Medical Vision–Language Pretraining with LLM-Guided Temporal Supervision

Liang Bai[1], **Zhi Wang**[1], **Huimin Yan**[1], **Xian Yang**[2*]

[1] Institute of Intelligent Information Processing, Shanxi University, Taiyuan, 030006, China
[2] Alliance Manchester Business School, The University of Manchester, Manchester, M13 9PL, UK
bailiang@sxu.edu.cn, 202322407045@email.sxu.edu.cn, yanhm0925@163.com, xian.yang@manchester.ac.uk

Abstract

Medical vision–language pretraining typically relies on static image–text pairs, overlooking temporal cues vital for understanding clinical progression. This limits model sensitivity to evolving semantics and reduces their effectiveness in real-world clinical reasoning. To address this challenge, we propose TAMM—a temporal alignment framework that leverages weak but semantically rich supervision from large language models (LLMs). Given temporally adjacent clinical reports, LLMs automatically generate (i) coarse-grained trend labels (e.g., improving or worsening), and (ii) fine-grained rationales explaining the supporting clinical evidence. These complementary signals inject temporal semantics without requiring manual annotation, and guide vision–language representation learning to capture trend-sensitive cross-modal alignment and rationale-grounded coherence. Experiments on multiple medical benchmarks demonstrate that TAMM improves retrieval and classification performance while yielding more interpretable, temporally consistent embeddings. Our results highlight the potential of leveraging LLM-derived supervision to equip vision–language models with temporal awareness critical for clinical applications.

Introduction

Pretraining medical vision–language models has become a powerful paradigm for learning generalizable cross-modal representations by aligning radiological images with associated free-text reports (Huang et al. 2021; Wang et al. 2022a; Zhou et al. 2022; Wan et al. 2023; Zhou et al. 2023). These methods typically operate on static image–text pairs drawn from individual patient visits, using contrastive objectives (Cheng et al. 2023; Zhou et al. 2023) or masked modeling (Moon et al. 2022; Zhang et al. 2025) to align visual and textual semantics. While effective for capturing cross-modal correspondences at a single timepoint, such formulations largely ignore the longitudinal nature of real-world patient data.

In actual clinical practice, diagnosis and treatment decisions are rarely made based on isolated observations. Physicians routinely examine longitudinal records—comparing past and current imaging studies alongside associated reports—to track disease progression or recovery. This process of temporal reasoning is critical for interpreting ambiguous findings, planning interventions, and evaluating response to therapy. For instance, determining whether a pulmonary opacity is resolving or worsening often hinges not on the current scan alone, but on subtle linguistic trends across time such as "slightly decreased" or "newly emerged." Although large-scale datasets (e.g., (Johnson et al. 2019)) contain sequences of temporally ordered image–report pairs, most vision–language pretraining methods ignore this structure, treating each instance in isolation. Some recent works (Bannur et al. 2023; Yang et al. 2024) attempt to incorporate temporal metadata or time-aware sampling, yet they often lack direct semantic supervision for modeling clinical evolution. Designing learning objectives that faithfully reflect medical trend dynamics remains a key challenge, largely due to the scarcity of annotated temporal labels.

To address this gap, we propose **TAMM**, a **T**emporal **A**lignment framework for **M**edical vision–language **M**odeling that introduces LLM-guided temporal supervision. Given temporally adjacent reports, a large language model (LLM) automatically derives two complementary signals: (i) *coarse-grained trend labels* indicating whether the patient's condition improved, worsened, or remained stable, and (ii) *fine-grained rationales* that provide textual evidence supporting the trend. These semantically rich annotations inject clinically meaningful temporal semantics into pretraining—without requiring human labeling. TAMM leverages these signals to jointly model temporal alignment and cross-modal semantic coherence. It integrates a contrastive objective to align image–text representations across adjacent timepoints, capturing directional progression, and a rationale-guided objective that explicitly grounds visual features to clinically salient textual changes. This dual supervision enables the model to detect fine-grained temporal semantics, improving robustness to representation drift and enhancing generalization on temporally structured tasks.

We evaluate TAMM on multiple medical benchmarks and show that it consistently improves retrieval and classification performance while producing more interpretable and temporally consistent representations. Our contributions are:

- We introduce LLM-guided temporal supervision that enables vision–language models to leverage longitudinal data without manual annotations.

*Xian Yang is the corresponding author.

- We propose TAMM, a unified framework that aligns multimodal representations across time through trend-aware contrastive learning and rationale-guided grounding.
- We demonstrate that TAMM enhances temporal sensitivity, coherence, and interpretability in downstream medical vision–language tasks.

