MMCLIP: Cross-Modal Attention Masked Modelling for Medical Language-Image Pre-Training

Biao Wu[1,4*] Yutong Xie[2*] Zeyu Zhang[3*] Vu Minh Hieu Phan[4]
Qi Chen[4] Ling Chen[1] Qi Wu[4†]

[1] University of Technology Sydney [2] Mohamed bin Zayed University of Artificial Intelligence
[3] The Australian National University [4] Adelaide University
[*] Equal Contribution [†] Corresponding Author

Abstract

Vision-and-language pretraining (VLP) in medicine leverages contrastive learning on image–text pairs, often enhanced with masked modeling. However, existing methods face two challenges: difficulty reconstructing key pathological features due to limited data, and reliance on either paired or image-only datasets without combining both. To address this, we propose **MMCLIP** (**M**asked **M**edical **C**ontrastive **L**anguage-**I**mage **P**re-training), which introduces two modules: *AttMIM*, masking image features highly correlated with text to improve reconstruction of fine medical details, and *EntMLM*, masking key medical entities in text and reconstructing them using visual cues. Furthermore, MMCLIP incorporates unpaired data through disease-kind prompts, achieving state-of-the-art performance in zero-shot and fine-tuning across five benchmarks. Code: https://github.com/AIGeeksGroup/MMCLIP.

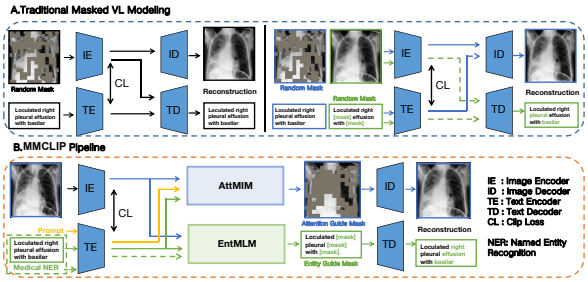


Figure 1: Modules A and B illustrate the distinctions between the existing VLP training frameworks and our proposed approach. Methods in module A only use the random mask strategy, and the reconstruction process does not involve multi-modal interaction. Module B shows that our method takes advantage of multi-modal interaction and uses an attention mechanism to guide the mask.

1 Introduction

Vision-and-language pretraining (VLP) has garnered increasing attention in the medical field, primarily due to its capability to transfer representations effectively across various downstream tasks, including zero- or few-shot recognition. This ability significantly reduces reliance on extensive annotated data and efficiently detects various pathologies. Most VLP approaches emphasize contrastive learning to understand the relationship between visual and linguistic elements (of Radiology et al., 2014) Masked modelling has progressed and yielded promising results in pure visual (He et al., 2022; Bao et al., 2021; Xie et al., 2022b), and textual (Kenton and Toutanova, 2019) self-supervised representation learning. Masked image modelling (MIM) reconstructs masked image segments using surrounding context, while masked language modelling (MLM) predicts masked words in sentences through adjacent context, both enhancing comprehension in their respective visual and linguistic domains. The success has inspired some researchers to explore the significance of MIM and MLM within the scope of VLP. This involves randomly masking elements in a single modality (Li et al., 2023; Sun et al., 2023) or across multiple modalities (Kwon et al., 2022; Chen et al., 2023) during VLP, as shown in Figure 1 A.

Applying MIM and MLM to the medical domain still presents significant challenges, primarily due to two key limitations. *First*, traditional random masking misaligns with the specific needs of medical data, which fails to capture pathological representations critical for clinical diagnosis. In MIM, random masking inadequately captures the subtle but critical variations in anatomical structures, crucial for precise medical diagnosis. Concurrently, in MLM, this approach can lead to a diminished focus on accurately predicting medical entities, generating less medically informative content. The latest research has started to explore attention-guided MIM (Wang et al., 2023; Kakogeorgiou et al., 2022), which are prone to finding crucial image regions for masking instead of ran-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2440–2455

dom masking. However, constrained by the scarcity of medical data, models still struggle to accurately identify important medical areas during masking. *Second*, MIM and MLM were initially conceived for single-modality self-supervised learning, not accounting for the strong interplay between medical images and clinical texts. This oversight limits their effectiveness in VLP scenarios where cross-modal interactions are vital. The advancement (Xie et al., 2023; Chen et al., 2022) has proven that integrating modality interaction in MIM/MLM can enhance the recognition ability of meaningful regions, but heavily depending on the medical image-text pairs.

This motivates us to propose novel multimodal **M**asked **M**edical **C**ontrastive **L**anguage-**I**mage **P**re-Training, named **MMCLIP**. Our MMCLIP is designed to accomplish three self-supervised learning tasks, as shown in Figure 1 B. First, we design a novel attention-masked image modelling module (AttMIM) for medical images. It detects discriminative features to mask by harnessing feature interactions between vision and language modalities. Specifically, our AttMIM blends masks generated through image self-attention, image-report cross-attention, and prompt-driven attention. While the first two modules mine instance-level interactions, the prompt-driven attention exploits global-level interactions between images and common disease terms, independent of paired data. As such, even having missing corresponding reports, our prompt-driven attention still localizes pathological features by leveraging the global affinity with frequent diseases. Secondly, we propose the entity-driven masked language modelling module (EntMLM) for medical reports, which masks informative entities and reconstructs the masked words. We also incorporate the learned image representations into the text decoder to aid the model in understanding medical entities with greater detail and nuance. Besides, we employ standard contrastive learning to align medical images and reports. To solve the inconsistent convergence rate problem between contrastive learning and masked modelling objectives, we first warm up the masked modelling tasks for specific iterations and then jointly train three tasks end-to-end.

Our MMCLIP is pre-trained on two large-scale medical datasets, one is MIMIC-CXR with paired chest X-ray images and reports, and the other is the PadChest with only chest X-ray images. The learned image representations are transferred to five downstream classification datasets under zero-shot and fine-tuning settings, all of which achieved SOTA respectively.

Our contributions are three-fold: (1) we present a novel cross-modal attention masking, which is the first work to explore the potential of attention masking for MIM and MLM within a unified medical VLP framework, offering a new perspective to enhance the accuracy of medical data representation learning; (2) we develop a novel blending masking strategy that integrates attention-guided masking to capture discriminative pathological features, and common disease-prompt masking to enable unsupervised learning without relying on paired reports; (3) we conduct extensive experiments on medical image classification downstream tasks with improved zero-shot and fine-tuning performance. Our method surpasses strong competitors like MedKLIP, GLORIA, and CheXzero.

2 Methodology

Figure 2 illustrates the pipeline of the proposed MMCLIP. MMCLIP excels in representation learning by using image-report contrastive learning and masked modelling across both paired and unpaired medical image-report data. Central to MMCLIP are two innovative modules: Attention-Masked Image Modeling (AttMIM) and Entity-Driven Masked Language Modeling (EntMLM).

AttMIM capitalizes on vision-language interactions to selectively mask discriminative features in medical images. It seamlessly integrates image self-attention, cross-attention with medical reports, and prompt-driven attention. This multifaceted approach allows for the effective identification of pathological features, even when corresponding reports are unavailable.

EntMLM, on the other hand, concentrates on masking and reconstructing significant entities within medical reports, utilizing insights from image representations. This integration facilitates a more nuanced and detailed comprehension of medical terminology, enhancing the capabilities of both visual and textual learners within the MMCLIP framework.

