
Anatomical Masking as Lightweight Visual Priors for Chest X-ray Classification

Jiwoo Park *

Ohio State University
Columbus, OH 43210
park.3620@osu.edu

Doeun Lee *

Ohio State University
Columbus, OH 43210
lee.11501@osu.edu

Wonyoung Kim *

Ohio State University
Columbus, OH 43210
kim.9519@osu.edu

Abstract

Pretrained medical vision-language models have shown strong performance across different imaging tasks, but they often miss the anatomical knowledge that radiologists use during clinical work. We test whether overlaying anatomical segmentation masks on chest X-ray images can serve as simple visual priors to improve classification without changing model architectures. Through fine-tuning BiomedCLIP on MIMIC-CXR across multiple training setups, we find an interesting pattern. Training on masked images does not always help, but evaluating on masked inputs at test time does improve accuracy. Our results suggest that anatomical masks can guide pretrained vision encoders toward relevant diagnostic regions, helping both accuracy and interpretation in automated chest X-ray analysis.

1 Introduction

Automated chest X-ray interpretation has become an important application of deep learning in medical imaging. Recent vision-language models (VLMs) pretrained on large biomedical datasets, such as BiomedCLIP [1], have shown good transferability across different medical imaging tasks. These models learn general-purpose representations from paired image-text data, which can then be adapted to specific classification tasks through fine-tuning.

Despite strong performance, current approaches for adapting pretrained VLMs to chest X-ray classification usually treat images as complete entities. They do not explicitly use the anatomical knowledge that radiologists routinely apply during interpretation. In clinical practice, radiologists systematically examine specific anatomical regions (lung fields, cardiac silhouette, pleural spaces, mediastinum) to identify pathological findings. The spatial location of abnormalities within these anatomical structures matters for differential diagnosis and radiological reporting.

This observation raises a question. Can we explicitly encode anatomical structure into the input space to improve both classification performance and interpretability of pretrained medical VLMs, without needing to change model architectures or add new training objectives? To answer this, we propose a simple approach that overlays anatomical segmentation masks onto chest X-ray images. This highlights clinically relevant regions during both model training and inference.

Our approach uses CheXmask [2], a public dataset of anatomical segmentations aligned with MIMIC-CXR [3]. We create masked variants of chest X-ray images that emphasize boundaries of major thoracic structures (lungs and heart). We test the impact of anatomical masking under different fine-tuning setups (linear probing with frozen encoders vs. joint fine-tuning of encoders and classifiers) and check performance on both original and masked test images. This controlled experimental design lets us separate the effects of anatomical masking during training vs. inference.

*Equal contribution

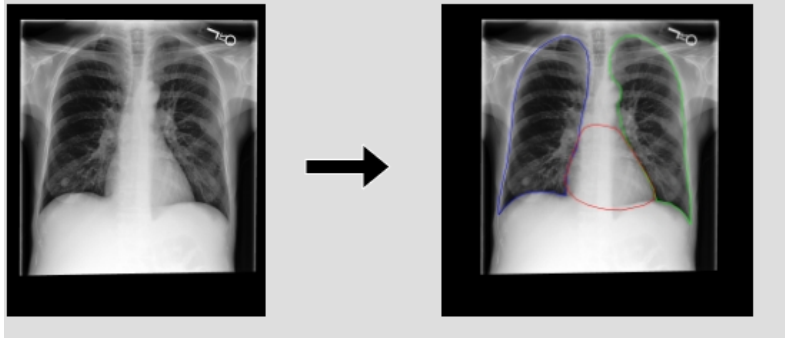


Figure 1: Illustration of anatomical masking. **Left** shows an original chest X-ray image from MIMIC-CXR. **Right** shows the masked variant with overlaid anatomical boundaries from CheXmask, highlighting lung fields and cardiac silhouette. The colored contours guide attention toward clinically relevant regions while keeping the underlying radiographic features.