Related Work

Temporal Modeling in Multimodal Medical Data. Prior work on medical vision-language pretraining (VLP) has primarily focused on *static* cross-modal alignment, typically aligning single image-report pairs using objectives such as contrastive learning or masked token reconstruction (Boecking et al. 2022; Huang et al. 2024; Zhang et al. 2023b; Zhou et al. 2022; Zhang et al. 2023a; Li et al. 2025). For instance, MedKLIP (Wu et al. 2023) and MedCLIP (Wang et al. 2022b) incorporate entity-level and descriptive cues, while MGCA (Wang et al. 2022a) captures intra-modal semantic consistency. MRM (Zhou et al. 2023) further improves cross-modal understanding by jointly reconstructing masked image patches and report tokens. However, all these methods treat each image-report pair independently and overlook the temporal evolution of disease states across longitudinal clinical visits. Recent studies have begun to explore temporal modeling in radiology (Bannur et al. 2023; Yang et al. 2024; Moon et al. 2022). For instance, ALTA (Moon et al. 2022) uses prior-view X-rays for temporal context but is limited to unimodal setups and lacks an explicit temporal objective. Med-ST (Yang et al. 2024) introduces cross-modal cycle consistency from historical image-text pairs but focuses mainly on bidirectional mapping, without explicitly modeling semantic progression. Our method addresses these limitations by reorganizing longitudinal data into temporally ordered sequences and leveraging LLM-generated trend signals to supervise representation changes over time.

LLMs and Reasoning-Guided Supervision. LLMs have shown strong capabilities in biomedical QA, report summarization, and zero-shot clinical reasoning (Singhal et al. 2022; Wang, Zhao, and Petzold 2023; Shaib et al. 2023; Yunxiang et al. 2023; Moor et al. 2023). Their use as supervisory signals in medical AI is emerging, particularly in weakly supervised settings (Phan et al. 2024; Wang et al. 2024). Prior studies have explored explanation-driven training with rationales (Wang et al. 2023; Gai et al. 2024) or enforcing consistency via bidirectional reasoning (Sultan et al. 2025; Chen et al. 2023). However, few have extended these ideas to longitudinal modeling, where temporal supervision is scarce and disease progression subtle. Our work introduces a novel use of LLMs to generate complementary temporal supervision signals.

Method

Method Overview

We consider a medical dataset $\mathcal{D}$ consisting of patient-specific sequences of diagnostic visits, where each visit contains an image–text pair. Formally, $\mathcal{D} = \{\mathcal{S}_1, \ldots, \mathcal{S}_{|\mathcal{D}|}\}$, where each $\mathcal{S}_i = \{(\mathbf{I}_{i,t}, \mathbf{T}_{i,t})\}_{t=1}^{|\mathcal{S}_i|}$ is a temporally ordered sequence of studies for patient i. Here, $\mathbf{I}_{i,t}$ and $\mathbf{T}_{i,t}$ denote the image and report at time t. Our objective is to pretrain a vision-language model that leverages such longitudinal data to learn representations sensitive to semantic trends—such as clinical improvement or deterioration—across time. To this end, we propose **TAMM**, a pretraining framework that leverages LLMs to provide informative temporal supervision, as shown in Figure 1. TAMM extracts two types of signals from temporally adjacent reports: coarse-grained trend labels (e.g., improved, stable, worsened) and fine-grained rationales describing the supporting clinical evidence. These signals are integrated into three key components. First, *LLM-Guided Temporal Signal Extraction* derives trend and rationale annotations from neighboring reports. Next, *Trend-Aware Representation Alignment* aligns cross-time representation shifts with the extracted trend labels. Finally, *Rationale-Conditioned Temporal Representation Learning* uses rationales to guide semantic alignment across time, grounding visual features to clinically meaningful textual changes.

LLM-Guided Temporal Signal Extraction

Understanding disease progression from longitudinal data requires identifying subtle and clinically meaningful changes between visits. However, manual annotation of such transitions is costly, inconsistent, and infeasible at scale. To address this, we leverage the zero-shot reasoning capabilities of LLMs to extract semantically meaningful temporal signals from report pairs.

