bias Experiments

In this section, we conduct experiments concerning the exposure bias, where we also try to quantify the exposure bias. All the experiments are tested in the testing data split. We first conduct preparatory experiments. We compare image-to-text generation with text-to-text generation. Concretely, we train the backbone image-to-text generation and text-to-text generation models utilizing the training data splits. In the backbone image-to-text generation, the image encoder is a ResNet-50 and the text decoder is a vanilla Transformer decoder. In the backbone text-to-text generation, the text encoder is a structured text encoder and the text decoder is a vanilla Transformer decoder. Subsequently, we evaluate the performance of these models using testing splits. Results of image-to-text generation (noted as "I2T") and text-to-text generation (noted as "T2T") are presented in Table VIII. In the table, we present the relative changes between different inference strategies including the teacher-forcing (i.e., predicting the succeeding word given the previous ground-truth sequence and the input) and normal sampling (generating the word conditioned on the previously generated sequence and the input) in terms of BLEU4 and CIDEr. Here, such relative change is quantified to measure exposure bias. As can be observed, in both datasets, image-to-text generation ("I2T") suffers from severe exposure bias, whereas text-to-text generation ("T2T") exhibits a moderate exposure bias. The phenomenon is attributed to the model's training strategy using the teacher-forcing strategy and a word-level loss function.

Moreover, we also conduct experiments to test whether our Token-Mixer can overcome such exposure bias. We compare the exposure bias of our Token-Mixer with the ablation model "Baseline w tailored decoder". The BLEU-4 and CIDEr of image-to-text generation on testing splits are presented in Table IX. As can be observed, our Token-Mixer shows less exposure bias on both datasets concerning both BLEU-4 and CIDEr. The phenomena demonstrate that our Token-Mixer suffers less from exposure bias compared with the "Baseline w tailored decoder".

