

## MedCutMix: A Data-Centric Approach to Improve Radiology Vision-Language Pre-Training with Disease Awareness

Authors	Wang, Sinuo;Xie, Yutong;Liu, Yuyuan;Wu, Qi
Citation	S. Wang, Y. Xie, Y. Liu, Q. Wu, "MedCutMix: A Data-Centric Approach to Improve Radiology Vision-Language Pre-Training with Disease Awareness," 2026, pp. 6291-6295.
DOI	<a href="https://doi.org/10.1109/icassp55912.2026.11462043">10.1109/icassp55912.2026.11462043</a>
Publisher	IEEE
Rights	Licence for published version: Creative Commons Attribution 4.0 International
Download date	2026-06-15 04:20:04
Item License	<a href="http://creativecommons.org/licenses/by/4.0/">http://creativecommons.org/licenses/by/4.0/</a>
Link to Item	<a href="https://hdl.handle.net/20.500.14634/2422">https://hdl.handle.net/20.500.14634/2422</a>

# MEDCUTMIX: A DATA-CENTRIC APPROACH TO IMPROVE RADIOLOGY VISION-LANGUAGE PRE-TRAINING WITH DISEASE AWARENESS

Sinuo Wang<sup>1</sup> Yutong Xie<sup>2</sup> Yuyuan Liu<sup>3</sup> Qi Wu<sup>1</sup>

<sup>1</sup>The University of Adelaide

<sup>2</sup>Mohamed bin Zayed University of Artificial Intelligence

<sup>3</sup>University of Oxford

## ABSTRACT

Vision-Language Pre-training (VLP) is drawing increasing interest for its ability to minimize manual annotation requirements while enhancing semantic understanding in downstream tasks. However, its reliance on image-text datasets poses challenges due to privacy concerns and the high cost of obtaining paired annotations. Data augmentation emerges as a viable strategy to address this issue, yet existing methods often fall short of capturing the subtle and complex variations in medical data due to limited diversity. To this end, we propose MedCutMix, a novel multi-modal disease-centric data augmentation method. MedCutMix performs diagnostic sentence CutMix within medical reports and establishes the cross-attention between the diagnostic sentence and medical image to guide attentive manifold mix within the imaging modality. Our approach surpasses previous methods across four downstream radiology diagnosis datasets, highlighting its effectiveness in enhancing performance and generalizability in radiology VLP.

**Index Terms**— Medical VLP, Disease-aware Augmentation

## 1. INTRODUCTION

The rise of self-supervised deep learning methods [1–3] has attracted significant attention due to their capacity to reduce the need for extensive manual annotations. This advantage is especially pronounced in the medical field, where the annotation process is costly but the multimodal nature of medical data often allows weak supervision through accompanying medical reports [4–6]. Leveraging the paired image-text data, the Vision-Language Pre-training (VLP) paradigm enables the model to learn general visual and textual representations, offering data-efficient solutions for downstream target tasks. In contrast to the general domain, medical VLP models necessitate exceeding global alignment and acquiring the ability to capture fine-grained representations. In radiology, clinicians often focus on the subtle yet crucial visual cues to perform the diagnosis, which cannot be adequately captured solely through the global alignment objective [7–11]. To address this, recent works propose attention-based local alignment [8] and disease-level alignment strategies [9]. In the masked image modeling pretraining, MedIM [12] proposes masking and reconstructing the image regions guided by reports to enhance the fine-grained semantics capturing.

Despite these advancements, these approaches persist in their inherent data-hunger nature and face the challenge of medical data scarcity. While widely adopted VLP models in general domains,

such as CLIP [1], are trained on vast datasets of 400M image-text pairs from the internet, replicating such efforts for medical data is currently hindered by privacy and legal concerns. We advocate a data-centric approach to enhance medical VLP by focusing on data augmentation to expand medical image-report pairs. It offers a promising solution to mitigate data scarcity constraints while simultaneously enriching data diversity without requiring real-world data acquisition, thus maintaining patient privacy.

