RESEARCH PRESENTATION · MEDICAL AI

MMCLIP

Cross-modal Attention Masked Modelling for Medical Language-Image Pre-Training

A disease-aware VLP framework using attention-guided masking for improved cross-modal medical representation learning.

AttMIM · EntMLM · Zero-Shot SOTA · Unpaired Data

Biao Wu · Yutong Xie · Zeyu Zhang · Vu Minh Hieu Phan · Qi Chen · Ling Chen · Qi Wu

University of Technology Sydney · Mohamed bin Zayed University of AI
The Australian National University · University of Adelaide

ARXIV:2407.19546V4 [CS.CV] · APRIL 2025



BACKGROUND & MOTIVATION

Limitations of Random Masking in Medical VLP

Suboptimal Image Masking (MIM)

Problem: Missing Subtle Lesions

- Medical findings (e.g., pneumonia, effusion) are **subtle and localized**, unlike large objects in natural images.
- Random masking often hides normal lung areas or irrelevant background, **ignoring actual lesion regions**.

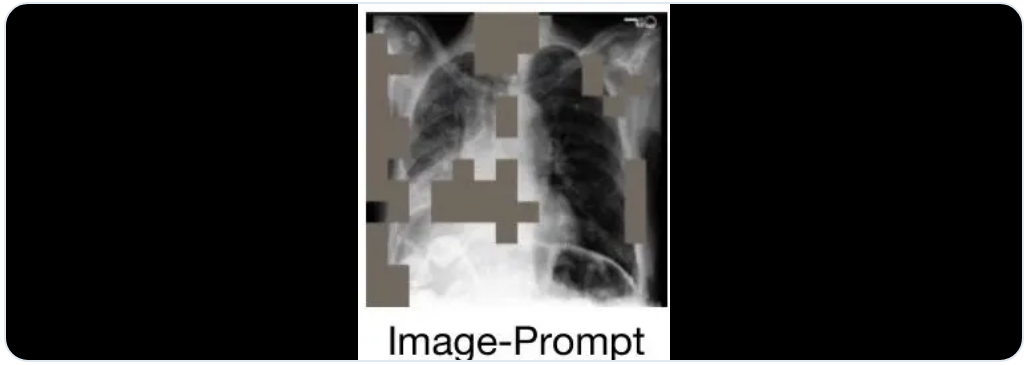


Figure 1: Traditional random masking in Chest X-ray data

Inefficient Language Masking (MLM)

Problem: Masking Common Words

- Standard MLM masks frequent tokens like **"with", "right", "the"** which lack diagnostic value.
- Critical **medical entities** (e.g., "cardiomegaly", "pneumothorax") are rarely masked, limiting semantic learning.
- Models struggle to reconstruct key pathological features due to the **scarcity of entity-focused training**.
- Medical text is highly structured; reconstruction must prioritize **entity accuracy** over textual fluency.

> *"Result: Models fail to capture the deep visual-textual links necessary for high-accuracy clinical diagnosis."*

CONTEXT & MOTIVATION

Limitations of Current Medical VLP

Random Masking Suboptimality

MIM: Visually Subtle Pathologies

Random patch masking often hides normal lung areas while missing subtle, localized lesions like pleural effusion or pneumonia.

MLM: Non-Diagnostic Focus

Masks common function words ("with", "right") instead of medically significant entities critical for diagnosis.

⚠ **Generic masking fails to force the model to learn fine-grained pathological features.**

Data Utilization Barriers

Strict Pair Dependency

Traditional CLIP frameworks rely heavily on image-report pairs, rendering unpaired data unusable for pre-training.

Real-World Data Scarcity

High-quality images (e.g., PadChest) often lack English reports or have noisy/missing text, leading to data waste.

⚠ **Current models cannot exploit large-scale image-only datasets for representation learning.**

METHODOLOGY

MMCLIP Framework Overview

Image-Text Contrastive Learning

Aligns chest X-rays with clinical reports in a shared representation space.