2.1 Image and Text Encoders

Image encoder. We employ the Vision Transformer (ViT) (Dosovitskiy et al., 2020) with a patch size of 16 as the image encoder, initialized with the weights of the clip-base, to process the given input medical images x. Initially, x is transformed into

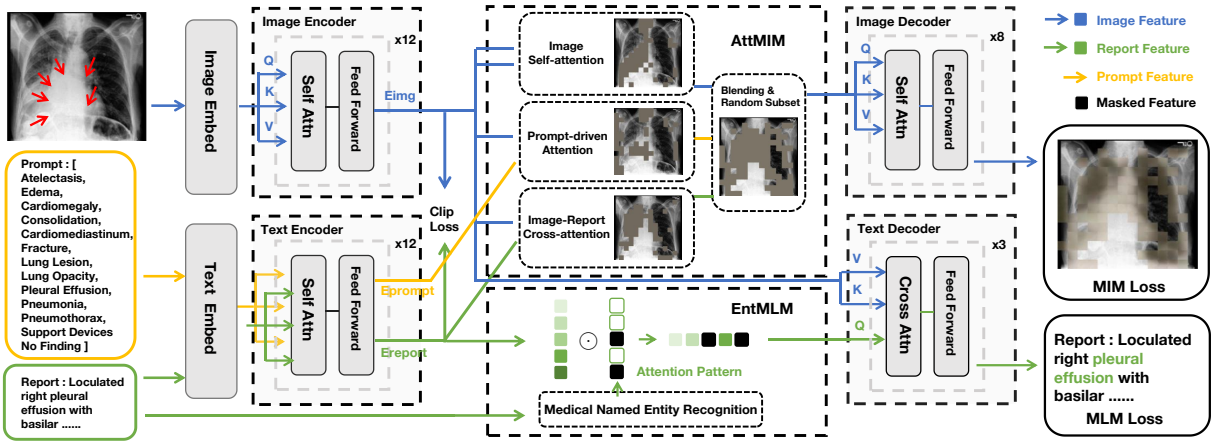


Figure 2: Our MMCLIP framework builds upon CLIP, integrating MIM and MLM modules, with a redesigned masking strategy to enhance model representation. The key contributions include three simple yet effective designs: generating masks through feature interaction, fusing masks to augment adaptability, and refining the text masking strategy with Medical Entity Recognition. Incorporating these designs into our multimodal pre-training framework significantly boosts the model's zero-shot performance.

a sequence of flattened patches. These patches are subsequently embedded and introduced into a 12-layer transformer followed by a linear projection layer, denoted as $F(\cdot)$ and $Linear(\cdot)$, resulting in encoded representations of visual tokens:

$$E_{\text{img}} = Linear(F(x)) \in \mathbb{R}^{N_{\text{img}} \times C}, \quad (1)$$

where C is the encoding dimension and N_{img} denotes the number of patches.

Text encoder. The text encoder also consists of a 12-layer transformer and a linear projection layer, initialized with the weights of the clip-base, denoted as $T(\cdot)$ and $Linear(\cdot)$. Our method involves the text t_{report} and t_{prompt} being tokenized using Byte Pair Encoding (BPE) into N_{report} and N_{prompt} tokens, which are then transformed into embeddings. These embeddings serve as the input for the text encoder. The encoder $T(\cdot)$ processes these to create subword features:

$$E_{\text{report}} = Linear(T(t_{report})) \in \mathbb{R}^{N_{\text{report}} \times C}, \quad (2)$$

$$E_{\text{prompt}} = Linear(T(t_{prompt})) \in \mathbb{R}^{N_{\text{prompt}} \times C} \quad (3)$$

where C represents the feature dimension.

2.2 Attention-masked Image Modeling

To address the limitations of random masking in medical imaging, we propose an attention-masked image modelling (AttMIM) module. It is divided into two phases: in the attention extraction phase, it uses an attention mechanism to focus on features that are relevant to the medical reports, disease prompts, and the image features themselves; in the attention-based mask generation and blending phase, it merges highly activated regions of these features to generate corresponding masks that accurately mask key pathology features (see Figure 3).

2.2.1 Attentions extraction

Image-report cross-attention effectively identifies image regions that are relevant to the diagnostic report. The detailed information contained within these reports directs the model's focus towards the most crucial pathological features. Given the diversity of human organs, their unique physiological structures, and the specific pathological details associated with each, diagnostic reports typically emphasize areas presenting abnormalities. This guidance enables the model to prioritize regions abundant in diagnostic information, ensuring a more focused and informed analysis of medical images.

Based on the cross-attention mechanism, the image features after interacting with the report features can be obtained as follows:

$$A_r = \text{softmax}\left(\frac{E_{\text{img}} E_{\text{report}}^T}{\sqrt{d_k}}\right) \cdot E_{\text{report}}, \quad (4)$$

where d_k is the dimension of E_{report}.

Prompt-driven attention can effectively recognize the image regions related to the disease prompts denoted as in Figure 3. Meanwhile, as shown in Figure 2, the disease prompt contains 14 common diseases and medical devices under the chest organ, and this rich variety of diseases

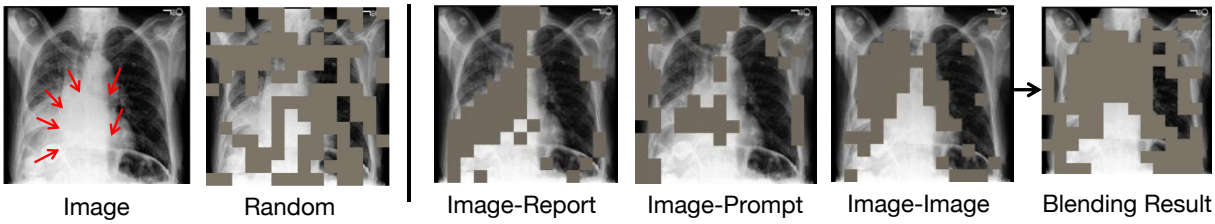


Figure 3: Different masking strategies result in varying concerns regarding mask application. Masks that are specifically tailored to the pathological characteristics of the lesion area are more effective than those applied randomly.

helps the model to focus on all potential lesions, thus improving the compatibility of MIM with all candidate regions where lesions may be present.

Besides, including disease prompts offers a significant advantage in scenarios where only images are available and the corresponding diagnostic reports are missing, acting as a bridge to fill the information gap. By supplementing image data, these prompts guide the model in identifying relevant pathological features, enabling effective representation learning and a deeper understanding of medical images without needing textual descriptions. Based on the cross-attention mechanism, the image features after interacting with the disease prompt features can be obtained as follows:

$$A_p = \text{softmax}\left(\frac{E_{\text{img}} E_{\text{prompt}}^T}{\sqrt{d_k}}\right) \cdot E_{\text{prompt}}, \quad (5)$$

where d_k is the dimension of E_{prompt}.

Image self-attention directs the model to focus on the most challenging parts of the image features during convergence, which tend to be more difficult regions (Wang et al., 2023), thus improving the model's ability and efficiency in understanding the image (Zhang et al., 2024; Ji et al., 2024). Based on the self-attention mechanism, the image features after interacting with the image features themselves can be obtained as follows:

$$A_i = \text{softmax}\left(\frac{E_{\text{img}} E_{\text{img}}^T}{\sqrt{d_k}}\right) \cdot E_{\text{img}}, \quad (6)$$

where d_k is the dimension of E_{img}.