Recognizing this gap, we propose MedCutMix, a novel multi-modal disease-centric data augmentation technique explicitly designed for medical vision-language pre-training. MedCutMix not only increases the volume of the training data for medical VLP but also expands the data diversity, specifically focusing on subtle yet crucial disease-related semantics multi-modally. To achieve this, we extract disease labels from radiology reports using the rule-based labeler. These labels help identify diagnostic sentences, which contain key medical findings relevant to the disease. To effectively integrate this textual information with images, we introduce a pairwise CutMix strategy: (1) *Text-level CutMix*: diagnostic sentences from source and target reports are mixed at the input level to augment textual diversity; and (2) *Feature-level Image Mixing*: we leverage cross-attention between diagnostic sentences and image features to identify disease-relevant image regions, which are then selectively mixed at the feature level, ensuring semantic consistency between the mixed image and mixed report. Our contributions include:

1. We propose MedCutMix, a novel disease-centric data augmentation framework designed for medical VLP, enhancing training data diversity while preserving critical disease-related semantics.
2. To ensure cross-modal consistency, we introduce a pairwise CutMix strategy that integrates text-level CutMix for augmenting diagnostic sentences and feature-level image mixing, where cross-attention aligns disease-relevant visual regions with textual information.
3. Our results show that MedCutMix successfully improves performance over previous methods across four downstream radiology diagnosis datasets, demonstrating its effectiveness in enhancing generalisation in radiology VLP.

## 2. METHOD

In this work, we propose MedCutMix, a multi-modal disease-centric data augmentation method tailored for medical VLP. As shown in Fig. 1, MedCutMix introduces a novel pairwise CutMix strategy,

transferring disease-related regions across image-report pairs to enhance the model's ability to recognize and understand diseases in diverse visual contexts. The method comprises four key steps: (1) Disease-centric label extraction and balanced sampling, (2) Disease-related sentence identification from medical reports, (3) Disease-relevant regions localization from medical images, and (4) Mixing disease-related regions between image-text pairs to maintain semantic consistency.

**Disease-centric label extraction and balanced sampling.** In this work, we use the MIMIC-CXR dataset [13] and focus on  $C$  common chest X-ray disease findings for pairwise CutMix. Let the pre-training dataset be denoted by  $\mathcal{D} = \{(x_i, y_i, l_{(i,c)}) \mid c = 1, 2, \dots, C\}$  where  $x_i \in \mathcal{X} \in \mathbb{R}^{H \times W}$  represents the input image resolution with 3 RGB channels, and  $y_i \in \mathcal{Y} \in \mathbb{R}^{1 \times N_{si}}$  denotes the  $N_{si}$  sentences in each report that describe the corresponding image  $x_i$ . The disease label  $l_{(i,c)}$  is extracted from MIMIC-CXR reports via the open-source rule-based tool-CheXpert labeler [14], where  $C$  denotes the total categories of the dataset. To address the class imbalance, we perform balanced sampling by selecting  $n = \frac{N_{max}}{c}$  instances for each disease, where  $N_{max}$  is a hyperparameter specifying the maximum number of newly created pairs. For each pair of randomly sampled image-report pairs, e.g.,  $(x_i, y_i)$  and  $(x_j, y_j)$ , both of which contain disease  $l_c$ , we identify diagnostic sentences and visual regions pertinent to disease  $l_c$ . We then perform input-level CutMix for the medical report and execute attentive feature-level mixup for the corresponding image modality.

**Disease-related sentence identification from medical reports.** Recognizing that radiological reports often comprise several sentences, each potentially providing independent information about different findings of the image [15]. Diagnostic sentences not only identify the disease but also convey additional details such as severity and other relevant clinical aspects, thus offering a more comprehensive understanding than a mere mention of disease name. Consequently, we extract disease-related content from medical reports at the sentence level. Specifically, we perform the semantic match for the disease  $l_c$  within different sentences in  $\mathbf{y}_i = \{\mathbf{y}_{(i,s)}\}_{s=1}^{N_{si}}$  and  $\mathbf{y}_j = \{\mathbf{y}_{(j,s)}\}_{s=1}^{N_{sj}}$ , where  $N_{si}$  and  $N_{sj}$  are the number of sentences. The sentence containing an exact match of the disease term  $l_c$  is cut out as the diagnostic sentence, denoted as  $\mathbf{y}_{(i,s)}^c = \text{Cutout}(\mathbf{y}_i, M_{ti})$  and  $\mathbf{y}_{(j,s)}^c = \text{Cutout}(\mathbf{y}_j, M_{tj})$ . Here,  $M_{ti}$  and  $M_{tj}$  are the binary masks that point out where the disease-related sentence in the reports, as follows:

$$\begin{aligned} M_{ti} &= \{M_{(ti,s)} = \mathbb{I}(l_c \in \mathbf{y}_{(i,s)})\} \in \{0, 1\}^{N_{si}}, \\ M_{tj} &= \{M_{(tj,s)} = \mathbb{I}(l_c \in \mathbf{y}_{(j,s)})\} \in \{0, 1\}^{N_{sj}}, \end{aligned} \quad (1)$$

where  $\mathbb{I}$  is an indicator function that assigns 1 to sentences containing  $l_c$  and 0 otherwise. The operation  $\text{Cutout}(\cdot, \cdot)$  selects a sentence from  $\mathbf{y}_i$  or  $\mathbf{y}_j$  where the corresponding binary mask  $M_{(ti,s)} = 1$  or  $M_{(tj,s)} = 1$ .

**Disease-relevant regions localization from medical images.** To achieve the disease semantic consistency across the augmented image and corresponding report, we guide the extraction of image regions using the diagnostic sentences. This process utilizes a warmed-up medical VLP model [9] with the Vision Transformer [16], ViT-B/16, as image encoder, which aligns the fine-grained correspondences between medical images' visual and semantic aspects with their corresponding radiological reports, ben-

efiting in better identifying the text-guided discriminative image regions. The base model accepts images  $x_i$  and  $x_j$ , along with their paired reports  $y_i$  and  $y_j$  as inputs. The model generates the image global embedding  $\mathbf{v}_i$  and  $\mathbf{v}_j$ , local patch embeddings  $\mathbf{E}_{(img,i)}$  and  $\mathbf{E}_{(img,j)}$ , as well as the global text representations  $\mathbf{t}_i$  and  $\mathbf{t}_j$ , local embeddings  $\mathbf{E}_{(text,i)}$  and  $\mathbf{E}_{(text,j)}$ . We also collect the intermediate image local features  $\mathbf{E}_{(img,i)}^k$  and  $\mathbf{E}_{(img,j)}^k$  from the  $k$ -th layer from the image encoder for future image feature-level mixup. To target on the  $l_c$  disease-related information, the diagnostic sentence embedding is derived by applying a masking operation over the report, represented as  $\mathbf{E}_{(sent,i)} = \text{Cutout}(\mathbf{E}_{(text,i)}, M_{ti})$ . Leveraging the intrinsic vision-language alignment capability of the warmed-up medical VLP model, the disease-related image regions are identified using the diagnostic sentence embeddings to attend to the image local embeddings. Specifically, we first compute the attention map  $\mathbb{C}$  for the  $i$ -th pair as follows:

$$\mathbb{C}_i = \sum \text{softmax}\left(\frac{\mathbf{E}_{(img,i)} \cdot (\mathbf{E}_{(sent,i)})^\top}{\tau_1}\right) \in \mathbb{R}^{N_{patch}}. \quad (2)$$

Here,  $N_{patch}$  denotes the number of image tokens obtained after flattening the 2D patchified image. The softmax function normalizes the elements along the image dimension to find the focused region matched to each word in the diagnostic sentence. The summation operation  $\sum$  performs on the text dimension to aggregate the attention related to the whole diagnostic sentence.  $\tau$  is the temperature to control the size of the attention area.

**Mixing disease-related regions between image-text pairs.** After identifying the disease-related sentences and their related image regions, we implement the mix augmentation. This text augmentation process operates as an input-level CutMix, involving cutting out the diagnostic sentence from the source report  $\mathbf{y}_i$  and pasting the diagnostic sentence to the target report  $\mathbf{y}_j$ , represented as:

$$\mathbf{y}_{aug} = \text{Paste}\left(\text{Cutout}(\mathbf{y}_i, M_{ti}), \mathbf{y}_j, M_{tj}\right), \quad (3)$$

where  $\text{Paste}(\cdot, \cdot, \cdot)$  denotes the operation of placing the cutout from  $\mathbf{y}_i$  into  $\mathbf{y}_j$  at the location specified by the mask  $M_{tj}$ . For the visual modality, we utilize the source attention  $\mathbb{C}_i$  to perform an attentive soft mix on the intermediate features obtained from the  $k$ -th layer of the image encoder, expressed as:

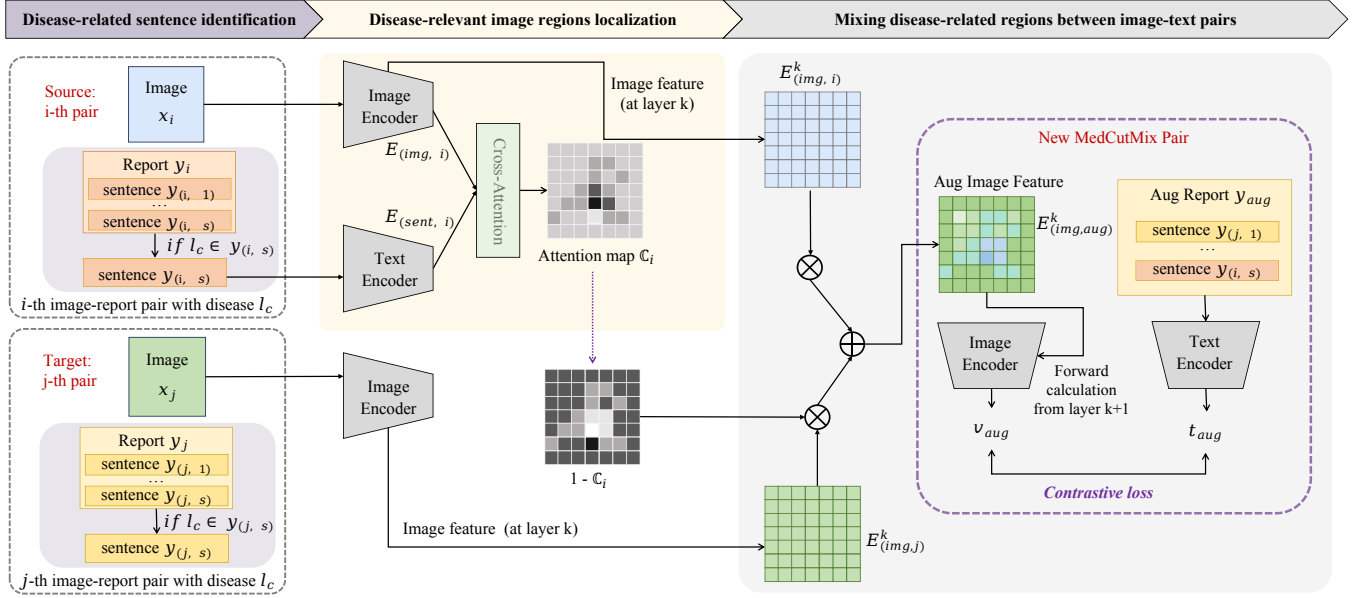
$$\mathbf{E}_{(img,aug)}^k = \mathbb{C}_i \odot \mathbf{E}_{(img,i)}^k + (1 - \mathbb{C}_i) \odot \mathbf{E}_{(img,j)}^k. \quad (4)$$

Following this, the augmented report  $\mathbf{y}_{aug}$  is encoded via the text encoder to yield the global text representation, denoted as  $\mathbf{t}_{aug}$ . whereas, the augmented image feature  $\mathbf{E}_{(img,aug)}^k$  is directly fed back into the image encoder to resume encoding from layer  $k + 1$  onward, and derive the global visual representation  $\mathbf{v}_{aug}$ . The newly augmented global representations are subsequently used to compute another image-text contrastive loss, as follows:

$$\begin{aligned} \ell_{(aug,i)}^{v2t} &= -\log \frac{\exp(\text{sim}(\mathbf{v}_{(aug,i)}, \mathbf{t}_{(aug,i)}) / \tau_2)}{\sum_{j=1}^N \exp(\text{sim}(\mathbf{v}_{(aug,i)}, \mathbf{t}_{(aug,j)}) / \tau_2)}, \\ \ell_{(aug,i)}^{t2v} &= -\log \frac{\exp(\text{sim}(\mathbf{t}_{(aug,i)}, \mathbf{v}_{(aug,i)}) / \tau_2)}{\sum_{j=1}^N \exp(\text{sim}(\mathbf{t}_{(aug,i)}, \mathbf{v}_{(aug,j)}) / \tau_2)}. \end{aligned} \quad (5)$$