Given each temporally adjacent report pair $(\mathbf{T}_{i,t}, \mathbf{T}_{i,t+1})$, we prompt an LLM to generate two outputs: (1) a discrete trend label $y_{i,t} \in \{-1, 0, 1\}$, representing worsening, stable, or improving in patient's condition; and (2) a natural language rationale $s_{i,t}^{\text{reason}}$ that highlights clinical observations supporting the assigned label (e.g., "increased effusion" or "reduction in lung opacity"). To enhance reliability, we apply a bidirectional prompting strategy: the LLM is also queried on the reversed report order $(\mathbf{T}_{i,t+1}, \mathbf{T}_{i,t})$ to produce a counter-trend label $y_{i,t}^{\text{counter}}$ and rationale $s_{i,t}^{\text{counter}}$. We retain only those samples satisfying a directional consistency check: $y_{i,t}^{\text{counter}} = -y_{i,t}$. This filters noisy or ambiguous responses, ensuring the extracted signals are semantically and temporally coherent. The resulting trend labels and rationales serve as pseudo-supervision for the subsequent modules in TAMM. The trend labels guide the alignment of representation shifts across time, while the rationales inform bidirectional text-guided visual representation.

Trend-Aware Representation Alignment

To model clinically meaningful temporal progression, we propose a trend-aware representation alignment module that aligns changes in multimodal representations with pseudo trend labels derived from LLMs. For each patient sequence $\mathcal{S}_i = \{(\mathbf{I}_{i,t}, \mathbf{T}_{i,t})\}_{t=1}^{|\mathcal{S}_i|}$, we use an image encoder f_{I} (ViT-B/16 (Dosovitskiy et al. 2020)) and a text encoder f_{T} (Bio-ClinicalBERT (Alsentzer et al. 2019)) to extract both global and fine-grained representations. Specifically,

$$[\mathbf{x}_{i,t}^{g}, \mathbf{X}_{i,t}] = f_{\text{I}}(\mathbf{I}_{i,t}), \quad [\mathbf{r}_{i,t}^{g}, \mathbf{R}_{i,t}] = f_{\text{T}}(\mathbf{T}_{i,t}), \quad (1)$$

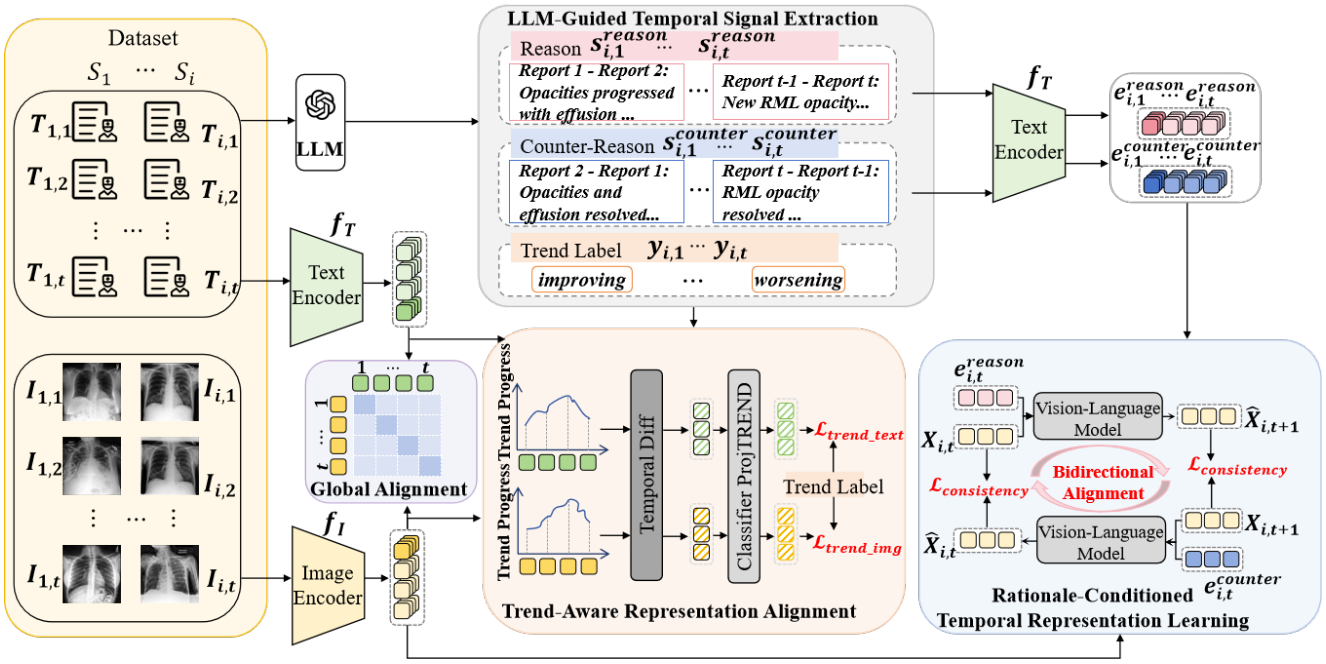

Figure 1: Overview of TAMM. Given temporally adjacent image–report pairs, TAMM uses an LLM to extract discrete trend labels and fine-grained rationales. These signals supervise two key modules: a trend-aware alignment module that aligns representation shifts with predicted trends, and a rationale-conditioned module that enforces bidirectional temporal consistency across modalities. A standard image–text alignment module is jointly trained to learn cross-modal embeddings.