2.2.2 Attention-based mask generation and blending.

We introduce the function $M(A, \lambda)$, as shown in Fig. 7, which retrieves indices corresponding to the top values in feature matrix A, with the number of indices determined as a proportion λ_1 of the matrix's size. This function effectively transforms these indexes into a mask corresponding to the input matrix, which is the high-activated attention masking results in the attention extraction. We blend the masks produced by different strategies in a union manner and define this result as M_b, formulated as:

$$M_b = M(Ar; \lambda_1) \cup M(Ap; \lambda_1) \cup M(Ai; \lambda_1). \quad (7)$$

Subsequently, a subset is randomly selected in λ_2 proportion to constitute the final outcome of AttMIM masking process. We defined it as $\mathbb{M}_{\text{AttMIM}}$:

$$\mathbb{M}_{\text{AttMIM}} = \text{RandomSubSet}(M_b, \lambda_2). \quad (8)$$

Finally, we apply the mask $\mathbb{M}_{\text{AttMIM}}$ to E_{img}, rendering it as the feature input for MIM, which can be expressed as

$$\text{MIM}_{\text{input}} = \mathbb{M}_{\text{AttMIM}} \cdot E_{\text{img}}. \quad (9)$$

The comparison results in Figure 3 show that the interactions of image self-attention, image-report cross-attention, and prompt-driven attention focus on different lesion areas to varying degrees. By integrating the strengths of these, AttMIM obtains a final mask that balances specificity and comprehensiveness.

2.3 Entity-driven Masked Language Modeling

Unlike natural language where fluency is emphasized, medical texts prioritize the accuracy of medical entity terms over textual smoothness. The typical autoregressive models in general MLM tasks may not be essential for medical MLM. Thus, we design an EntMLM module to target disease entities for masking.

For EntMLM, we adopt ScispaCy a specialized Named Entity Recognition (NER) tool for biomedical and scientific text analysis. It excels in identifying medically relevant entities, such as diseases,

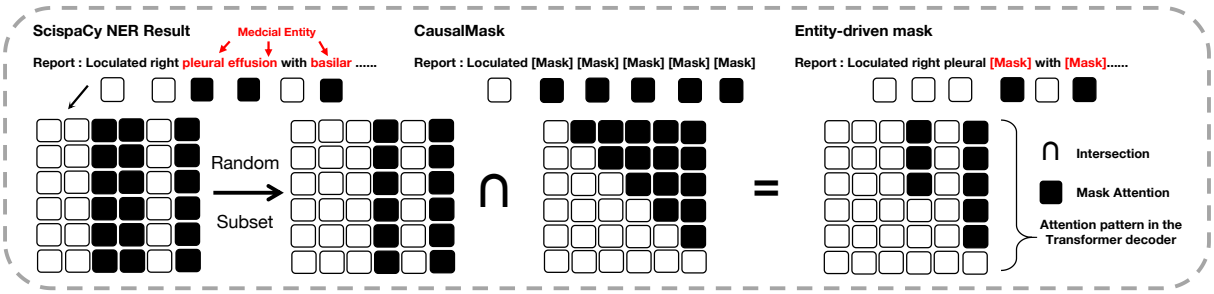


Figure 4: Illustration of entity-driven mask generation, which combines the CausalMask with the mask guided by the medical NER to generate the final result.

drugs, medical terms, treatments, and symptoms, from healthcare-related documents like medical records and research papers. As shown in Figure 4, we first identify medical entities in the report using ScispaCy, then retrieve their position indexes, and use these indexes to guide the mask generation. The resulting mask can be denoted as M_e.

Notably, this mask is not directly applied to the original report text but is instead integrated with an upper triangular mask in the text decoder, which we defined as CausalMask, where their intersection is taken. CausalMask plays a key role in self-supervised learning tasks for natural language processing, especially in generative tasks. It facilitates the learning of the structure of the language and the generation of coherent text by ensuring that the model's predictions rely only on past and present information. The implementation of this approach enhances the ability of the model to process and generate natural language more efficiently. Hence, the mask result of EntMLM, we defined it as $\mathbb{M}_{\text{EntMLM}}$:

$$\mathbb{M}_{\text{EntMLM}} = \text{RandomSubSet}(M_e, \lambda_3) \cap \text{CausalMask}. \tag{10}$$

As shown in Figure 4, this approach allows the model to temper its focus on text fluency, thereby enhancing its integration with image features and improving its understanding of medical semantic entities. Finally, we apply the mask $\mathbb{M}_{\text{EntMLM}}$ to E_{report}, rendering it as the feature input for MLM, formulated as:

$$\text{MLM}_{\text{input}} = \mathbb{M}_{\text{EntMLM}} \cdot E_{\text{report}}. \tag{11}$$

2.4 Objective Functions

Image-Report Alignment. Inspired by (Li et al., 2021), our approach focuses on aligning image and text features prior to their fusion by the multimodal decoder. This pre-alignment strategy simplifies the process of cross-modal learning for the multimodal decoder, facilitating a more effective integration of visual and textual information. We follow (Radford et al., 2021) and compute the objective alignment function $Loss_{\text{align}}$ by exploiting the fine-grained correspondences between E_{img} and E_{report}, formulated as:

$$\text{Loss}_{\text{align}} = -\frac{1}{N}\Bigg(\underbrace{\sum_{i=1}^{N}\log\frac{\exp(x_i^\top y_i/\sigma)}{\sum_{j=1}^{N}\exp(x_i^\top y_j/\sigma)}}_{\text{image-to-text}} + \underbrace{\sum_{i=1}^{N}\log\frac{\exp(y_i^\top x_i/\sigma)}{\sum_{j=1}^{N}\exp(y_i^\top x_j/\sigma)}}_{\text{text-to-image}}\Bigg) \tag{12}$$

where x_i and y_j are the E_{img} in the i-th pair and the E_{report} in the j-th pair, respectively. N is the batch size, and σ is the temperature to scale the logits.

Image Reconstruction. The image decoder, inspired by (He et al., 2022), consists of an 8-layer transformer, which is essential for pre-training image reconstruction. This component integrates encoded visible patches and mask tokens to reconstruct the image. We defined this decoder as $D_{\text{MIM}}(\cdot)$. Reconstruction accuracy is quantified by the Mean Squared Error loss computed solely on masked patches, shown as:

$$Loss_{\text{MIM}} = \left\| y^{\text{MIM}}, x \right\|^2, \quad \text{where } y^{\text{MIM}} = D_{\text{MIM}}(\text{MIM}_{\text{input}}). \tag{13}$$

Report Reconstruction. The text decoder in our model, following (Yu et al., 2022), features a 3-layer transformer. The multi-modal decoder combines visual and textual data, enhancing the representations. It uses a combination of causal and

disease-guided masks and integrates visual encoder outputs via alignment from contrastive learning and the cross-attention mechanism.