The final contrastive loss is the average of the two, which is computed as:

$$\mathcal{L}_{ITC_{aug}} = \frac{1}{2N} \sum_{i=1}^N (\ell_{(aug,i)}^{v2t} + \ell_{(aug,i)}^{t2v}). \quad (6)$$



**Fig. 1.** Overview of the MedCutMix pipeline. After disease-centric label extraction and balanced sampling, the pipeline proceeds with the following steps: (1) identification of disease-related sentences from medical reports, (2) localization of disease-relevant regions in medical images, and (3) mixing disease-related regions between image-text pairs to maintain semantic consistency.

Since MedCutMix operates as a plug-in and runs simultaneously with the base model’s training, we reuse the original objectives of the base model [9], which include a token-wise alignment loss ( $\mathcal{L}_{CTA}$ ), instance-wise alignment loss ( $\mathcal{L}_{ITA}$ ), and a prototype-level alignment loss ( $\mathcal{L}_{CPA}$ ). Therefore, the final loss is expressed as follows:

$$\mathcal{L} = \mathcal{L}_{CTA} + \mathcal{L}_{ITA} + \mathcal{L}_{CPA} + \mathcal{L}_{ITC_{aug}}. \quad (7)$$

MedCutMix strengthens contrastive pre-training by emphasizing disease-specific regions, enhancing the semantic diversity of the original dataset. Additionally, as disease-related image regions are guided by corresponding diagnostic reports, the CutMix areas between images and reports remain semantically aligned, ensuring both the coherence and quality of the augmented data.

### 3. EXPERIMENTS

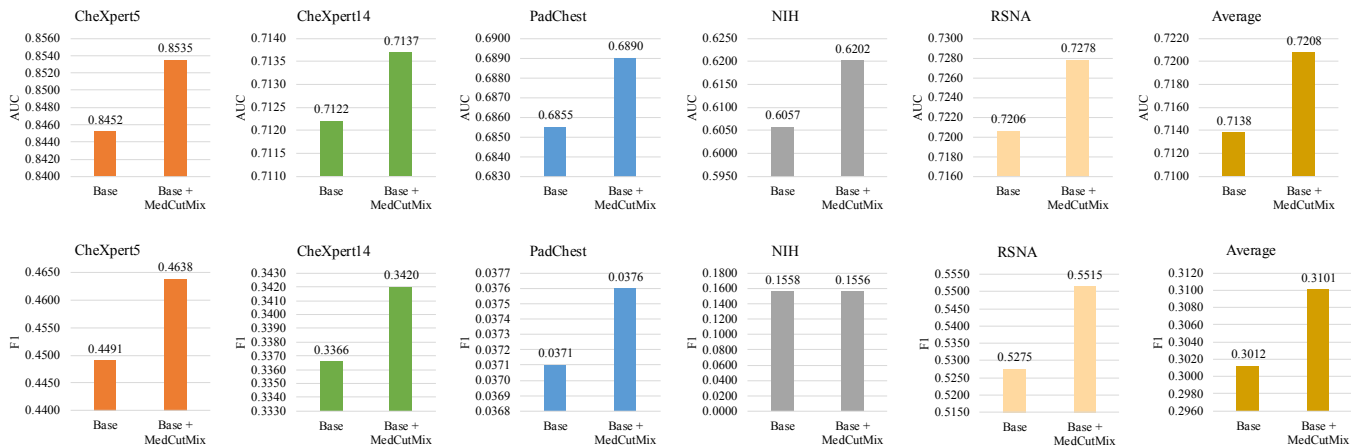
#### 3.1. Experimental Details

**Pre-training Setup.** For our pre-training phase, we utilize the MIMIC-CXR-JPG dataset [13]. Adhering to protocols established by previous research [9], we select only frontal-view chest images and extract the impression and findings from the accompanying radiological reports, yielding over 210,000 radiograph-report pairs. We follow the base framework [9] and use the ViT-B/16 [17] as the image encoder and employ BioClinicalBERT [18] as the text encoder. The batch size is set to 72 on a single GPU, while 2 GPUs are used for pre-training. MedCutMix is applied after 5 epochs of warm-up. In MedCutMix augmentation, we focus on common chest X-ray disease findings suggested by [14] for pairwise augmentation. We empirically set the sentence temperature  $\tau_1$  to 0.005, while the intermediate image features are extracted from the 11th layer of the ViT.