where $\mathbf{x}^g_{i,t}, \mathbf{r}^g_{i,t} \in \mathbb{R}^d$ denote global feature vectors, and $\mathbf{X}_{i,t} \in \mathbb{R}^{M\times d}, \mathbf{R}_{i,t} \in \mathbb{R}^{W\times d}$ are patch-level and token-level feature matrices.

To capture temporal dynamics, we compute first-order differences between adjacent timepoints:

$$\Delta\mathbf{x}_{i,t} = \mathbf{x}^g_{i,t+1} - \mathbf{x}^g_{i,t}, \quad \Delta\mathbf{r}_{i,t} = \mathbf{r}^g_{i,t+1} - \mathbf{r}^g_{i,t}. \tag{2}$$

These difference vectors are passed through a shared trend classifier $\mathrm{Proj}_{\mathrm{TREND}}$ to produce logits over three categories:

$$\hat{\mathbf{y}}^{(x)}_{i,t} = \mathrm{Proj}_{\mathrm{TREND}}(\Delta\mathbf{x}_{i,t}), \quad \hat{\mathbf{y}}^{(r)}_{i,t} = \mathrm{Proj}_{\mathrm{TREND}}(\Delta\mathbf{r}_{i,t}), \tag{3}$$

where $\hat{\mathbf{y}}^{(x)}_{i,t}, \hat{\mathbf{y}}^{(r)}_{i,t} \in \mathbb{R}^3$ represent predicted logits for *worsening*, *stable*, and *improving*.

Pseudo labels $y_{i,t} \in \{-1, 0, 1\}$ are generated via LLM-based bidirectional prompting and are mapped to class indices $\{0, 1, 2\}$. We define the classification losses for the image and text modalities as:

$$\mathcal{L}_{\text{trend\_img}} = \frac{1}{|\mathcal{S}_i|-1}\sum_{t=1}^{|\mathcal{S}_i|-1} -\log\left(\frac{\exp(\hat{\mathbf{y}}^{(x)[y_{i,t}+1]}_{i,t})}{\sum_{j=0}^{2}\exp(\hat{\mathbf{y}}^{(x)[j]}_{i,t})}\right), \tag{4}$$

$$\mathcal{L}_{\text{trend\_text}} = \frac{1}{|\mathcal{S}_i|-1}\sum_{t=1}^{|\mathcal{S}_i|-1} -\log\left(\frac{\exp(\hat{\mathbf{y}}^{(r)[y_{i,t}+1]}_{i,t})}{\sum_{j=0}^{2}\exp(\hat{\mathbf{y}}^{(r)[j]}_{i,t})}\right). \tag{5}$$

The loss averages both modalities:

$$\mathcal{L}_{\text{trend}} = \frac{1}{2}\left(\mathcal{L}_{\text{trend\_img}} + \mathcal{L}_{\text{trend\_text}}\right). \tag{6}$$

This trend-aware alignment objective guides the model to encode temporally sensitive features aligned with clinically meaningful changes, supporting better longitudinal understanding and downstream prediction.

Rationale-Conditioned Temporal Representation Learning

We introduce a rationale-conditioned temporal representation learning module that encourages the model to predict visual features at adjacent timepoints, guided by natural language rationales generated by LLMs. For each report pair $(\mathbf{T}_{i,t}, \mathbf{T}_{i,t+1})$, the LLM provides two explanations: a forward rationale $s^{\text{reason}}_{i,t}$ describing the change from t to $t{+}1$, and a backward (counterfactual) rationale $s^{\text{counter}}_{i,t}$ describing the inverse transition. We encode these using the same text encoder f_{T}:

$$\mathbf{e}^{\text{reason}}_{i,t} = f_{\mathrm{T}}(s^{\text{reason}}_{i,t}), \quad \mathbf{e}^{\text{counter}}_{i,t} = f_{\mathrm{T}}(s^{\text{counter}}_{i,t}), \tag{7}$$