Moreover, when introducing multi-modal information, the model can leverage visual information to predict masked textual entities, significantly aiding the model's understanding during the pre-training phase. This allows the decoder to predict text while simultaneously incorporating image context. Such an approach guarantees efficient and flexible integration of diverse modalities in multi-modal learning tasks. Meanwhile, the training of text generation is to maximize the conditional likelihood of the paired text y under the forward autoregressive factorization:

$$Loss_{\text{MLM}} = -\sum_{t=1}^{T} \log P_\theta(y_t|y_{<t}, x), \quad \text{where } y_t \in \text{MLM}_{\text{input}}. \tag{14}$$

Final objective function. In the medical field, the scenario of unpaired data is common and significant, mainly due to data access limitations, ethical and privacy considerations, and the diversity and complexity of medical data. These factors make it difficult to obtain perfectly paired medical data in practical applications and also highlight the need to develop advanced algorithms capable of handling such data. MMCLIP Integrating prompt-driven attention and image self-attention facilitates the generation of masks for image data, which lacks corresponding reports. This approach improves the training process, supports compatibility with unpaired image data, and ensures that even data solely involved in the MIM task contributes to enhanced model representation learning. Our whole training process mixes paired data and unpaired data, thus the final objective function consists of these two parts as well.

As shown in Figure 2, the paired data effectively contributes to the MIM task by the image-report cross-attention, prompt-driven attention, and image self-attention together. Besides, it aids in the MLM and alignment tasks through the combined use of image and report representations, demonstrating the advantages of multimodal feature interaction. The objective function can be defined as:

$$Loss_{\text{pair}} = Loss_{\text{align}} + Loss_{\text{MIM}} + Loss_{\text{MLM}}. \tag{15}$$

Meanwhile, the unpaired data is unable to complete the MLM and alignment tasks due to the lack of corresponding reports, but can effectively participate in the MIM task with the help of prompt-driven attention and image self-attention.

The objective function can be formulated as:

$$Loss_{\text{unpair}} = Loss_{\text{MIM}}. \tag{16}$$

The final objective is the combination of $Loss_{\text{pair}}$ and $Loss_{\text{unpair}}$ as:

$$Loss_{\text{final}} = Loss_{\text{pair}} + Loss_{\text{unpair}}. \tag{17}$$

3 Experiments

3.1 Dataset

The experimental data is divided into pre-training data and downstream task data. The pre-training data includes MIMIC (Johnson et al., 2019) data that can complete multi-modal tasks and Padchest (Bustos et al., 2020) data used as unpaired image-only data. The downstream task data includes binary classification datasets CovidX (Pavlova et al., 2022) and Pneumonia (Anouk Stein, 2018), as well as multi-label classification of 14 types of diseases datasets CheXpert (Irvin et al., 2019) and Xray14 (Wang et al., 2017). Please refer to *Appendix* for details.

3.2 Main Results

Zero-shot Classification. We compare our MMCLIP with different pre-training methods, including ConVIRT (Zhang et al., 2022), GLORIA (Huang et al., 2021), BioViL (Bannur et al., 2023), BioViL-T(Bannur et al., 2023), CheXzero (Tiu et al., 2022), and MedKlip (Wu et al., 2023b), for zero-shot classification. The results in Table 1 and Table 2 show that our MMCLIP exhibits superior zero-shot performance across all these scenarios on three datasets, even for the unseen class recognition on the PadChest dataset. This may be because it is difficult for existing models to understand the feature information contained in the whole image and report, while the proposed AttMIM and EntMLM can utilize the rich multimodal information of the medical data itself to help our MMCLIP to understand the more critical medical knowledge, and even understand the complex medical features that have not been mentioned in the report of the training set.

Models	CheXpert			PadChest-Seen 8			Pneumonia		
	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC
ConVIRT (Zhang et al., 2022)	52.10	35.61	57.43	74.31	23.58	80.12	79.21	55.67	75.08
GLORIA (Huang et al., 2021)	54.84	37.86	60.70	74.56	24.02	80.75	70.37	48.19	70.54
BioViL (Bannur et al., 2023)	60.01	42.10	66.13	71.17	20.75	79.58	84.12	54.49	74.43
BioViL-T (Bannur et al., 2023)	70.93	47.21	69.96	76.17	27.41	85.32	86.03	62.56	80.04
CheXzero (Tiu et al., 2022)	87.90	61.90	81.17	76.75	28.68	87.37	85.13	61.49	78.34
MedKlip (Wu et al., 2023b)	89.97	63.67	84.32	84.97	35.17	90.68	86.94	63.42	80.02
MMCLIP	90.53	67.05	86.04	85.79	33.50	91.19	86.16	62.57	80.46
+ Ensemble	**91.45**	**69.18**	**86.28**	**86.31**	**35.61**	**92.16**	**86.95**	63.30	**80.51**

Table 1: Zero-shot performance of various pre-training models on three datasets.

Specifically, in Table 1, compared to the second-best MedKlip, the AUC scores of our MMCLIP have increased as follows: from 89.97 to 90.53 on the CheXpert dataset, from 84.97 to 85.79 on the PadChest-seen dataset, and from 85.94 to ours 86.16 on the RSNA Pneumonia dataset. Meanwhile, we used the ensemble learning method to combine the weights obtained from 6 different groups of parameters, which can improve the performance further on most datasets.

We comparatively analyze the zero-shot performance of MMCLIP on the PadChest dataset for unseen classes. Only 8 categories in PadChest strictly match the visible categories in the MIMIC data, so we only report the test results of 8 visible categories in Table 1. There are 12 categories that either cannot be strictly matched, but cannot be counted as unseen classes, or do not appear in the test set, so we report at most 173 unseen classes in Table 2. Our analysis of the 10 and 20 most numerous of these unseen classes reveals that the zero-shot performance of MMCLIP has a 4% to 5% lift.

PadChest	Unseen 10	Unseen 20	Unseen 173
CheXzero (Tiu et al., 2022)	59.75	63.11	65.76
MedKlip (Wu et al., 2023b)	60.07	53.73	64.16
MMCLIP	**65.09**	**67.14**	**70.00**

Table 2: Comparison of zero-shot AUC scores on the PadChest dataset. Unseen 10 represents the 10 categories with the highest number of unseen categories.

Fine Tuning. Existing multi-modal frameworks are constrained as they require data that include both text and images, which limits the potential for the image encoder to leverage image-only data to improve its representation capabilities. To address this, we introduce a training approach designed to work effectively with unpaired, single-modal data. This new method allows the image encoder to better use image-only data, enhancing its representation power and overall performance. Our experiments involved two datasets, where we used 1%, 10%, and 100% of the data for fine-tuning, aligning with the methods in previous works (Wu et al., 2023b). As shown in Figure 5, our model significantly outperforms existing models in terms of AUC scores across all datasets. Meanwhile, the performance of the model was further improved by incorporating image data from PadChest, demonstrating that compatibility with unpaired data is vital for representation learning.

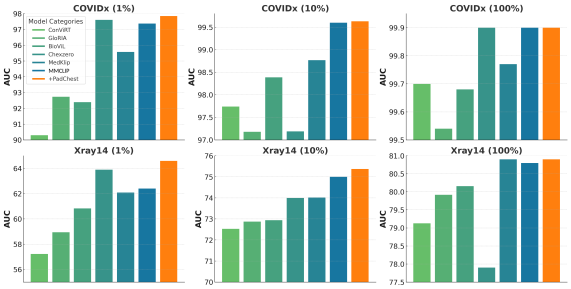


Figure 5: Performance comparison under different labeled data for fine-tuning. Unpaired data like PadChest can work with pair data to improve image encoder performance on the fine-tuning task.