AttMIM (Masked Image Modeling)

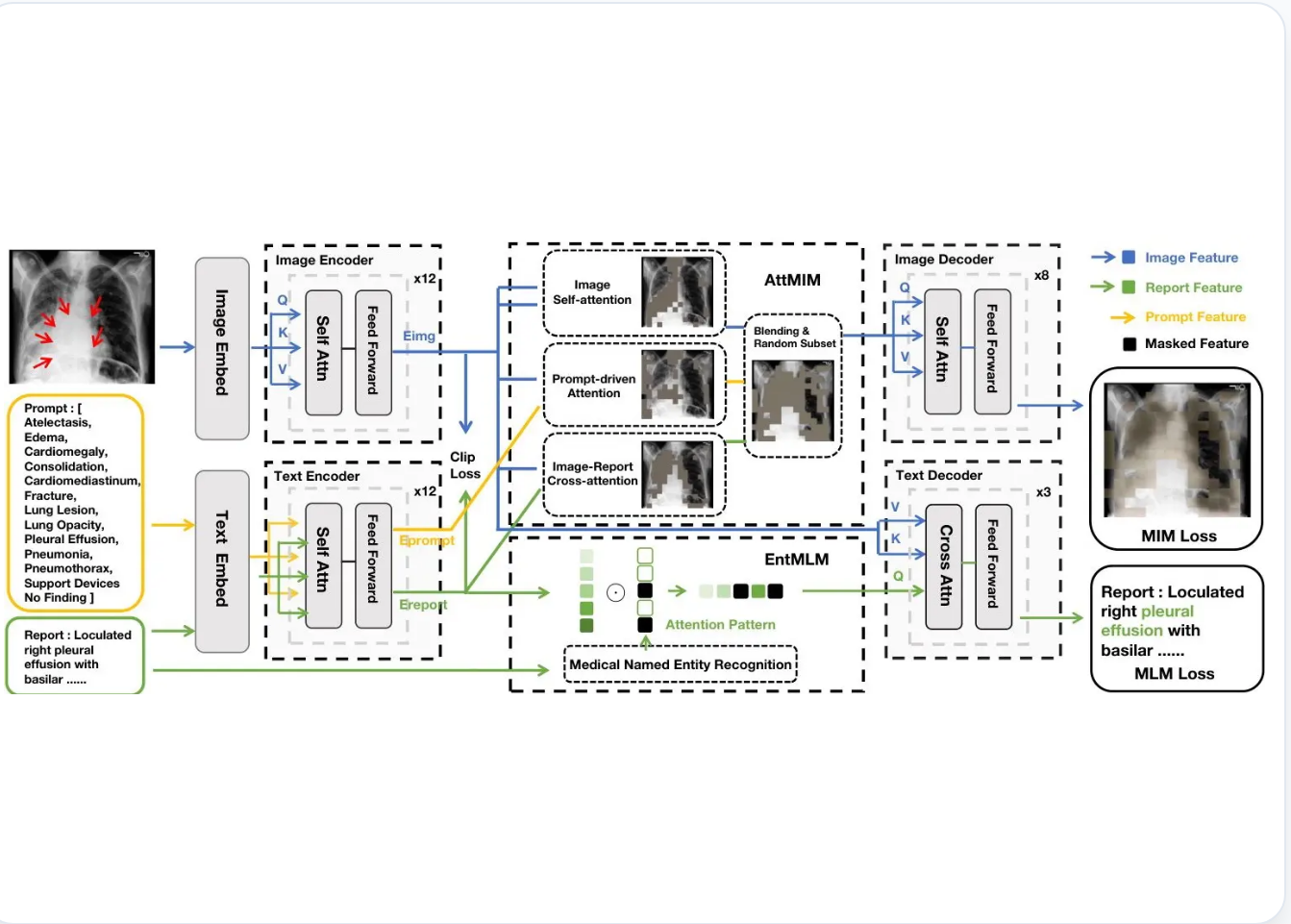
Disease-aware masking guided by cross-modal and self-attention.

EntMLM (Masked Language Modeling)

Entity-driven reconstruction of clinical terms with visual guidance.

CORE COMPONENTS

- **Encoders:** ViT-B/16 (Image) & 12-layer Transformer (Text)
- **Decoders:** 8-layer (Image) & 3-layer (Text) with cross-attention
- **Disease Prompts:** 14 common findings for global guidance



MMCLIP Architecture: Unified framework for multi-objective pre-training on paired and unpaired data.