where $\mathbf{e}^{\text{reason}}_{i,t}, \mathbf{e}^{\text{counter}}_{i,t} \in \mathbb{R}^d$ are semantic embeddings of the corresponding rationales. We use these embeddings to guide the prediction of visual patch features across time. Given fine-grained image representations $\mathbf{X}_{i,t} \in \mathbb{R}^{M\times d}$, we predict features at the adjacent timepoints using a vision-language fusion model $f_{\mathrm{VLM}}(\cdot,\cdot)$ (Yan et al. 2025):

$$\hat{\mathbf{X}}_{i,t+1} = f_{\mathrm{VLM}}(\mathbf{X}_{i,t}, \mathbf{e}^{\text{reason}}_{i,t}), \quad \hat{\mathbf{X}}_{i,t} = f_{\mathrm{VLM}}(\mathbf{X}_{i,t+1}, \mathbf{e}^{\text{counter}}_{i,t}), \tag{8}$$

where $\hat{\mathbf{X}}_{i,t+1}, \hat{\mathbf{X}}_{i,t} \in \mathbb{R}^{M \times d}$ are predicted features conditioned on the respective rationales.

To enforce semantic fidelity, we minimize a bidirectional alignment loss between the predicted and original visual features:

$$\mathcal{L}_{\text{consistency}} = \frac{1}{|\mathcal{S}_i| - 1} \sum_{t=1}^{|\mathcal{S}_i|-1} \Big(\|\hat{\mathbf{X}}_{i,t+1} - \mathbf{X}_{i,t+1}\|_2^2 + \|\hat{\mathbf{X}}_{i,t} - \mathbf{X}_{i,t}\|_2^2 \Big) \quad (9)$$

This module promotes temporally coherent visual features grounded in reasoning, with a bidirectional design capturing dynamic disease changes for improved interpretability and clinical alignment.

Overall Objective

TAMM is trained with a unified objective that jointly captures temporal dynamics and cross-modal alignment. The temporal loss combines coarse-grained trend supervision with fine-grained reasoning consistency:

$$\mathcal{L}_{\text{temporal}} = \mathcal{L}_{\text{trend}} + \lambda_{\text{cons}} \mathcal{L}_{\text{consistency}}, \quad (10)$$

where λ_{cons} controls the trade-off between the two temporal signals.

To encourage semantic alignment between images and reports at each visit, we adopt a CLIP-style symmetric InfoNCE loss (Radford et al. 2021) over global embeddings. Let $\ell_{i,t}^{\text{InfoNCE}}$ denote the bidirectional contrastive loss for each image-report pair $(\mathbf{I}_{i,t}, \mathbf{T}_{i,t})$, as defined in (Radford et al. 2021). The global alignment loss is computed by averaging over visits within each patient:

$$\mathcal{L}_{\text{global}} = \sum_i \frac{1}{|\mathcal{S}_i|} \sum_t \ell_{i,t}^{\text{InfoNCE}}. \quad (11)$$

The overall training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{global}} + \mathcal{L}_{\text{temporal}}. \quad (12)$$

This joint formulation enables TAMM to learn image-text representations that are both semantically aligned and temporally coherent with underlying clinical trajectories.

Experiments

Experimental Setting

Pretraining Data. We pretrain TAMM on the MIMIC-CXR dataset (Johnson et al. 2019), which includes chest radiographs and paired free-text reports. To support temporal modeling, we reorganize the data into patient-level sequences sorted by study time. Each sequence contains up to four consecutive image–report pairs (Yang et al. 2024). Shorter sequences are retained. This restructuring yields approximately 92,255 sequences for pretraining.

Downstream Tasks. We evaluate TAMM on diverse vision-language and temporal reasoning benchmarks with datasets disjoint from pretraining. For *cross-modal retrieval*, we use MIMIC-5×200 (Wang et al. 2022b) and report P@1, P@2, P@5, and P@10 for both image-to-text and text-to-image retrieval. For *zero-shot classification*, we evaluate on COVIDx (Wang, Lin, and Wong 2020) and RSNA Pneumonia (Shih et al. 2019), reporting accuracy and F1 scores. For *temporal image classification*, we fine-tune on Chest ImaGenome (Wu et al. 2021) and test on MS-CXR-T (Bannur et al. 2023), using macro-average accuracy over progression types. For *temporal sentence similarity classification*, we conduct zero-shot binary classification on MS-CXR-T (Bannur et al. 2023) sentence pairs, reporting accuracy.