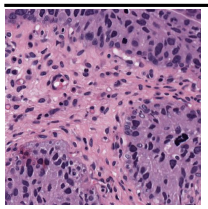
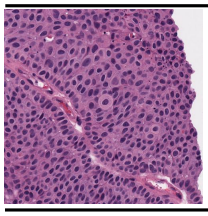
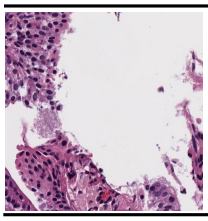
	Input Images	Reference Reports	RATCHET [26]	R2Gen [24]	CMN [36]	Our Token-Mixer
(a)		Left-sided dual-chamber pacemaker leads terminating in the right atrium and right ventricle are noted. There is mild enlargement of the cardiac silhouette which is stable. The aorta remains tortuous and diffusely calcified. Pulmonary vascularity is normal. The lungs remain hyperinflated. No pleural effusion or pneumothorax is seen. There is minimal atelectasis at the lung bases but no areas of focal consolidation. No acute osseous abnormality is present.	Left chest wall pacer device is again seen with leads extending to the region the right atrium and right ventricle. The heart remains mildly enlarged. The lungs are clear without focal consolidation large effusion or pneumothorax. No signs of congestion or edema. Bony structures are intact. No free air below the right hemidiaphragm.	Dual-lead left-sided pacemaker is again seen with leads extending to the expected positions of the right atrium and right ventricle. No focal consolidation, pleural effusion or evidence of pneumothorax is seen. The cardiac and mediastinal silhouettes are stable..	*Port-A-Cath resides over the left chest wall with catheter tip extending into the region of the low svc.* The lungs are clear without focal consolidation large effusion or pneumothorax. The cardiomediastinal silhouette appears normal. The imaged bony structures are intact. No free air below the right hemidiaphragm.	Left-sided dual-chamber pacemaker device is noted with leads terminating in the right atrium and right ventricle. Moderate enlargement of the cardiac silhouette is unchanged. The aorta remains tortuous and diffusely calcified. The mediastinal and hilar contours are similar. Pulmonary vasculature is not engorged. Lungs are hyperinflated. No focal consolidation pleural effusion or pneumothorax is present. No acute osseous abnormality is visualized.
(b)		Opacity over the right mid-to-lower lateral lung appears similar likely corresponding to known loculated pleural effusion catheter within the effusion appears similarly positioned. Right Port-A-Cath terminates in the low svc similar to prior. No new consolidation left effusion pneumothorax or pulmonary edema is detected. Heart size is persistently enlarged likely exaggerated by low lung volumes.	Right chest wall Port-A-Cath is again seen with catheter tip extending to the region of the low svc. There is a right pleural effusion *with compressive atelectasis.* There is no pneumothorax. The heart is mildly enlarged. mediastinal contour is normal. bony structures are intact.	Right-sided Port-A-Cath tip terminates in the low svc. Heart size is mildly enlarged. Mediastinal contour is unchanged. There is mild pulmonary edema with perihilar haziness and vascular indistinctness. More focal consolidative opacity is seen *within the right upper lobe concerning for pneumonia.* No pleural effusion or pneumothorax is present. There are no acute osseous abnormalities.	Right-sided Port-A-Cath tip terminates in the low svc. Heart size remains moderately enlarged. Mediastinal contour is unchanged. *There is mild pulmonary edema* new in the interval with a small right pleural effusion. Patchy opacities in the lung bases may reflect areas of atelectasis. No pneumothorax is identified. No acute osseous abnormalities are detected.	Right-sided Port-A-Cath tip terminates in the low svc unchanged. Heart size remains mildly enlarged. Mediastinal and hilar contours are similar. Pulmonary vasculature is not engorged. Small to moderate size right pleural effusion is re-demonstrated. Patchy opacities in the lung bases likely reflect areas of atelectasis. No pneumothorax is present. there are No acute osseous abnormalities.
(c)		The patient is status post sternotomy. The heart is moderately enlarged. Layering pleural effusions are present. These are difficult to directly compare to the prior study because of suspected differences in positioning but the appearance is probably fairly similar. Coinciding compressive atelectasis is likely. mild interstitial opacification suggests mild vascular congestion new since the prior study. Prior vertebroplasties have been performed.	The patient is status post median sternotomy and CABG. Heart size is mildly enlarged. *The aorta is tortuous.* The mediastinal and hilar contours are similar. Pulmonary vasculature is not engorged. Patchy opacities in the lung bases likely reflect areas of atelectasis. No pleural effusion or pneumothorax is present. There are no acute osseous abnormalities.	