Cross-modal Attention Masking (AttMIM)

DISEASE-AWARE GUIDANCE

Image-Report Cross-Attention

Radiology reports guide the model to identify image regions relevant to specific diagnostic descriptions.

$$A_r = softmax(E_{img} \cdot E_{report}{}^T / \sqrt{d_k}) \cdot E_{report}$$

Prompt-Driven Attention

14 common chest X-ray findings provide global guidance, enabling support for unpaired image-only data.

$$A_p = softmax(E_{img} \cdot E_{prompt}{}^T / \sqrt{d_k}) \cdot E_{prompt}$$

Image Self-Attention

Identifies visually informative or challenging regions inherently from the image without external guidance.

$$A_i = softmax(E_{img} \cdot E_{img}{}^T / \sqrt{d_k}) \cdot E_{img}$$

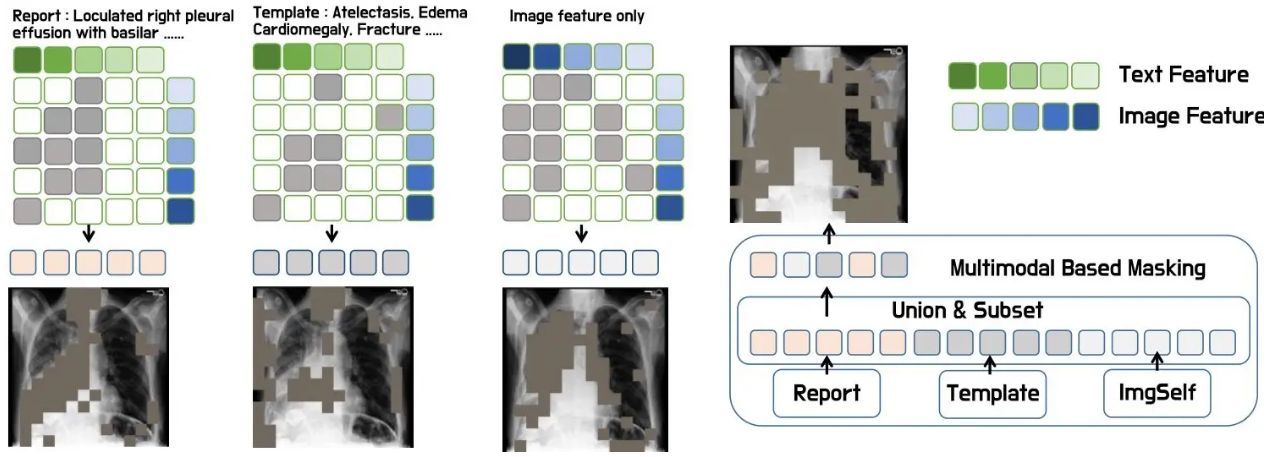
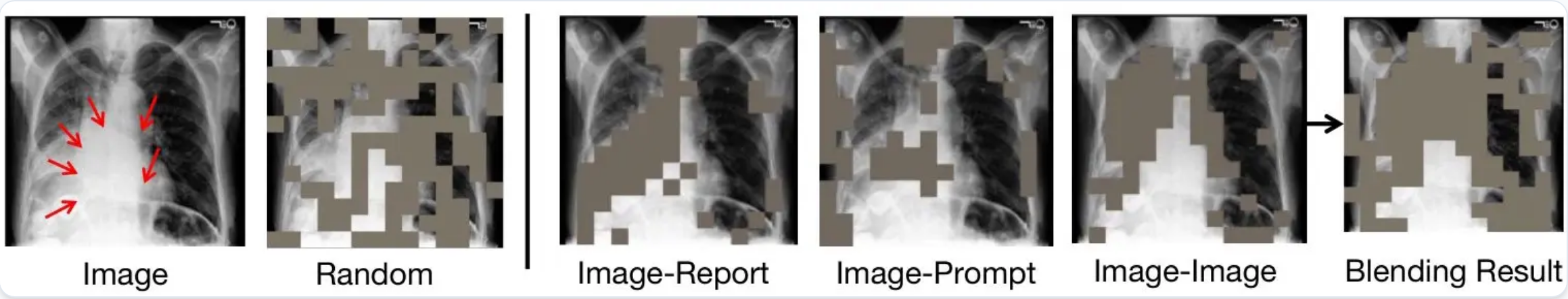


Figure 2: Multimodal masking process combining Report, Prompt (Template), and Image features to generate a comprehensive disease-aware mask.