Baselines We compare TAMM against recent state-of-the-art methods, grouped by whether they incorporate temporal dynamics. *Non-temporal vision-language models* focus on static image–text alignment: MGCA (Wang et al. 2022a) aligns radiographs and reports at multiple granularities; MedCLIP (Wang et al. 2022b) enhances contrastive learning with semantic matching; CARZero (Lai et al. 2024) utilizes LLM-based prompts and cross-attention for zero-shot classification; PRIOR (Cheng et al. 2023) integrates cross-modal reconstruction with global-local supervision; and MAVL (Phan et al. 2024) employs LLMs and disease decomposition for better attribute grounding. In contrast, *temporal modeling baselines* explicitly model longitudinal progression: Med-ST (Yang et al. 2024) introduces cycle consistency across time to capture disease evolution, and ALTA (Lian et al. 2025) adopts masked pretraining with temporal views and efficient tuning. These methods provide strong references to evaluate TAMM's temporal reasoning capability.

Implementation Details. We initialize TAMM using pretrained MGCA weights for faster convergence and improved stability. The model is trained for 8 epochs on 4 NVIDIA A40 GPUs. We use the AdamW optimizer (Loshchilov and Hutter 2017) with a weight decay of 1×10^{-6} and an initial learning rate of 1×10^{-5}. A cosine annealing scheduler with 40% warm-up is applied, gradually reducing the learning rate to 1×10^{-8}. All images are resized to 256×256, followed by a random crop to 224×224 for data augmentation. For downstream tasks, we either fine-tune a lightweight classification head or evaluate the pretrained model in a zero-shot manner, depending on the benchmark. For the *cross-modal retrieval* and *temporal sentence similarity classification* tasks, no standard deviation is reported because these are evaluated in a zero-shot manner on the pretrained model without additional training. In contrast, the *temporal image classification* and *zero-shot classification* tasks require fine-tuning a lightweight classification head. For these, we perform training with three different random seeds and report the mean and standard deviation of the results to capture variability due to training randomness. This evaluation protocol follows the standard setup used in prior baselines for fair comparison.

Main Results

To comprehensively evaluate TAMM, we divide our evaluation into two categories: standard vision-language benchmarks and temporal trend understanding tasks.

1) Standard Vision-Language Benchmarks:

Method	Text → Image				Image → Text			
	P@1	P@2	P@5	P@10	P@1	P@2	P@5	P@10
MGCA	74.50	<u>76.30</u>	71.16	60.85	63.62	63.88	<u>64.11</u>	<u>61.95</u>
MedCLIP	45.75	47.10	48.63	43.07	50.31	48.37	48.30	48.09
PRIOR	47.13	48.11	47.53	47.24	49.50	52.55	51.95	36.55
MAVL	50.00	50.00	42.91	37.27	50.00	50.00	49.47	50.00
CARZero	52.44	50.00	49.47	50.01	50.00	50.00	51.65	47.38
Med-ST	<u>75.25</u>	76.18	<u>71.36</u>	**65.04**	<u>64.50</u>	<u>63.97</u>	62.48	59.00
ALTA	69.80	65.85	52.60	37.72	56.80	55.40	46.80	46.80
Our	**78.69**	**79.26**	**73.07**	<u>64.03</u>	**66.44**	**67.63**	**66.05**	**62.50**

Table 1: Cross-modal retrieval results on the MIMIC-CXR 5×200 benchmark, reported as precision (%) at top-k (P@k). The best results are in bold, while the second-best results are underlined.

Method	RSNA Pneumonia		COVIDx	
	ACC	F1	ACC	F1
MGCA	85.06 ± 0.05	76.56 ± 0.10	92.13 ± 0.10	80.10 ± 0.65
MedCLIP	84.17 ± 0.16	<u>77.38 ± 0.35</u>	90.19 ± 0.09	75.60 ± 2.10
PRIOR	81.30 ± 0.35	72.02 ± 1.09	90.01 ± 0.19	**87.97 ± 0.68**
MAVL	79.99 ± 0.83	73.14 ± 0.48	88.77 ± 0.92	85.09 ± 1.71
CARZero	84.28 ± 0.07	76.94 ± 0.40	91.92 ± 0.20	85.83 ± 0.98
Med-ST	<u>85.15 ± 0.04</u>	76.96 ± 0.25	<u>92.27 ± 0.17</u>	80.22 ± 0.35
ALTA	84.24 ± 0.20	76.54 ± 0.61	91.53 ± 0.19	77.62 ± 1.26
Our	**85.48 ± 0.05**	**78.21 ± 0.14**	**93.03 ± 0.06**	<u>87.32 ± 0.38</u>

Table 2: Results for zero-shot classification on the RSNA Pneumonia and COVIDx datasets, reported as Accuracy and F1 (%). The best results are in bold, while the second-best results are underlined.