Our findings show an interesting asymmetry. While training on masked images does not consistently improve performance and may even hurt it by removing contextual information, evaluating on masked images at test time does improve classification accuracy across multiple fine-tuning setups. This suggests that anatomical masks work as lightweight visual priors that can guide the model’s attention toward relevant diagnostic regions without needing explicit attention mechanisms or additional supervision.

2 Related Work

Medical Vision-Language Models. The success of CLIP [4] in natural images has inspired adaptations for medical imaging. BiomedCLIP [1] extends CLIP to biomedical domains by pretraining on PMC-15M, a large dataset of image-caption pairs from PubMed Central articles. Other models like MedCLIP [5] and BioViL [6] have also shown the value of contrastive learning for medical image-text representation. While these models work well for zero-shot and few-shot tasks, their adaptation to fine-grained diagnostic classification often needs task-specific fine-tuning.

Anatomical Priors in Medical Imaging. Adding anatomical knowledge to deep learning models for medical imaging has been explored through different methods. Spatial attention mechanisms [7] let models focus on discriminative regions without explicit anatomical guidance. More recent work has used anatomical segmentations as auxiliary supervision or multi-task learning objectives. But these approaches usually need architectural changes or additional training objectives. Our work tests whether simply overlaying anatomical masks on input images can provide similar benefits without changing model architectures.

Interpretability in Chest X-ray Classification. Interpretability matters for clinical adoption of AI systems. Gradient-based visualization methods [8] and attention mechanisms have been widely used to identify regions that contribute to model predictions. CheXplain [9] and other work have shown that models often attend to anatomically relevant regions. Our approach complements these post-hoc interpretability methods by explicitly encoding anatomical structure as input priors, which may improve both performance and alignment with clinical reasoning.

3 Method

3.1 Problem Formulation

We formulate chest X-ray classification as a multi-label binary classification problem. Given a chest X-ray image $\mathbf{x} \in \mathbb{R}^{H \times W}$ and a set of disease labels $\mathbf{y} \in \{0, 1\}^C$ where C is the number of disease categories, our goal is to learn a classifier $f : \mathbb{R}^{H \times W} \rightarrow [0, 1]^C$ that predicts the probability of each disease being present. We focus on five major pulmonary pathologies (Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion).

3.2 Anatomical Masking Strategy

For each chest X-ray image \mathbf{x} , we get a corresponding anatomical segmentation mask $\mathbf{m} \in \{0, 1\}^{H \times W \times K}$, where K denotes the number of anatomical regions. In our case, we use lungs and heart. The masks come from CheXmask [2], which provides pixel-level segmentations aligned with MIMIC-CXR images.

To create a masked image variant $\tilde{\mathbf{x}}$, we overlay the anatomical boundaries on the original image. We extract the contours of each anatomical region and superimpose them as colored overlays.

$$\tilde{\mathbf{x}} = \mathbf{x} + \alpha \cdot \text{Contour}(\mathbf{m}), \quad (1)$$

where $\text{Contour}(\cdot)$ extracts the boundary pixels of each segmented region, and α controls the overlay intensity. This process keeps the original radiographic information while visually emphasizing anatomical structures relevant to clinical diagnosis (Figure 1).

Unlike hard masking approaches that completely remove regions outside anatomical structures, our soft overlay strategy keeps contextual information (e.g., mediastinal contours, rib shadows) that may provide additional diagnostic cues. This design choice reflects the clinical practice of holistic image interpretation while directing attention to key anatomical regions.

3.3 Model Architecture and Fine-tuning

We use BiomedCLIP [1] as our pretrained foundation model. BiomedCLIP has a Vision Transformer (ViT) encoder for images and a PubMedBERT encoder for text, jointly trained on PMC-15M image-text pairs using contrastive learning. While BiomedCLIP provides strong zero-shot capabilities through text-based queries, we focus on supervised fine-tuning for multi-label disease classification.

We test two fine-tuning strategies to understand how anatomical masking interacts with different adaptation methods.