3.3 Ablation Study

We noticed the original CheXpert validation set's limitations due to its small size and inconsistent performance tendency with the test set. To improve evaluation reliability, we crafted a new validation set with 2200 image-text pairs from existing datasets, which was used for all subsequent experiments, offering a more stable basis for performance assessment.

Improvements from AttMIM. As shown in Table 3, when we randomly mask 75% of the features according to the setting of MIM in MAE (He et al., 2022) and directly splice this module with CLIP

Init	Mask Method	AUC	F1
no warm up	clip (Baseline)	72.43	34.24
no warm up	+ MIM Random	72.85	34.49
warm up	+ MIM Random	73.98	34.80
warm up	+ Image Report	74.67	34.71
warm up	+ Prompt-driven	74.91	34.59
warm up	+ Image Self	74.84	34.70
warm up	+ AttMIM	**74.96**	**35.22**

Table 3: Zero-shot AUC on PadChest. 'Unseen 10' refers to the 10 most frequently unseen categories.

Mask Method	Mask Ratio	AUC	F1
AttMIM	-	74.96	35.22
+ MLM Full	1.0	74.75	34.79
+ MLM Random	0.5	75.52	34.51
+ EntMLM	0.5	74.98	34.49
+ EntMLM	0.3	75.44	34.88
+ EntMLM	0.1	74.67	34.74
+ EntMLM	0.2	**75.78**	**35.02**

Table 4: Performance of different text features masking strategies on the new CheXpert validation set.

and initialize it with the weights of MAE decoder, we can see that the model performance improves by 0.42% of AUC score when we train it directly. However, CLIP and MIM rely on different training strategies, with the latter tending to require smaller learning rates and longer training epochs. To harmonize the two, we first train the image encoder via the MIM reconstruction task for 15,000 iterations, as warm Up process. With the Warm Up, the model further improves the AUC by 1.1%. Meanwhile, we comparatively validate varying single strategies with mask ratios of 0.75, and the AttMIM strategy with the fusion of the three strategies. Experimental results show that three attention-based mask strategies can help the model understand key features better than random. As a fusion of the three strategies, AttMIM can more comprehensively improve related capabilities.

AttMIM has two adjustable parameters: the mask ratio of a single strategy and the final mask ratio after multi-strategy blending. As shown in Figure 6a in the appendix, after comprehensively comparing the values of AUC and F1, we concluded that the best settings are 0.7 and 0.75, which means that all single strategies mask 70% of the image features, and randomly select a subset after merging these masks, and finally mask 75% of image features.

Improvements from EntMLM. After splicing the MLM module with the existing structure, we first tried the Full mask and the Random mask based on the Causal Mask. The former did not make any modifications and essentially used MLM to complete the captioning task. The latter only randomly selected half of the Causal Mask to complete MLM tasks. As shown in Table 4, we can find that a random mask 50% is better than a Full mask. This may be because the task of predicting medical entities in medical texts is difficult. When there is a 50% probability of seeing these entity words, it can help the model better learn the knowledge in the report. This highlights the importance of paying more attention to medical entity words. Since the medical text is a highly structured task focusing on the accuracy of medical entities rather than text fluency, the EntMLM we proposed mainly masks and reconstructs medical entity words. Experimental results confirmed this, highlighting the importance of mask medical entities.

EntMLM has only one learnable parameter, unlike MLM where the mask ratio is a percentage of the sentence length. The ratio of EntMLM is the mask ratio for the medical entity words detected by NER. In Figure 6b, when the mask ratio is 0.2, the comprehensive indicators of AUC and F1 reach the best.

4 Conclusion

In this paper, we propose MMCLIP, an advanced medical VLP framework that aims to overcome major challenges in medical image and text analysis by introducing an innovative cross-modal attention masking modeling mechanism. By fusing AttMIM and EntMLM, MMCLIP can accurately capture key pathological features in medical data and improve the model's ability to understand complex medical scenes. In addition, we also use disease prompts to make the model compatible with unpaired data. Through pre-training on two large-scale medical datasets and validation on multiple downstream tasks, MMCLIP demonstrates its superior performance, not only achieving state-of-the-art results in the zero-shot setting but also in the fine-tuning setting. This result proves the practicality and effectiveness of MMCLIP in deep learning applications in the medical field.

References

et al. Anouk Stein. 2018. Rsna pneumonia detection challenge.

Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. 2023. Foundational models defining a new era in vision: A survey and outlook. *arXiv preprint arXiv:2307.13721*.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, and 1 others. 2023. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15016–15027.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797.

Cheng Chen, Aoxiao Zhong, Dufan Wu, Jie Luo, and Quanzheng Li. 2023. Contrastive masked image-text modeling for medical visual representation learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 493–503. Springer.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. 2022. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 679–689. Springer.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. 2021. 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*, pages 3942–3951.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas Montine, and James Zou. 2023. Leveraging medical twitter to build a visual–language foundation model for pathology ai. *bioRxiv*, pages 2023–03.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, and 1 others. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Yiping Ji, Hemanth Saratchandran, Cameron Gordon, Zeyu Zhang, and Simon Lucey. 2024. Sine activated low-rank matrices for parameter efficient learning. *arXiv preprint arXiv:2403.19243*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. 2022. What to hide from your students: Attention-guided masked image modeling. In *European Conference on Computer Vision*, pages 300–318. Springer.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Gukyeong Kwon, Zhaowei Cai, Avinash Ravichandran, Erhan Bas, Rahul Bhotika, and Stefano Soatto. 2022. Masked vision and language modeling for multi-modal representation learning. *arXiv preprint arXiv:2208.02131*.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language

representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, and 1 others. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer.

Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. 2023. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23390–23400.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-clip: Contrastive language-image pre-training using biomedical documents. *arXiv preprint arXiv:2303.07240*.

Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, and 1 others. 2019. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The lancet digital health*, 1(6):e271–e297.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. 2022. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pages 529–544. Springer.

Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rueckert. 2022. Radiological reports improve pre-training for localized imaging tasks on chest x-rays. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 647–657. Springer.

American College of Radiology, American Association of Physicists in Medicine (AAPM), and 1 others. 2014. Society for imaging informatics in medicine (siim). *Practice guideline for digital radiography (Resolution 42, adopted in 2007)[visionato il 4 dicembre 2012]. Disponibile su: www. siimweb. org*.

Maya Pavlova, Naomi Terhljan, Audrey G Chung, Andy Zhao, Siddharth Surana, Hossein Aboutalebi, Hayden Gunraj, Ali Sabri, Amer Alaref, and Alexander Wong. 2022. Covid-net cxr-2: An enhanced deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Frontiers in Medicine*, 9:861680.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. 2022. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406.

Haochen Wang, Kaiyou Song, Junsong Fan, Yuxi Wang, Jin Xie, and Zhaoxiang Zhang. 2023. Hard patches mining for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10375–10385.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. 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*, pages 2097–2106.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*.