Mask Blending and Selection Strategy



ORIGINAL SCAN | RANDOM MASK | IMAGE-REPORT | IMAGE-PROMPT | IMAGE-SELF | FINAL BLENDING

Multi-Attention Fusion

$$M_b = M(A_r; \lambda_1) \cup M(A_p; \lambda_1) \cup M(A_i; \lambda_1)$$

- **M(A, λ):** Selects top-λ activated features as mask.
- **Union (U):** Fuses report, prompt, and image guidance.
- Ensures comprehensive coverage of pathological regions.

Final Sampling & Params

$$M_{AttMIM} = \text{RandomSubSet}(M_b, \lambda_2)$$

Downsamples the blended union to a fixed mask ratio.

$\lambda_1 = 0.70$ (Strategy) **$\lambda_2 = 0.75$ (Final)**

Maintains high reconstruction difficulty while targeting specific disease-related features.

Support for Unpaired Image-only Data

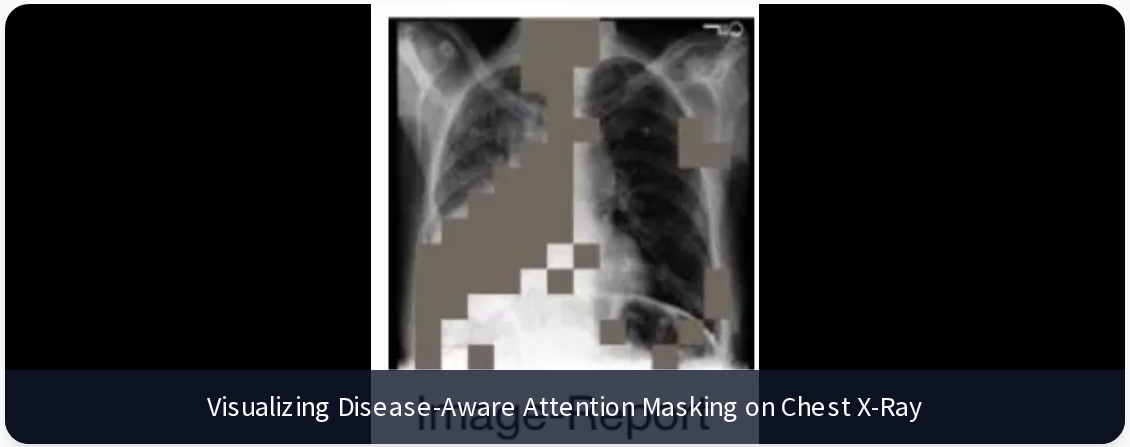
Prompt-Driven Attention

- Utilizes 14 pre-defined clinical findings (e.g., Atelectasis, Effusion) to provide global guidance.
- Replaces report-driven cross-attention when paired textual data is unavailable.
- Enables pre-training on large-scale datasets with non-English reports (e.g., PadChest's Spanish reports).

A_p = Prompt Attention A_i = Self-Attention

Unpaired Data Objective

- **Loss Calculation:** For image-only data, the model optimizes only the Masked Image Modeling loss.
- **Formula:** $Loss_{unpair} = Loss_{MIM}$
- Allows Spanish-report data (160k images) to contribute to robust feature extraction.



Visualizing Disease-Aware Attention Masking on Chest X-Ray

ABLATION: IMPACT OF ATTENTION STRATEGIES

Attention Strategy	AUC (%)	F1 Score
CLIP Baseline	72.43	34.24
Prompt-Driven (A_p)	74.91	34.59
AttMIM (Fused Strategy)	**74.96**	**35.22**

METHODOLOGY > ENTMLM

Entity-Driven Masked Language Modeling (EntMLM)

Medical Rationale

Medical reports prioritize **accuracy of clinical terms** over general textual fluency. Random masking often hides redundant function words rather than informative pathology.

- Avoids masking words like "with", "the", "right".
- Focuses on high-value diagnostic entities.

NER-Guided Logic

Uses **ScispaCy** (biomedical NER) to identify critical entities within radiology reports for preferential masking.