**Downstream Setup.** The efficacy of MedCutMix is validated through a comparative assessment of the transferability of baseline models with and without MedCutMix. Our evaluation focus on zero-shot image classification across various datasets, with results presented as the macro average of AUROC and F1 scores across all categories. **CheXpert** [14] features 224,316 chest X-ray images from 65,240 patients, collected at Stanford Hospital. The official validation set consists of 200 chest radiographic studies annotated by three board-certified radiologists, while the official test set includes 500 studies annotated by five board-certified radiologists. Model selection is based on zero-shot AUROC performance on the validation set. During testing, results are reported both for the 5 observations in the official test set for competition tasks and for the full set of 14 disease labels. **NIH** [19] was utilized in a multi-label classification setup, comprising a total of 100k frontal-view X-ray images of about 32,000 patients annotated with 14 Chest X-Ray CXR diseases. We used official NIH test set for evaluation. **PadChest** [20] comprises over 160,000 chest X-ray images from approximately 67,000 patients at Hospital San Juan (Spain). It has 193 disease image labels, including 174 radiographic 254 findings and 19 differential diagnoses. We adopt 39,053 chest X-rays annotated by board-certified 255 radiologists for zero-shot evaluation. **RSNA Pneumonia** [21] was used in its stage 2 version, followed [9]. It comprises approximately 29,700 frontal-view chest radiographs. The task involves binary classification, distinguishing each chest image as either normal or pneumothorax positive. We employed the data split provided by [9].

#### 3.2. Experimental Results

The results Fig. 2 show the results of the base model without and with MedCutMix, which demonstrate the effectiveness of MedCutMix, a disease-centric data augmentation technique, in improv-



**Fig. 2.** Performance comparison of the base model without and with MedCutMix across multiple radiology datasets. The top row shows the improvement in AUC scores, while the bottom row presents F1-score gains.

ing model performance across multiple radiology diagnosis tasks. Notably, MedCutMix enhances AUC scores, particularly in the CheXpert 5-Class task (from 0.8452 to 0.8535), while also yielding gains in CheXpert 14-Class (0.7122 to 0.7137), NIH (0.6057 to 0.6202), and RSNA (0.7206 to 0.7278), leading to an overall average AUC improvement from 0.7138 to 0.7208. Furthermore, F1 scores also see a significant boost, with CheXpert 5-Class increasing from 0.4491 to 0.4638 and RSNA improving from 0.5275 to 0.5515. These consistent gains highlight MedCutMix’s ability to generate more informative and diverse training samples, enhancing both model generalization and robustness.

### 3.3. Discussions

To evaluate the impact of hyper-parameter settings on MedCutMix, we explored the effects of varying the maximum number of mixed samples ( $N_{max}$ ) and the mixed layer ( $k$ ).

**Analysis of the number of mixed samples.** The results in Table 1 highlight the impact of varying  $N_{max}$  on model performance, revealing a balance between augmentation benefits and potential drawbacks. The highest average AUC (0.7208) is achieved at  $N_{max} = 40$ , demonstrating that moderate augmentation enhances generalization by introducing meaningful diversity while preserving essential clinical features. However, as  $N_{max}$  increases to 300, performance declines significantly (AUC drops to 0.7054), likely due to an overabundance of synthetic samples disrupting the real data distribution and reducing feature fidelity. Dataset-specific effects are also observed, where different datasets exhibit varying sensitivity to synthetic augmentation. These findings underscore that while moderate augmentation strengthens model robustness, excessive artificial mixing may degrade clinical relevance and hinder learning.

**Analysis of the number of mixed layers.** Table 2 presents the impact of varying the intermediate layers ( $k$ ) at which visual features are mixed, with the number of mixed samples held constant. Specifically, we observed highest performance is achieved when mixing occurs at the final layer and diminishes when applied to earlier layers. Nonetheless, all variations still outperform the baseline, which does not employ MedCutMix.