Biao Wu, Yutong Xie, Zeyu Zhang, Jinchao Ge, Kaspar Yaxley, Suzan Bahadir, Qi Wu, Yifan Liu, and Minh-Son To. 2023a. Bhsd: A 3d multi-class brain hemorrhage segmentation dataset. In *International Workshop on Machine Learning in Medical Imaging*, pages 147–156. Springer.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023b. Medklip: Medical knowledge enhanced language-image pre-training. *medRxiv*, pages 2023–01.

Yutong Xie, Lin Gu, Tatsuya Harada, Jianpeng Zhang, Yong Xia, and Qi Wu. 2023. Medim: Boost medical image representation via radiology report-guided masking. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 13–23. Springer.

Yutong Xie, Jianpeng Zhang, Yong Xia, and Qi Wu. 2022a. Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier. In *European Conference on Computer Vision*, pages 558–575. Springer.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. 2022b. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3208–3216.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2022. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR.

Zeyu Zhang, Xuyin Qi, Mingxi Chen, Guangxi Li, Ryan Pham, Ayub Zuhair, Ella Berry, Zhibin Liao, Owen Siggs, Robert Mclaughlin, and 1 others. 2024. Jointvit: Modeling oxygen saturation levels with joint supervision on long-tailed octa. *arXiv preprint arXiv:2404.11525*.

Zeyu Zhang, Xuyin Qi, Bowen Zhang, Biao Wu, Hien Le, Bora Jeong, Minh-Son To, and Richard Hartley. 2023a. Segreg: Segmenting oars by registering mr images and ct annotations. *arXiv preprint arXiv:2311.06956*.

Zeyu Zhang, Bowen Zhang, Abhiram Hiwase, Christen Barras, Feng Chen, Biao Wu, Adam James Wells, Daniel Y Ellis, Benjamin Reddi, Andrew William Burgan, Minh-Son To, Ian Reid, and Richard Hartley. 2023b. Thin-thick adapter: Segmenting thin scans using thick annotations. *OpenReview*.

Hong-Yu Zhou, Xiaoyu Chen, Yinghao Zhang, Ruibang Luo, Liansheng Wang, and Yizhou Yu. 2022. Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. *Nature Machine Intelligence*, 4(1):32–40.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. 2021. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*.

Appendix

A Implementation Details

We initialize the image and text encoders with official clip pre-training weights. The experiments are conducted using a single NVIDIA A100. SGD is employed as the network optimizer with a momentum of 0.9 and an initial learning rate of 5e-5. The input resolution is set to 224 × 224, and the batch size is 64. For a warm-up, we first optimize the masked modeling objectives for 15,000 iterations and then jointly train all objectives end-to-end. After pre-training, we keep only the image and text encoders for downstream tasks. Using the CLIP (Radford et al., 2021) approach, zero-shot classification is done by matching the image and category name features from both encoders. Meanwhile, the image encoder can be used independently for fine-tuning classification tasks. About evaluation metrics, following previous works (Tiu et al., 2022; Xie et al., 2023), we adopt the area under the ROC curve (AUC), ACC, and F1 scores, which are standard metrics to evaluate classification tasks.

B Pre-training Dataset

MIMIC-CXR (Johnson et al., 2019) encompasses more than 227k investigations with paired imagery and report data, originating from 65,379 individuals across varied scanning instances. Each investigation might contain one or two images (diverse scanning perspectives), summing up to a total of 377,110 images.

PadChest (Bustos et al., 2020) is a comprehensive, chest x-ray dataset with 160k+ images from 67k patients, acquired at San Juan Hospital, Spain (2009-2017), inclusive of six varied positional views and rich associated patient/report data. Reports, categorized into 174 radiographic findings, 19 differential diagnoses, and 104 anatomical locations, utilize UMLS terminology and incorporate a hierarchical taxonomy. Limited by the fact that PadChest pairs are labelled in Spanish, this experiment only uses the image of this dataset.

C Datasets for Downstream Tasks

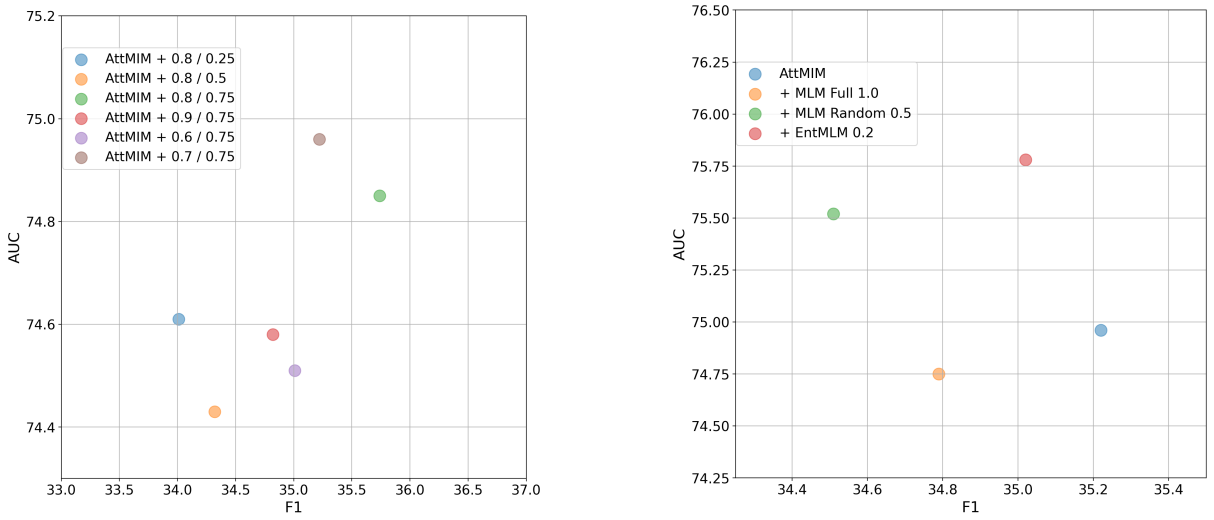
CheXpert (Irvin et al., 2019), is a large-scale dataset, containing 224,316 labelled chest X-rays from 65,240 patients. It covers 14 conditions including Fracture, Edema, Consolidation, Enlarged Cardiom, Cardiomegaly, Lung Lesion, Lung Opacity, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Support Devices, and No Finding. This dataset is involved in zero-shot and linear probing in our downstream tasks, both using the officially released test set with 500 images, which is consistent with the existing work (Wu et al., 2023b; Xie et al., 2023).

COVIDx (Pavlova et al., 2022) datasets is used for COVID-19 diagnosis. COVIDx has 29,986 images from 16,648 patients, used for the binary classification task of COVID symptoms, with a 70%/20%/10% split for training, validation, and testing. This dataset is utilized for fine-tuning purposes, following the existing work (Wu et al., 2023b).

Pneumonia(Anouk Stein, 2018) comprises over 260,000 frontal view chest X-ray images along with their corresponding pneumonia opacity masks, as well as support binary classification for pneumonia symptom, which were collected by the Radiological Society of North America. Following the settings of previous work(Wu et al., 2023b), this dataset is divided into three parts: 60% for training, 20% for validation, and 20% for testing.