- Identifies: Diseases, Drugs, Symptoms.
- Targets: Treatments and Medical terms.

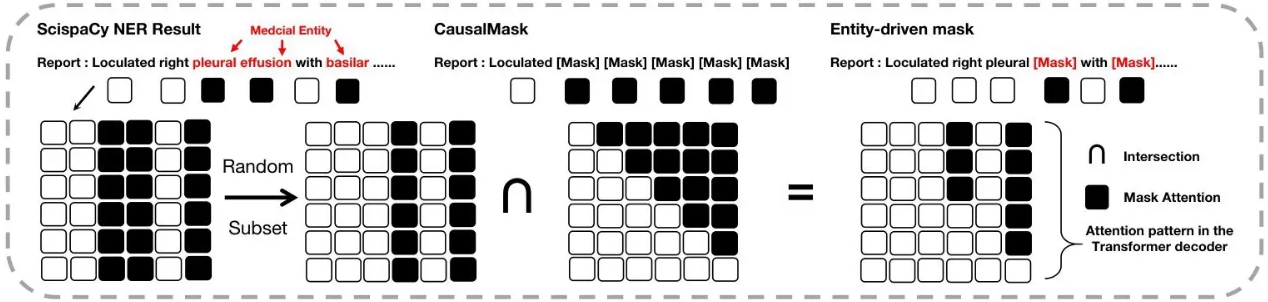
Tool: ScispaCy NER

Masking Strategy

Integrates entity selection with autoregressive constraints to ensure robust cross-modal reconstruction.

- **Ratio:** 0.2 of identified entity words.
- **Constraint:** Intersection with CausalMask.
- **Decoding:** Visual-guided reconstruction.

VISUALIZING ENTMLM: NER INTERSECTION & CAUSAL MASKING



CONTRIBUTION 3 > ENTMLM

Visual-Guided Entity Reconstruction

Cross-Modal Decoder Mechanism

Unlike standard language models that rely solely on textual context, the MMCLIP text decoder employs a **Visual-Guided Reconstruction** strategy to bridge the semantic gap.

- **Cross-Attention Integration:** The 3-layer transformer decoder queries high-level visual features from the Image Encoder.
- **Contextual Enrichment:** Masked medical entities (e.g., "pleural effusion") are reconstructed using both preceding tokens and spatially relevant image patches.
- **Autoregressive Discipline:** Causal masking ensures predictions are made based on available visual cues and previous textual information.

Strengthening Pathology-Language Links

This forced interaction compels the model to map specific medical terminology directly to visual pathological manifestations in the chest X-ray.

PRECISE ALIGNMENT | SEMANTIC GROUNDING | DISEASE-AWARENESS

Reconstruction Pipeline

1. Identify Medical Entity (ScispaCy)
2. Apply Selective Entity Masking
3. Extract Visual Pathology Cues
4. Reconstruct via Cross-Attention

Core Value Added

- Ensures **lexical accuracy** for clinical findings rather than general fluency.
- Mitigates noise from irrelevant background text in radiology reports.
- Boosts **Zero-Shot performance** by refining fine-grained image-word mapping.

Training Objectives and Strategy

Two-Phase Training

PHASE 1: WARM-UP

Execute masked modeling objectives for **15,000 iterations**.

PHASE 2: JOINT TRAINING

Jointly optimize all objectives end-to-end to ensure balanced representation learning.

> ***Note:*** *Warm-up addresses inconsistent convergence rates between contrastive alignment and reconstruction tasks.*

PAIRED IMAGE-REPORT DATA

$$\text{Loss}_{\text{pair}} = L_{\text{align}} + L_{\text{MIM}} + L_{\text{MLM}}$$

- Contrastive Alignment
- AttMIM (3 Attention Types)
- EntMLM (Visual-Guided)

UNPAIRED IMAGE-ONLY DATA

$$\text{Loss}_{\text{unpair}} = L_{\text{MIM}}$$

Enables leveraging large-scale datasets like PadChest without paired reports.

- Prompt-Driven Attention
- Image Self-Attention

Final Training Objective: $L = \text{Loss}_{\text{pair}} + \text{Loss}_{\text{unpair}}$

Experimental Results: Zero-Shot Classification

MMCLIP demonstrates superior zero-shot performance across three major chest X-ray datasets, significantly outperforming medical CLIP baselines.