Linear Probing (LP). We freeze the pretrained vision encoder and train only a randomly initialized linear classifier on top of the extracted image features. This setting isolates the quality of pretrained representations and tests whether anatomical masks can improve downstream performance without changing the encoder. Given an image \mathbf{x} (original or masked), we extract features $\mathbf{z} = \text{Encoder}(\mathbf{x})$ and apply a linear classifier $\mathbf{y} = \sigma(\mathbf{W}\mathbf{z} + \mathbf{b})$, where σ is the sigmoid function for multi-label classification.

Joint Fine-tuning (JFT). We jointly update both the vision encoder and the linear classifier during training. This allows the encoder to adapt to task-specific patterns and potentially learn to use anatomical cues more effectively. This setting tests whether end-to-end optimization can benefit from anatomical masking during training.

For both strategies, we train models on either original images or masked images, giving us four training setups (LP-Original, LP-Masked, JFT-Original, JFT-Masked). Each trained model is then evaluated on both original and masked test images, giving eight inference setups in total.

3.4 Training Objectives and Optimization

For multi-label classification, we use binary cross-entropy loss.

$$\mathcal{L} = -\frac{1}{C} \sum_{c=1}^C [y_c \log \hat{y}_c + (1 - y_c) \log(1 - \hat{y}_c)], \quad (2)$$

where y_c and \hat{y}_c denote the ground-truth and predicted probabilities for disease c .

4 Experimental Setup

4.1 Datasets and Preprocessing

MIMIC-CXR-JPG. We use the MIMIC-CXR-JPG dataset [3], a large collection of chest X-ray images with associated structured radiology reports and disease labels extracted using rule-based

Dataset	Modality	Description
MIMIC-CXR-JPG [3]	X-ray	Frontal chest X-rays with diagnostic labels
CheXmask [2]	Segmentation	Anatomical masks (lungs, heart) aligned with MIMIC-CXR

Table 1: Datasets used in this study. MIMIC-CXR-JPG provides chest X-ray images with diagnostic labels, while CheXmask provides corresponding anatomical segmentations.

NLP systems. We work with frontal-view images (PA and AP projections) and focus on five common pulmonary pathologies (Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion).

CheXmask. CheXmask [2] provides pixel-level anatomical segmentations for a subset of MIMIC-CXR images, covering major thoracic structures including left lung, right lung, and heart. The segmentations are generated using deep learning models and validated against expert annotations. We pair each MIMIC-CXR image with its corresponding CheXmask segmentation to create masked image variants.

Data Preprocessing. All images are resized to 224×224 pixels using bicubic interpolation to match the input resolution of BiomedCLIP’s vision encoder. We normalize pixel values to $[0, 1]$ and apply channel-wise standardization using ImageNet statistics. For masked images, we overlay anatomical boundaries as RGB contours with distinct colors for each structure (red for lungs, blue for heart). We keep only samples with at least one positive label among the five target diseases, and split the data into training, validation, and test sets.

4.2 Evaluation Metrics and Baselines

We evaluate model performance using classification accuracy averaged across all five disease categories. As a baseline, we report the zero-shot performance of the pretrained BiomedCLIP model using text prompts (e.g., a chest X-ray showing [disease]) without any fine-tuning.

4.3 Implementation Details

All experiments are implemented in PyTorch 2.6.0 using the Hugging Face Transformers library and OpenCLIP. We use the official BiomedCLIP checkpoint (microsoft/BiomedCLIP-PubMedBERT_256-vit_base_patch16_224) as our pretrained model. Training and evaluation are done on NVIDIA A100 GPUs with CUDA 11.8. Data preprocessing and analysis are done using NumPy, pandas, and PIL for image handling.

5 Results

5.1 Overall Performance Comparison

Table 2 shows the classification accuracy across different fine-tuning strategies and input setups. Several key observations emerge from these results.

Fine-tuning improves over zero-shot baseline. The pretrained BiomedCLIP model gets 23.7% accuracy on original test images without any fine-tuning. This shows limited zero-shot capability for this multi-label classification task. All fine-tuned variants do much better than this baseline, with the best model getting 45.1% accuracy.