Cross-modal Retrieval on MIMIC-5×200. We evaluate TAMM's ability to align visual and textual information on the MIMIC-5×200 benchmark, a widely used setup for cross-modal retrieval. As shown in Table 1, TAMM consistently outperforms all baseline methods across all retrieval ranks in both text-to-image and image-to-text directions. This demonstrates TAMM's robust and comprehensive alignment capabilities. The improvements are not limited to a single metric or direction, but are broadly observed, indicating the effectiveness of temporal trend modeling in enhancing multimodal understanding under real-world clinical settings.

Zero-shot Classification Tasks. To further examine TAMM's transferability, we evaluate its visual encoder on two downstream classification tasks—RSNA Pneumonia and COVIDx—under a strict zero-shot setting. As reported in Table 2, TAMM achieves the highest accuracy on both datasets (85.48% on RSNA Pneumonia and 93.03% on COVIDx), without any task-specific fine-tuning. It also obtains the best F1 score on RSNA Pneumonia (78.21%) and competitive performance on COVIDx (87.32%), closely approaching the top result. These results suggest that TAMM learns generalizable and discriminative visual features that effectively transfer to unseen classification tasks.

2) Temporal Trend Understanding Tasks:

Temporal Image Classification. To evaluate the model's ability to capture disease progression trends from longitudinal chest X-rays, we conduct temporal image classification on the Med-ST benchmark. Specifically, the prediction target is to classify each case into one of three progression categories: improving, stable, or worsening, for each pathology. As shown in Table 3, our method achieves the highest macro accuracy across all five pathological conditions (Consolidation, Edema, Pleural Effusion, Pneumonia, and Pneumothorax). In each individual category, TAMM consistently outperforms other methods, achieving the best or second-best scores. These results clearly demonstrate the effectiveness of our temporal modeling strategy, which enables the model to integrate temporal cues and progression semantics into its decision-making process, making it well-suited for complex temporal classification tasks.

Temporal Sentence Similarity Classification. To further evaluate the temporal language understanding capability of TAMM, we conduct zero-shot sentence similarity classification on the MS-CXR-T dataset. Specifically, this experiment focuses on examining the text encoder component of each model to assess its ability to capture semantic changes and temporal trends in textual descriptions. As presented in Figure 2, TAMM achieves the highest accuracy (88.88%), outperforming all baseline methods by substantial margins. The RadGraph subset used in this experiment poses a relatively complex semantic challenge, further validating the robustness of our approach. By leveraging temporal modeling, TAMM is able to perceive subtle changes and better understand the underlying semantic context, thereby demonstrating its strong potential for temporal language reasoning in medical applications.

Case Study of Retrieval Outputs To better illustrate TAMM's semantic understanding capabilities, we present a case study of text-to-image and image-to-text retrieval examples in Figure 3. For each retrieval direction, we show the top two retrieved items. These qualitative results highlight TAMM's ability to retrieve semantically and clinically relevant matches that are better aligned with the query compared to baseline models. Specifically, the retrieved images exhibit similar pathological patterns, and the retrieved texts highlight consistent disease descriptions and findings. In contrast, many baseline methods tend to confuse visually or semantically similar conditions, such as misidentifying cardiomegaly as pleural effusion, indicating limited capacity in fine-grained disease understanding. This semantic and

Method	Consolidation	Edema	Pl. Effusion	Pneumonia	Pneumothorax
MGCA	44.99 ± 0.47	62.71 ± 0.24	57.17 ± 0.69	63.73 ± 0.89	52.16 ± 1.02
MedCLIP	53.40 ± 0.87	42.85 ± 0.01	49.96 ± 0.28	67.10 ± 0.22	55.46 ± 0.02
PRIOR	41.85 ± 1.51	43.97 ± 1.03	59.21 ± 1.19	41.41 ± 3.24	45.33 ± 2.13
MAVL	43.79 ± 0.00	42.85 ± 0.00	48.66 ± 0.00	67.10 ± 0.00	55.46 ± 0.00
CARZero	48.95 ± 1.25	51.92 ± 1.64	50.19 ± 1.72	57.44 ± 1.35	47.84 ± 1.78
Med-ST	60.57 ± 1.18	67.35 ± 0.32	58.47 ± 1.50	65.00 ± 0.34	54.18 ± 0.81
ALTA	52.76 ± 1.86	66.92 ± 0.89	50.92 ± 1.72	66.82 ± 0.19	51.04 ± 0.41
Our	**65.51 ± 0.83**	**69.22 ± 0.01**	**63.18 ± 0.45**	**72.71 ± 0.77**	**60.50 ± 0.01**

Table 3: Temporal image classification results. Accuracy (%) are reported across five conditions. The best results are in bold, while the second-best results are underlined.