Xray14 (Wang et al., 2017) includes 112,120 frontal view X-ray images, each annotated with 14 disease labels derived from corresponding radiology reports. Following the methodology outlined in previous work(Wu et al., 2023b), the dataset is divided into three subsets: 70% of the data is allocated for training, 10% for validation, and 20% for testing. This division is carried out at the patient level to ensure that images from the same patient are not present across different subsets, thereby avoiding any overlap between the training, validation, and testing groups.

PadChest testset encompasses 193 different categories, which can be used to evaluate zero-shot classification performance for unseen classes. We first identify 20 categories that were present during the pretraining phase. Then, we pinpoint categories not seen during pretraining within the test set. Next, we rank these unseen categories based on their frequency and select groups of the top 10 and top 20 most frequent unseen categories, as well as a group including all 173 unseen categories. Performance tests for zero-shot classification are conducted separately on each of these groups.



(a) AUC and F1 under different mask ratios of AttMIM.

(b) AUC and F1 under different mask ratios of EntMLM.

Figure 6: AUC and F1 under different mask ratios of (a) AttMIM and (b) EntMLM.

D Related Works

D.1 Medical Vision Language Pretraining

Driven by the effectiveness of self-supervised pre-training methods in NLP and CV, there is promising potential for customizing VLP methods to medical imaging analysis, such as lesion detection (Liu et al., 2019) and segmentation (Wu et al., 2023a; Zhang et al., 2023b,a). The most recent research exploring VLP is now centred around: improving feature extraction capabilities (Lu et al., 2019; Li et al., 2019); improving model structure (Li et al., 2020; Su et al., 2019); improving training methods (Li et al., 2019; Chen et al., 2020); improving model compatibility with different modalities (Xie et al., 2022a); introducing preprocessing modules (Mu et al., 2022); introducing prior knowledge (Wu et al., 2023b). As an application and extension of VLP in the medical field, Medical VLP focuses on understanding the content of medical images and texts. The latest outstanding research mainly focuses on the following three directions: 1) Model structures, particularly improvements in dual encoders (Jia et al., 2021; Yao et al., 2021) and fusion encoders (Li et al., 2019; Kim et al., 2021; Tan and Bansal, 2019; Yu et al., 2021); 2) Scaling up training data, by collecting high-quality medical multi-modal datasets from various platforms to enhance representational learning (Huang et al., 2023; Lin et al., 2023; Awais et al., 2023); 3) utilization of medical-specific prior knowledge to enhance model representational performance through prompts (Wu et al., 2023b; Wang et al., 2022).

D.2 Mask Image or Language Modelling

The development of MIM in computer vision, represented by MAE (He et al., 2022), SimMIM (Xie et al., 2022b), and iBOT (Zhou et al., 2021), marks a significant advance in self-supervised learning. MAE and SimMIM innovate in image masking, while iBOT combines these advances with the pre-training technology of self-distillation. Recent approaches have extensively explored various fusions of SSL techniques. For example, CTITM (Chen et al., 2023) utilizes MIM and MLM to augment CLIP, thereby enhancing the expressive power of the visual coder. MaskVLM (Kwon et al., 2022) not only effectively integrates CLIP, MIM, and MLM, but also innovatively explores strategies for images and text to guide each other in generating masks, thus optimizing the synergy between modalities. However, these methods have limitations: they either don't make full use of the available multimodal information, depending instead on features from a single mode to perform tasks, or their exploration of masking strategies is confined to using random methods.

D.3 Attention Masked modelling

Traditional MIM pre-training methods typically require the model to predict the content of masked image patches using predefined masking strategies such as random masking, block-wise masking, and uniform masking. However, we believe that merely solving these tasks is insufficient; the model also needs to learn how to create challenging tasks.

In the realm of attention-based mask modeling,

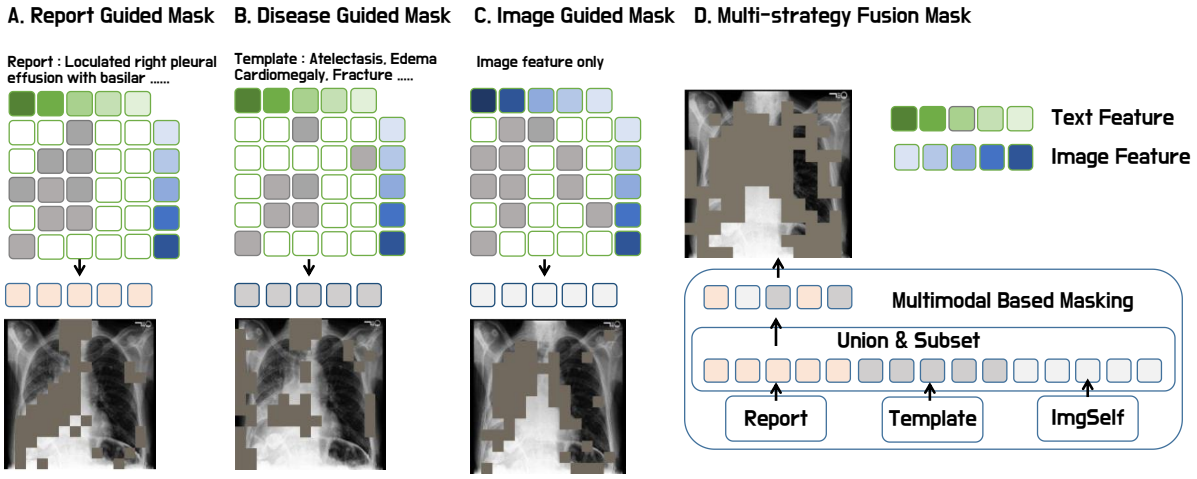


Figure 7: The working mechanism of function M: The features obtained from different text data correspond to image features through the Cross Attention layer, querying the corresponding activated features to identify different regions of interest. These regions of interest are then combined, and a 25% subset is randomly selected to obtain the final Mask result. Unlike MediM, which only uses dot products to activate cross-modal response features, MMCLIP employs cross-attention for deeper feature activation. This allows the model to induce more precise key areas from text features.

hard patch mining (HPM) stands out as a groundbreaking training paradigm for MIM, designed to enhance a model's comprehension of image content (Wang et al., 2023). HPM method first has the model generate a challenging mask and then trains the model to predict the masked patches, similar to traditional methods. This approach enables the model to learn where it is worth applying masks while simultaneously learning how to solve the problems. Also, HPM introduces new metrics to measure the difficulty of the pre-training task by reconstructing the loss, which could help the model to focus more on challenging problems.

Meanwhile, traditional random masking strategies, effective in general domains, fall short in medical contexts where precision is crucial (Chen et al., 2022; Huang et al., 2021; Müller et al., 2022; Zhang et al., 2022; Wu et al., 2023b; Zhou et al., 2022). MedIM (Xie et al., 2023) mitigates the unique challenges posed by healthcare data by improving mask strategies.

However, current methods either rely solely on single-mode features for their attention mechanisms (Kwon et al., 2022) or have limited interaction between multimodal features, such as only employing feature matching (Xie et al., 2023). Our proposed method solves these problems well by employing cross-attention for multimodal feature interaction while performing attention-masked modelling on features that incorporate multimodal information, which allows the model to understand image and text information in a more integrated way.