Inference-time masking consistently improves performance. Across all fine-tuning setups, evaluating on masked test images gives higher accuracy than evaluating on original images. This pattern is particularly strong in the zero-shot setting and certain joint fine-tuning setups. This suggests that anatomical masks work as effective visual priors that guide the model toward relevant diagnostic regions.

Training on masked images shows mixed results. When comparing models trained on original vs. masked images within the same fine-tuning setup, results are less clear. For linear probing, training

Fine-tuning Method	Training Data	Test Accuracy	
		Original	Masked
<i>Baseline (Zero-shot)</i>	–	0.237	0.347
Linear Probing (Frozen Encoder)	Original	0.446	0.451
	Masked	0.421	0.422
Joint Fine-tuning (Encoder + Head)	Original	0.401	0.404
	Masked	0.348	0.432

Table 2: Classification accuracy comparison across fine-tuning strategies and input setups. Models are evaluated on both original and masked test sets. Bold indicates higher performance between original vs. masked test inputs for each training setup. Results show consistent improvement when evaluating on masked images, particularly in the zero-shot and masked-training settings.

on original images generally works better. But for joint fine-tuning evaluated on masked test images, training on masked data can give better performance. This suggests that the impact of training-time masking depends on whether the encoder is updated during fine-tuning.

5.2 Analysis by Disease Category

While we see consistent improvements from inference-time masking across most disease categories, the size of improvement varies by pathology. Diseases with spatially well-defined presentations (such as cardiac enlargement) appear to benefit more from anatomical masking, while pathologies with variable spatial distributions show more modest gains. This suggests that how well anatomical priors work depends on how closely the disease presentation aligns with the masked anatomical structures.

5.3 Effect of Fine-tuning Setup

Comparing linear probing vs. joint fine-tuning shows different sensitivities to anatomical masking. Linear probing gets strong overall accuracy, suggesting that the pretrained BiomedCLIP encoder already captures features relevant for chest X-ray classification. Joint fine-tuning shows variable performance across setups, with best results when both training and testing use masked images. This indicates that end-to-end optimization may benefit more from consistent anatomical emphasis across training and inference.

The performance differences between linear probing and joint fine-tuning suggest potential overfitting or disruption of pretrained features during encoder fine-tuning in some setups. This aligns with recent findings in transfer learning work showing that fine-tuning large pretrained models can sometimes hurt performance on specialized medical imaging tasks with limited data.

6 Discussion

6.1 Inference-Time Masking as Visual Prior

Our main finding is that anatomical masking improves performance at test time but not consistently during training. This suggests an interesting idea. Pretrained medical VLMs already encode spatial knowledge about anatomically relevant regions, and masking helps reveal or amplify these latent spatial biases without needing explicit supervision. This fits with recent work on attention visualization in medical imaging models, which shows that even models trained without explicit anatomical supervision tend to focus on relevant diagnostic regions.

The fact that zero-shot performance improves a lot with masked inputs (from 23.7% to 34.7%) provides strong evidence for this idea. Since the encoder has never been fine-tuned on the target task, the improvement must come from the masking itself redirecting the model’s feature extraction toward more informative regions. This suggests that anatomical masks work as a form of attention guidance that uses existing spatial knowledge rather than teaching new patterns.

6.2 Why Training on Masked Images Shows Mixed Results

The inconsistent benefits of training-time masking, particularly in some linear probing setups, can be explained by several factors. First, removing contextual information through masking may eliminate subtle cues that pretrained encoders have learned to use. For instance, surrounding anatomical structures and spatial context provide information that helps in localizing abnormalities.

Second, the contour overlays introduce artificial visual patterns that do not appear in natural radiographs, potentially creating a domain shift that makes learning harder. The encoder must adapt to both the diagnostic task and the new visual artifacts, which may complicate optimization, especially with limited data.