The patient is rotated to the left. The patients chin overlies the medial lung apices partially obscuring the view. Given this no focal consolidation is seen. there is no large pleural effusion or pneumothorax. The cardiac and mediastinal silhouettes are grossly stable. There is severe compression of a lower thoracic vertebral body grossly stable.	The patient is status post median sternotomy and CABG. The heart size is mildly enlarged. *The aorta is tortuous and diffusely calcified. There is mild pulmonary edema with* small bilateral pleural effusions. patchy opacities in the lung bases may reflect atelectasis. No pneumothorax is identified. There are no acute osseous abnormalities.	The patient is status post median sternotomy and CABG. The heart size remains moderately enlarged. *The aorta is tortuous and diffusely calcified.* The mediastinal and hilar contours are unchanged. There is mild pulmonary vascular congestion. Small bilateral pleural effusions are demonstrated with bibasilar airspace opacities likely reflective of atelectasis. No pneumothorax is present. There are moderate multilevel degenerative changes in the thoracic spine.
(d)		Severe pleomorphism and cytologic atypia is present. Moderate crowding of the nuclei can be seen. There is marked disorganization and lack of cellular polarity toward the surface urothelium. Mitosis is frequent throughout the tissue. The nucleoli are mostly inconspicuous. High grade.	Slight variability in nuclear size shape and outline consistent with mild pleomorphism. Nuclei are mildly crowded together. Polarity is not completely lost toward the surface urothelium. Mitosis is rare throughout the tissue. The nucleoli are mostly inconspicuous. *Low grade.*	Moderate pleomorphism and cytologic atypia is present. Mild crowding of the nuclei can be seen. Polarity is completely lost. Mitosis is rare throughout the tissue. The nucleoli are mostly inconspicuous. High grade.	Moderate pleomorphism and cytologic atypia is present. Nuclei are mildly crowded together. Polarity is not completely lost toward the surface urothelium. Mitosis is infrequent throughout the tissue. The nucleoli are mostly inconspicuous. High grade.	Mild pleomorphism and cytologic atypia is present. The nuclei are crowded to a mild degree. Architecturally the cells show complete lack of polarity toward the surface urothelium. Mitosis is rare throughout the tissue. The nucleoli are mostly inconspicuous. High grade.
(e)		Moderate pleomorphism and cytologic atypia is present. Nuclei are moderately crowded together. There is no loss of polarity. Mitosis are exceedingly rare and limited only to the basal layer of urothelium. The nuclei have inconspicuous nucleoli. Low grade.	Mild pleomorphism and cytologic atypia is present. Mild nuclear crowding is seen. Polarity is not completely lost toward the surface urothelium. Mitosis is rare. The nucleoli are mostly inconspicuous. Low grade.	Moderate pleomorphism and cytologic atypia is present. Mild crowding of the nuclei can be seen. Polarity is not completely lost toward the surface urothelium. *Mitosis is frequent throughout the tissue. Nucleoli is prominent. High grade.*	Moderate pleomorphism and cytologic atypia is present. Nuclei are mildly crowded together. *Polarity is completely lost.* There are infrequent mitotic figures throughout the tissue. The nucleoli are mostly inconspicuous. *High grade.*	Mild pleomorphism and cytologic atypia is present. The nuclei are crowded to a moderate degree. Polarity is not completely lost toward the surface urothelium. Mitosis is rare throughout the tissue. The nucleoli are mostly inconspicuous. Low grade.
(f)		Mild pleomorphism and cytologic atypia is present. No nuclear crowding is seen. insufficient information. Mitosis are exceedingly rare and limited only to the basal layer of urothelium. Nucleoli is absent to inconspicuous. Normal.	Mild pleomorphism and cytologic atypia is present. *Moderate crowding of the nuclei can be seen.* Polarity is not completely lost toward the surface urothelium. Mitosis are exceedingly rare and limited only to the basal layer of urothelium. The nucleoli are mostly inconspicuous. *Low grade.*	Moderate pleomorphism and cytologic atypia is present. *Nuclei are moderately crowded together.* Insufficient information. Mitosis is rare throughout the tissue. The nucleoli are mostly inconspicuous. *Low grade.*	*Moderate pleomorphism and cytologic atypia is present. Moderate nuclear crowding is seen.* Polarity is not lost. Mitosis is rare throughout the tissue. Visible and prominent nucleoli are not observed or exceedingly rare. *Low grade.*	Mild pleomorphism and cytologic atypia is present. The nuclei are normally crowded. Polarity is not lost. Mitosis is rare throughout the tissue. The nucleoli are mostly inconspicuous. Normal.