E Further Analysis

E.1 Different Evaluation Protocols

We further analyze the performance of our method across different evaluation protocols and datasets. As shown in Table 5, our model achieves strong results compared to existing approaches, even though prior methods follow different reporting conventions. For instance, CheXzero reports only ensemble zero-shot performance, while our method provides both single-model and ensemble results. Under the same data setup, our single model already achieves competitive performance, and the ensemble variant further improves the results. In addition, compared with methods such as MedKLIP, which incorporate external medical knowledge during training, our approach relies solely on paired image–report data while still achieving comparable or superior performance, suggesting the effectiveness of the proposed pre-training strategy.

We also observe that the evaluation in Table 1 is based on 14 CheXpert-style labels, some of which are generated by automatic labelers and may contain synthetic noise. To better assess clinical reliability, we further evaluate on ChestX-pert (5 classes), which provides fully human-annotated gold-standard labels. On this benchmark, our method shows more pronounced improvements

Method	ChestXpert (5)	PadChest (All)	PadChest (Unseen 173)	PadChest (Unseen 10)	PadChest (Unseen 20)	Mean
Gloria	53.44	51.12	50.77	50.80	50.24	51.27
MGCA	84.28	66.19	66.04	60.11	60.98	67.52
CheXzero	87.51	66.43	65.76	59.75	63.11	68.51
MedKlip	88.76	65.45	64.16	60.07	53.73	66.43
Ours	**91.42**	**71.54**	**70.00**	**65.09**	**67.14**	**73.04**

Table 5: Zero-shot classification performance across ChestXpert and PadChest benchmarks (%).

Method	CheXpert			Pneumonia			Knee Osteoarthritis		Osteoarthritis	
	1%	10%	100%	1%	10%	100%	10%	100%	10%	100%
MedKlip	64.70	68.04	69.40	74.70	79.43	81.84	30.86	37.92	34.32	35.62
CheXzero	81.18	82.72	83.15	73.94	81.25	82.44	40.82	49.40	34.56	35.39
Ours	**83.65**	**85.43**	**86.29**	**89.01**	**89.38**	**90.34**	**41.32**	**49.94**	**34.91**	**36.57**

Table 6: Performance comparison under different label fractions across multiple datasets (%).

compared to other approaches, indicating its advantage under more reliable supervision. To further examine generalization ability, we conduct zero-shot evaluations on ChestXpert and multiple subsets of the PadChest dataset, including the full label space as well as several unseen-class splits (173 unseen classes, 10 unseen major classes, and 20 unseen major classes). Across all settings, our method consistently achieves the best performance. Notably, the performance gain becomes more significant on unseen categories, which represent a more challenging setting for semantic generalization. These results demonstrate that our method not only performs well on standard benchmarks, but also generalizes effectively to more diverse and previously unseen conditions.

E.2 Cross-Domain Transferability

We further evaluate the transferability of our method beyond chest X-ray (CXR) data. To this end, we conduct linear probing experiments on two musculoskeletal imaging tasks, Knee Osteoarthritis and Osteoarthritis, which differ significantly from CXR in terms of anatomy, visual appearance, and disease semantics. These datasets provide a more challenging setting for assessing cross-domain generalization under supervised protocols.

Despite being pre-trained exclusively on CXR-style data, as shown in Table 6, our method consistently outperforms strong baselines across both datasets and under different label fractions (10% and 100%). This demonstrates that the proposed attention-guided masking strategy improves the robustness and semantic alignment of the learned visual representations, enabling effective transfer beyond thoracic imaging. We also note that the 1% setting is not included for these datasets. To ensure consistency, we adopt a unified batch size across all experiments. However, the number of samples in the 1% split is smaller than a single batch, making training infeasible under this configuration. Therefore, we report results only for the 10% and 100% settings.

Mask Method	Time Cost / Epoch (h)
Baseline	1.80
+ Random Mask	2.20
+ Image Report	2.25
+ Prompt-driven	2.25
+ Image Self	2.30
+ AttMIM	2.40

Table 7: Training time per epoch under different masking strategies.

E.3 Computational Overhead

The additional computation introduced by the proposed masking mechanism arises only during the pre-training stage, where the masking map is generated using three attention sources, including self-attention, cross-attention, and disease-prompt attention. In practice, this overhead is modest. As shown in Table 7, the per-epoch training time increases from 1.80 hours for the baseline to 2.40 hours with AttMIM, representing a relatively small incremental cost compared to the overall pre-training budget. Importantly, the masking mechanism is applied only during pre-training. For all downstream tasks, including zero-shot inference and fine-tuning, the model directly uses the pre-trained visual encoder without introducing any additional computation or latency compared to the baseline. Therefore, the proposed method maintains the same efficiency at inference time while achieving consistent performance improvements. Overall, this design provides a favorable trade-off between training cost and downstream effectiveness.

E.4 Pathology-Aware Reconstruction

We further analyze whether the observed performance gains come from the proposed method itself rather than from additional pre-training data.

Method	Dataset	ChestXpert (5)	PadChest (All)
MedKlip	MIMIC	88.76	65.45
MMCLIP (Single)	MIMIC	89.10	68.49
MMCLIP (Single)	MIMIC + PadChest	89.15	69.43
MMCLIP (Ensemble)	MIMIC + PadChest	**91.42**	**71.54**

Table 8: Comparison of different training data configurations and model variants (%).

To this end, we conduct a controlled comparison in which both MedKLIP and our model are pretrained exclusively on the MIMIC dataset. Under this strictly matched setting, as shown in Table 8, our method already outperforms MedKLIP on both ChestXpert (89.10% vs. 88.76%) and PadChest (68.49% vs. 65.45%), indicating that the improvement primarily stems from the proposed method rather than data scale. When additional PadChest data is incorporated, the performance is further improved (ChestXpert: +0.05; PadChest: +0.94), suggesting that although extra data provides complementary benefits, the main contribution lies in the method itself.

We also examine whether the performance improvements are associated with meaningful model behavior rather than unrelated factors. Quantitatively, as shown in Table 3, different masking strategies consistently improve over the CLIP baseline. While random masking yields only marginal gains, pathology-aware strategies such as Image-Report, Prompt-driven, and Image-Self lead to larger and more stable improvements. The proposed AttMIM achieves the best performance (AUC 74.96, F1 35.22), indicating that the gains are well aligned with the design of the masking mechanism. Qualitatively, Figure 3 provides direct visual evidence supporting this observation. The visualization shows that different masking strategies focus on distinct yet clinically meaningful regions, including dominant lesions, additional abnormalities, and structurally challenging boundaries. The final combined strategy integrates these complementary behaviors, resulting in a more comprehensive and balanced focus on pathological regions.

Overall, both the quantitative improvements and the qualitative visualizations consistently demonstrate that the proposed masking mechanism effectively guides the model to focus on clinically relevant regions during reconstruction, thereby leading to robust and interpretable performance gains.

F Limitations

Although MMCLIP shows strong performance across multiple tasks, some limitations remain. The model is primarily trained on chest X-ray data, so its applicability to other imaging modalities and anatomical regions still needs further validation. In addition, its reliance on predefined disease prompts and entity recognition tools may limit robustness when handling rare conditions or diverse reporting styles. The attention-guided masking strategy, while beneficial for feature learning, also increases computational cost, which may restrict use in resource-limited settings. Finally, as current experiments are mainly conducted on public datasets, broader evaluation in real-world clinical environments is necessary to confirm its reliability.