Finally, the difference between LP and JFT suggests that the frozen encoder in linear probing is less sensitive to training-time input variations, while joint fine-tuning can adapt to (but also potentially overfit to) the specific characteristics of masked training data.

6.3 Implications for Clinical Deployment

From a practical standpoint, our findings suggest that inference-time masking could be deployed as a lightweight enhancement to existing chest X-ray classification systems without needing model retraining. Given that anatomical segmentation models for chest X-rays are readily available and computationally cheap, overlaying masks at test time represents a low-cost intervention that could improve diagnostic accuracy.

Also, the visual highlighting of anatomical boundaries could improve interpretability for clinical users. Radiologists reviewing model predictions could benefit from seeing which anatomical regions the model is emphasizing, potentially increasing trust and making error detection easier. This aligns with growing emphasis on human-AI collaboration in medical imaging, where interpretability is as important as raw performance.

6.4 Limitations and Future Directions

Several limitations should be noted. First, our study focuses on a limited set of five pulmonary diseases from a single dataset (MIMIC-CXR). Whether these findings extend to other pathologies, imaging modalities, or patient populations remains to be checked. Second, we use CheXmask segmentations which, while validated, are automatically generated and may contain errors that propagate to downstream classification.

Third, our masking strategy is relatively simple (binary overlays of anatomical boundaries). More advanced approaches could weight different anatomical regions based on disease-specific priors or learn optimal masking strategies through meta-learning. Fourth, we evaluate only classification accuracy. Clinical metrics such as sensitivity, specificity, and alignment with radiologist decision-making would provide deeper insights into real-world utility.

Future work could explore several promising directions. Adaptive masking strategies that selectively emphasize regions based on preliminary model predictions could provide finer-grained spatial guidance. Combining anatomical masking with explicit attention mechanisms or multi-task learning objectives could further improve both performance and interpretability. Finally, extending this approach to other medical imaging modalities (e.g., CT, MRI) where anatomical segmentations are available could reveal whether these findings extend beyond chest X-rays.

7 Conclusion

We have tested the role of anatomical segmentation masks as lightweight visual priors for chest X-ray classification using pretrained medical vision-language models. Through systematic experiments across multiple fine-tuning setups and input configurations, we show that anatomical masking provides consistent benefits at test time, improving classification accuracy across most disease categories and training settings.

Our key finding is that inference-time masking outperforms training-time masking. This suggests that pretrained medical VLMs already encode spatial knowledge about relevant diagnostic regions, and that explicit anatomical overlays can effectively guide attention toward these areas without needing

architectural changes or additional supervision. This insight has practical implications for deploying chest X-ray classification systems in clinical settings, where lightweight, post-hoc interventions are often preferable to complex model retraining.

While our study focuses on a specific dataset and disease categories, the proposed approach is general and could be extended to other medical imaging domains where anatomical segmentations are available. By bridging the gap between data-driven learning and clinical anatomical knowledge, anatomical masking represents a promising direction for developing more accurate, interpretable, and trustworthy medical AI systems.

References

- [1] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2025.
- [2] Nicolas Gaggion, Candelaria Mosquera, Martina Aineseder, Lucas Mansilla, Diego Milone, , and Enzo Ferrante. Chexmask database: a large-scale dataset of anatomical segmentation masks for chest x-ray images (version 1.0.0). *PhysioNet*, 2025.
- [3] Alistair Johnson, Matthew Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. Mimic-cxr-jpg - chest radiographs with structured labels (version 2.1.0). *PhysioNet*, 2024.
- [4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.
- [5] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text, 2022.
- [6] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. Learning to exploit temporal structure for biomedical vision-language processing, 2023.
- [7] Xizhou Zhu, Dazhi Cheng, Zheng Zhang, Stephen Lin, and Jifeng Dai. An empirical study of spatial attention mechanisms in deep networks, 2019.
- [8] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, October 2019.
- [9] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang “Anthony” Chen. Chexplain: Enabling physicians to explore and understand data-driven, ai-enabled medical imaging analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI ’20, page 1–13. ACM, April 2020.