Fig. 4. Examples of reports generated by different methods (i.e., our Token-Mixer, RATCHET, R2Gen and CMN) on chest X-ray images and bladder pathology images. Correct disease captions are highlighted in blue, while incorrect disease descriptions are shown in red italics. Zoom in for details.

H. Human Evaluations

In this section, we conduct human evaluations to further assess the efficacy of our Token-Mixer. Specifically, we randomly select 100 samples from the testing set of MIMIC-CXR for evaluation. Three doctors are invited to compare reports generated by our Token-Mixer against reports generated by R2Gen, CMN or RATCHET. Each comparison is presented in a structured manner, encompassing the report generated by our Token-Mixer, the report generated by the comparison method, the reference report and the corresponding image. During the evaluation, doctors are instructed to rate a "win", "loss" or "tie" for each comparison in terms of correctness and coverage. "Win" means that the report matches better with the reference report and the image in correctness or coverage. Correctness refers to the accuracy of generated normalities and abnormalities. The coverage means that the coverage rate of generated abnormalities compared with the reference abnormalities. In cases that the reports exhibit similar correctness or coverage, a "tie" will be rated. Finally, we collect results

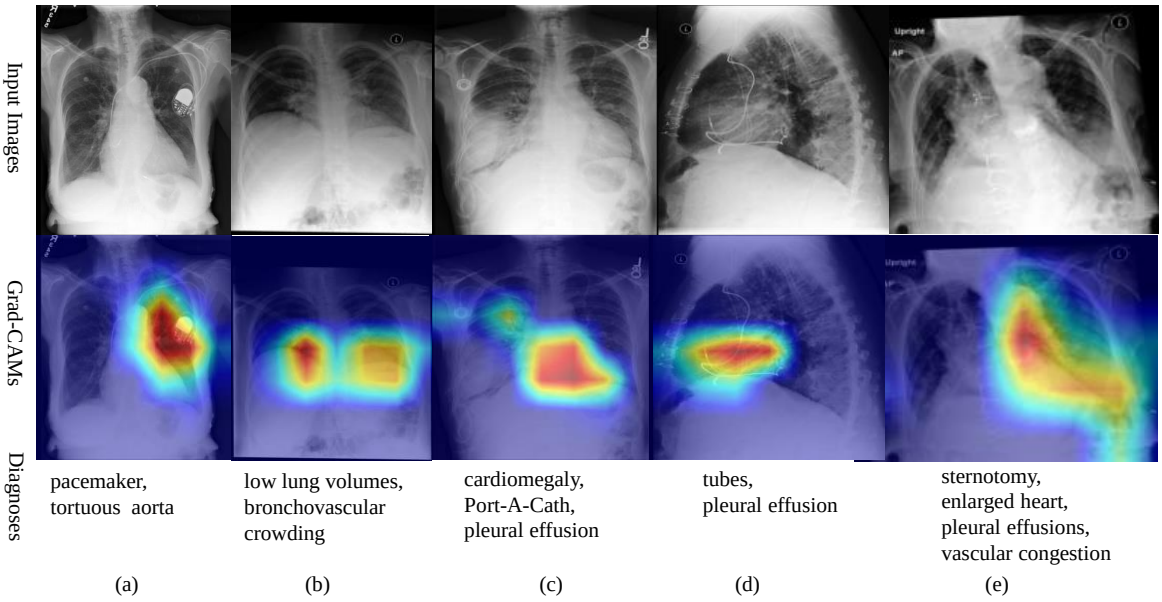


Fig. 5. Examples of Grad-CAMs when generating the complete reports on chest X-ray images. Zoom in for details.

TABLE IX
EXPOSURE BIAS RESULTS OF OUR TOKEN-MIXER AND THE BASELINE MODEL IN TERMS OF IMAGE-TO-TEXT GENERATION.

Datasets	Models	Inference Strategies	BLEU-4	CIDEr
MIMIC-CXR	Baseline w tailored decoder	teacher-forcing	0.318	1.472
		normal sampling	0.113	0.169
		relative change	0.645	0.885
	Token-Mixer	teacher-forcing	0.290	1.266
		normal sampling	0.124	0.163
		relative change	0.572	0.871
IU X-Ray	Baseline w tailored decoder	teacher-forcing	0.512	4.138
		normal sampling	0.166	0.393
		relative change	0.676	0.905
	Token-Mixer	teacher-forcing	0.431	2.510
		normal sampling	0.190	0.482
		relative change	0.559	0.808

TABLE X
HUMAN EVALUATIONS ON GENERATED REPORTS ABOUT THE CORRECTNESS AND COVERAGE. VALUES ARE IN PERCENTAGE %. ("TM" DENOTES THE "TOKEN-MIXER".)

Metrics	TM vs. RATCHET			TM vs. R2Gen			TM vs. CMN		
	Loss	Tie	Win	Loss	Tie	Win	Loss	Tie	Win
Correctness	39.0	14.3	46.7	39.7	15.0	45.3	36.0	14.0	50.0
Coverage	37.7	18.7	43.7	37.0	19.7	43.3	38.7	16.0	45.3

from doctors and compute average percentages for "win", "tie" and "loss" outcomes. Results are shown in Table X. As can be observed, our Token-Mixer achieves better performance in both correctness and coverage.

V. CONCLUSION

We propose a novel, straightforward and effective Token-Mixer framework to enhance the cross-modality alignment for medical image report generation. It aligns the process of image-to-text generation with text-to-text generation, in which the image and text tokens are aligned in a shared embedding space via the token mixing strategy. Meanwhile, we introduce the alternative training strategy and propose a novel tailored text decoder that seamlessly integrates with the Token-Mixer framework. Extensive experiments on three publicly available datasets have verified the impressive performance. Looking ahead, we believe automatic generation of medical image reports holds significant potential for practical implementation in clinical diagnosis. It will improve the efficiency and accuracy of medical diagnosis by providing a second opinion report for reference. Furthermore, we envision the versatility of our Token-Mixer framework extending beyond medical image report generation. We believe it is promising for other image-to-text generation tasks.

REFERENCES

[1] Z. Zhang, P. Chen, M. McGough, F. Xing, C. Wang, M. Bui, Y. Xie, M. Sapkota, L. Cui, J. S. Dhillon, N. Ahmad, F. K. Khalil, S. I. Dickinson, X. Shi, F. Liu, H. Su, J. Cai, and L. Yang, "Pathologist-level interpretable whole-slide cancer diagnosis with deep learning," *Nature Machine Intelligence.*, vol. 1, no. 5, pp. 236–245, 2019.

[2] P. Kisilev, E. Walach, E. Barkan, B. Ophir, S. Alpert, and S. Y. Hashoul, "From medical image to automatic medical report generation," *IBM Journal of Research and Development*, vol. 59, no. 2/3, 2015.