MODEL ARCHITECTURE	CHEXPERT AUC	PADCHEST-SEEN AUC	PNEUMONIA AUC
ConVIRT	52.10	74.31	79.21
GLORIA	54.84	74.56	70.37
BioViL	60.01	71.17	84.12
BioViL-T	70.93	76.17	86.03
CheXzero	87.90	76.75	85.13
MedKLIP	89.97	84.97	86.94
MMCLIP (Ours)	90.53	85.79	86.16
MMCLIP + Ensemble SOTA	▲ 91.45	▲ 86.31	▲ 86.95

Unseen Class Generalization

MODEL	UNSEEN-10 AUC	UNSEEN-20 AUC	UNSEEN-173 AUC
CheXzero	59.75	63.11	65.76
MedKLIP	60.07	53.73	64.16
MMCLIP SOTA	**65.09**	**67.14**	**70.00**

Source: Evaluation on PadChest unseen labels (Section 7, Table 2).

Superior Generalization

MMCLIP achieves a significant performance lift of **4.24% – 5.84%** in AUC across all unseen class sets compared to the best previous SOTA models.

Rare Finding Robustness

By moving beyond random masking, the model learns more robust pathological representations that generalize to rare medical findings not explicitly seen during training.

EVALUATION > DOWNSTREAM PERFORMANCE

Fine-Tuning and Downstream Performance

Data-Efficient Representation Learning

1% Labeled Data

MMCLIP exhibits superior few-shot capability, significantly outperforming baselines in extremely low-resource settings.

10% Labeled Data

Maintains a consistent lead over SOTA models, demonstrating rapid convergence and robust feature extraction.

100% Labeled Data

Achieves peak performance benchmarks on diagnostic classification, validating the quality of pre-trained weights.

KEY ADVANTAGE

The PadChest Boost

Incorporating 160k+ unpaired images from PadChest provides a **measurable performance lift** across all labeling proportions.

PERFORMANCE GAINS ON:

- **CheXpert Dataset**
- **CovidX Dataset**
- **Multi-label Chest X-ray Tasks**

MMCLIP transforms silent image-only data into active training signals for downstream tasks.

Ablation Study Findings

AttMIM & Training Strategy

METHOD CONFIGURATION	AUC (%)	F1 (%)
CLIP Baseline (No Warm-up)	72.43	34.24
+ MIM Random (No Warm-up)	72.85	34.49
+ MIM Random (With Warm-up)	73.98	34.80
+ AttMIM (Fused Strategy)	**74.96**	**35.22**

Key Insight: Warm-up iterations (+1.13% AUC) solve convergence gaps, while AttMIM blending provides superior coverage of pathological regions.

EntMLM Parameter Sensitivity

METHOD + MASK RATIO	AUC (%)	F1 (%)
AttMIM Only (No EntMLM)	74.96	35.22
+ EntMLM (Ratio 0.5)	74.98	34.49
+ EntMLM (Ratio 0.3)	75.44	34.88
+ EntMLM (Ratio 0.2)	**75.78**	**35.02**

Key Insight: Focusing on medical entities via NER-guided masking at a **0.2 ratio** maximizes the cross-modal alignment between visual pathology and clinical language.

CONCLUSION

Summary of Contributions

Disease-Aware Masking (AttMIM)

Replaces random patch hiding with cross-modal attention guidance from multiple sources.

- Image-Report Cross-Attention
- Disease-Prompt Guidance
- Image Self-Attention Fusing

Unpaired Data Compatibility

Enables the effective utilization of massive image-only datasets without paired reports.

- Prompt-Driven Attention
- Leverages PadChest (160k images)
- Scalable Representation Learning

Entity-Focused Language Model

Prioritizes medical clinical terminology over generic text reconstruction via EntMLM.

- NER-Guided Masking Selection
- Visual-Guided Reconstruction
- Stronger Pathology-Text Linking

Core Value Achievement

MMCLIP transforms generic VLP into a **disease-aware framework**, achieving SOTA on zero-shot, unseen-class recognition, and fine-tuning across 5 major medical datasets.