**Table 1.** AUC scores under different maximum MedCutMix samples ( $N_{max}$ ).

$N_{max}$	Datasets					Avg. AUC
	CheXpert5	CheXpert14	PadChest	NIH	RSNA	
0	0.8452	0.7122	0.6855	0.6057	0.7206	0.7138
30	0.8492	0.7158	<b>0.7011</b>	0.6196	0.6412	0.7054
40	0.8535	0.7137	0.6890	<b>0.6202</b>	0.7278	<b>0.7208</b>
50	0.8534	0.7137	0.6890	<b>0.6202</b>	0.7276	0.7208
100	<b>0.8574</b>	<b>0.7415</b>	0.6960	0.6183	0.6097	0.7046
300	0.8360	0.6768	0.6781	0.6068	<b>0.7294</b>	0.7054

**Table 2.** AUC scores under different mixing layers ( $k$ ).

Layer $k$	Datasets					Avg. AUC
	CheXpert5	CheXpert14	PadChest	NIH	RSNA	
5	0.8528	0.7100	0.6912	0.6206	0.7230	0.7195
8	0.8532	0.7130	0.6891	0.6201	0.7272	0.7205
11	0.8534	0.7137	0.6890	0.6202	0.7276	<b>0.7208</b>

## 4. CONCLUSION

This paper presents MedCutMix, a novel multi-modal disease-centric data augmentation technique specifically tailored to address data scarcity in VLP within the radiology domain. Our approach performs CutMix of diagnostic sentences within reports and mix feature-level representations of disease-attended images. The method effectively enhances the semantic diversity of augmented medical data while preserving cross-modal semantic coherence, thus addressing the challenges posed by privacy concerns and labeling costs. Our comprehensive evaluation demonstrates that MedCutMix outperforms the advanced medical VLP baseline in four datasets. These results affirm the efficacy of MedCutMix in improving the performance and generalizability of models within the medical VLP.

## 5. REFERENCES

- [1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [2] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi, “Align before fuse: Vision and language representation learning with momentum distillation,” *Advances in neural information processing systems*, vol. 34, pp. 9694–9705, 2021.
- [3] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *International conference on machine learning*. PMLR, 2022, pp. 12888–12900.
- [4] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz, “Contrastive learning of medical visual representations from paired images and text,” in *Machine Learning for Healthcare Conference*. PMLR, 2022, pp. 2–25.
- [5] Hong-Yu Zhou, Chenyu Lian, Liansheng Wang, and Yizhou Yu, “Advancing radiograph representation learning with masked record modeling,” *arXiv preprint arXiv:2301.13155*, 2023.
- [6] Zhihong Chen, Shizhe Diao, Benyou Wang, Guanbin Li, and Xiang 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. 23403–23413.
- [7] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun, “Medclip: Contrastive learning from unpaired medical images and text,” *arXiv preprint arXiv:2210.10163*, 2022.
- [8] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena 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.
- [9] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu, “Multi-granularity cross-modal alignment for generalized medical visual representation learning,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 33536–33549, 2022.
- [10] Zhongyi Shui, Jianpeng Zhang, Weiwei Cao, Sinuo Wang, Ruizhe Guo, Le Lu, Lin Yang, Xianghua Ye, Tingbo Liang, Qi Zhang, and Ling Zhang, “Large-scale and fine-grained vision-language pre-training for enhanced ct image understanding,” 2025.
- [11] Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rueckert, “Joint learning of localized representations from medical images and reports,” in *European conference on computer vision*. Springer, 2022, pp. 685–701.
- [12] Yutong Xie, Lin Gu, Tatsuya Harada, Jianpeng Zhang, Yong Xia, and Qi Wu, “Medim: Boost medical image representation via radiology report-guided masking,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 13–23.
- [13] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng, “Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs,” *arXiv preprint arXiv:1901.07042*, 2019.
- [14] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al., “Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison,” in *Proceedings of the AAAI conference on artificial intelligence*, 2019, vol. 33, pp. 590–597.
- [15] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al., “Making the most of text semantics to improve biomedical vision–language processing,” in *European conference on computer vision*. Springer, 2022, pp. 1–21.
- [16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [18] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott, “Publicly available clinical bert embeddings,” *arXiv preprint arXiv:1904.03323*, 2019.
- [19] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers, “Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.
- [20] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya, “Padchest: A large chest x-ray image dataset with multi-label annotated reports,” *Medical image analysis*, vol. 66, pp. 101797, 2020.
- [21] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya 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, pp. e180041, 2019.