[3] Z. Han, B. Wei, S. Leung, J. Chung, and S. Li, "Towards automatic report generation in spine radiology using weakly supervised framework," in *Medical Image Computing and Computer Assisted Intervention*, vol. 11073, 2018, pp. 185–193.

[4] P. Pino, D. Parra, C. Besa, and C. Lagos, "Clinically correct report generation from chest x-rays using templates," in *Machine Learning in Medical Imaging - 12th International Workshop*, vol. 12966, 2021, pp. 654–663.

[5] M. Kong, Z. Huang, K. Kuang, Q. Zhu, and F. Wu, "TranSQ: transformer-based semantic query for medical report generation," in *Medical Image Computing and Computer Assisted Intervention*, 2022, pp. 610–620.

[6] Y. Yang, J. Yu, J. Zhang, W. Han, H. Jiang, and Q. Huang, "Joint embedding of deep visual and semantic features for medical image report generation," *IEEE Transactions on Multimedia*, vol. 25, pp. 167–178, 2023.

[7] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," in *4th International Conference on Learning Representations*, 2016, pp. 1–16.

[8] H. Qin and Y. Song, "Reinforced cross-modal alignment for radiology report generation," in *Findings of the Association for Computational Linguistics*, 2022, pp. 448–458.

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

[10] B. Yan and M. Pei, "Clinical-BERT: vision-language pre-training for radiograph diagnosis and reports generation," in *Thirty-Sixth AAAI Conference on Artificial Intelligence*, 2022, pp. 2982–2990.

[11] 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*, vol. 139, 2021, pp. 8748–8763.

[12] C. Jia, Y. Yang, Y. Xia, Y. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, 2021, pp. 4904–4916.

[13] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "ImageBind: One embedding space to bind them all," *CoRR*, vol. abs/2305.05665, 2023.

[14] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*, 2022.

[15] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models," *CoRR*, vol. abs/2301.12597, 2023.

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

[17] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in *Proceedings of the Machine Learning for Healthcare Conference*, vol. 182, 2022, pp. 2–25.

[18] H. 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.

[19] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: contrastive learning from unpaired medical images and text," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 3876–3887.

[20] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, and J. Zou, "A visual–language foundation model for pathology image analysis using medical twitter," *Nature Medicine*, vol. 29, pp. 2307–2316, 2023.

[21] 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*, vol. 26, pp. 4706–4721, 2024.

[22] P. Cheng, L. Lin, J. Lyu, Y. Huang, W. Luo, and X. Tang, "PRIOR: prototype representation joint learning from medical images and reports," *CoRR*, vol. abs/2307.12577, 2023.

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

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

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[26] B. Hou, G. Kaissis, R. M. Summers, and B. Kainz, "RATCHET: medical transformer for chest x-ray diagnosis and reporting," in *Medical Image Computing and Computer Assisted Intervention*, 2021, pp. 293–303.

[27] Z. Wang, L. Liu, L. Wang, and L. Zhou, "METransformer: radiology report generation by transformer with multiple learnable expert tokens," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2023, pp. 11 558–11 567.

[28] Z. Huang, X. Zhang, and S. Zhang, "KiUT: knowledge-injected u-transformer for radiology report generation," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2023, pp. 19 809–19 818.

[29] S. Bannur, S. L. Hyland, Q. Liu, F. Pérez-García, M. Ilse, D. C. Castro, B. Boecking, H. Sharma, K. Bouzid, A. Thieme, A. Schwaighofer, M. Wetscherek, M. P. Lungren, A. V. Nori, J. Alvarez-Valle, and O. Oktay, "Learning to exploit temporal structure for biomedical vision-language processing," *CoRR*, vol. abs/2301.04558, 2023.

[30] 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*, vol. 26, pp. 904–915, 2024.

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

[32] 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 Conference on Computer Vision and Pattern Recognition*, June 2023, pp. 3334–3343.

[33] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. L. Yuille, and D. Xu, "When radiology report generation meets knowledge graph," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020, pp. 12 910–12 917.

[34] S. Yan, W. K. Cheung, K. Chiu, T. M. Tong, K. C. Cheung, and S. See, "Attributed abnormality graph embedding for clinically accurate x-ray report generation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 8, pp. 2211–2222, 2023.

[35] D. You, F. Liu, S. Ge, X. Xie, J. Zhang, and X. Wu, "AlignTransformer: hierarchical alignment of visual regions and disease tags for medical report generation," in *Medical Image Computing and Computer Assisted Intervention*, vol. 12903, 2021, pp. 72–82.

[36] Z. Chen, Y. Shen, Y. Song, and X. Wan, "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*, 2021, pp. 5904–5914.

[37] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, "Exploring and distilling posterior and prior knowledge for radiology report generation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 753–13 762.

[38] Z. Wang, H. Han, L. Wang, X. Li, and L. Zhou, "Automated radiographic report generation purely on transformer: A multicriteria supervised approach," *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2803–2813, 2022.

[39] L. Wang, M. Ning, D. Lu, D. Wei, Y. Zheng, and J. Chen, "An inclusive task-aware framework for radiology report generation," in *Medical Image Computing and Computer Assisted Intervention*, vol. 13438. Springer, 2022, pp. 568–577.

[40] G. Liu, Y. Liao, F. Wang, B. Zhang, L. Zhang, X. Liang, X. Wan, S. Li, Z. Li, S. Zhang, and S. Cui, "Medical-VLBERT: Medical visual language BERT for COVID-19 CT report generation with alternate learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 9, pp. 3786–3797, 2021.

[41] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *CoRR*, vol. abs/2304.08485, 2023.

[42] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. K. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, N. Schärli, A. Chowdhery, P. A. Mansfield, B. A. y Arcas, D. R. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. K. Barral, C. Semturs, A. Karthikesalingam, and V. Natarajan, "Large language models encode clinical knowledge," *Nature*, vol. 620, pp. 172–180, 2023.

[43] OpenAI, "GPT-4 technical report," *CoRR*, vol. abs/2303.08774, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.08774

[44] C. Wu, J. Lei, Q. Zheng, W. Zhao, W. Lin, X. Zhang, X. Zhou, Z. Zhao, Y. Zhang, Y. Wang, and W. Xie, "Can GPT-4V(ision) Serve Medical Applications? Case Studies on GPT-4V for Multimodal Medical Diagnosis," *CoRR*, vol. abs/2310.09909, 2023.

[45] S. Wang, Z. Zhao, X. Ouyang, Q. Wang, and D. Shen, "ChatCAD: interactive computer-aided diagnosis on medical image using large language models," *CoRR*, vol. abs/2302.07257, 2023.

[46] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day," *CoRR*, vol. abs/2306.00890, 2023.

[47] T. Tu, S. Azizi, D. Driess, M. Schaekermann, M. Amin, P. Chang, A. Carroll, C. Lau, R. Tanno, I. Ktena, B. Mustafa, A. Chowdhery, Y. Liu, S. Kornblith, D. J. Fleet, P. A. Mansfield, S. Prakash, R. Wong, S. Virmani, C. Semturs, S. S. Mahdavi, B. Green, E. Dominowska, B. A. y Arcas, J. K. Barral, D. R. Webster, G. S. Corrado, Y. Ma-

tias, K. Singhal, P. Florence, A. Karthikesalingam, and V. Natarajan, "Towards generalist biomedical AI," *CoRR*, vol. abs/2307.14334, 2023.
[48] Z. Chen, M. Varma, J.-B. Delbrouck, M. Paschali, L. Blankemeier, D. V. Veen, J. M. J. Valanarasu, A. Youssef, J. P. Cohen, E. P. Reis, E. B. Tsai, A. Johnston, C. Olsen, T. M. Abraham, S. Gatidis, A. S. Chaudhari, and C. Langlotz, "CheXagent: Towards a foundation model for chest x-ray interpretation," *CoRR*, vol. abs/2401.12208, 2024.
[49] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," in *5th International Conference on Learning Representations*, 2017, pp. 1–15.
[50] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, 2020, pp. 1877–1901.
[51] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR: A large publicly available database of labeled chest radiographs," *Scientific Data*, vol. 6, p. 317, 2019.
[52] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. K. Antani, G. R. Thoma, and C. J. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association*, vol. 23, no. 2, pp. 304–310, 2016.
[53] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the Annual Meeting on Association for Computational Linguistics*, 2002, pp. 311–318.
[54] R. Vedantam, C. L. 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.
[55] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL*, July 2004, pp. 74–81.
[56] A. Lavie and A. Agarwal, "METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proceedings of the Second Workshop on Statistical Machine Translation*, 2007, pp. 228–231.
[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[58] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019, pp. 1–19.
[59] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2261–2269.
[60] 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," in *9th International Conference on Learning Representations*, 2021.
[61] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. P. Lungren, "CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT," 2021.
[62] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-CAM: why did you say that? visual explanations from deep networks via gradient-based localization," *CoRR*, vol. abs/1610.02391, 2016.