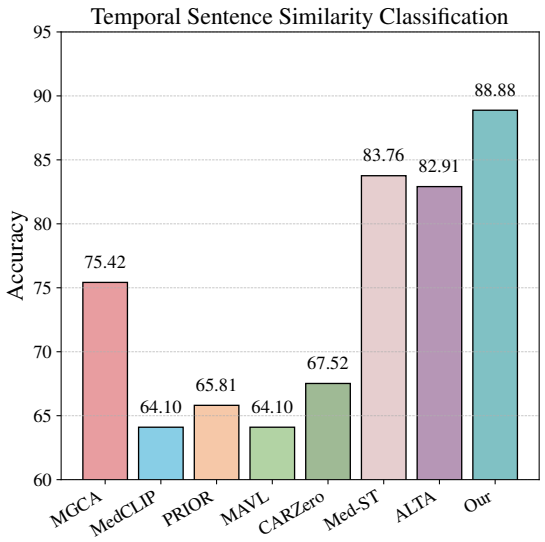

Figure 2: Accuracy (%) for the temporal sentence similarity classification task. Our model achieves the best performance, outperforming all baselines by a clear margin.

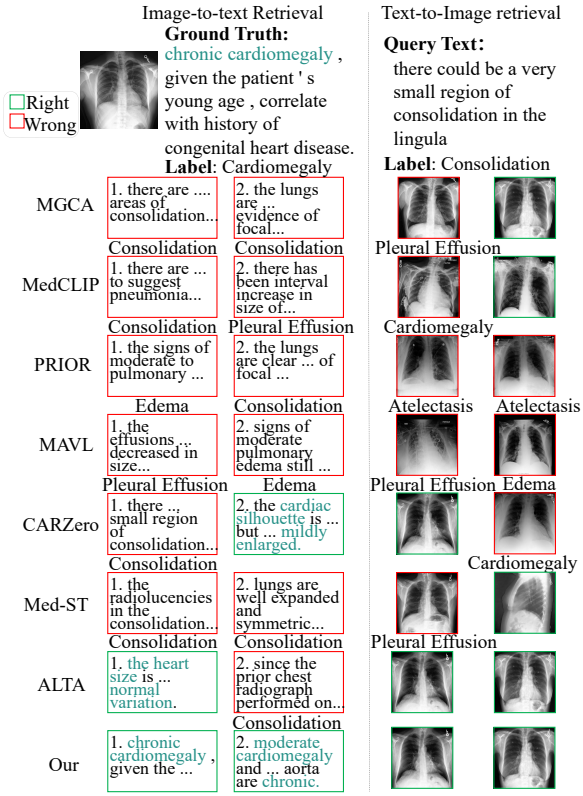

Figure 3: Qualitative results of text-to-image and image-to-text retrieval tasks. We show the top two retrieved images and top two retrieved texts for TAMM, with comparisons to seven baselines.We note the categories below wrongly retrieved samples.

pathological consistency demonstrates that TAMM effectively understands disease-relevant features in both modalities and aligns them accurately during retrieval. These visual comparisons, alongside the quantitative results, provide strong evidence of TAMM's superiority in cross-modal alignment and retrieval accuracy.

Additional case studies, including qualitative results for temporal image classification and phrase grounding, are provided in *Appendix: Additional Results for Visualization*.

Ablation Analysis of Loss Functions

We ablated TAMM's three core loss components—global alignment, trend alignment, and trend consistency—across four downstream tasks: cross-modal retrieval, zero-shot classification, temporal sentence similarity, and temporal image classification. Detailed analyses of the first two tasks are provided in *Appendix: Additional Results for Ablation*. Table 4 summarizes the results on the temporal tasks.

As shown in Table 4, removing any loss leads to a consistent performance drop, highlighting their complementary roles. Temporal modeling tasks are particularly sensitive to trend-related objectives. The trend alignment loss plays a central role in capturing disease progression and temporal similarity; its removal results in substantial accuracy drops in both temporal sentence and image classification tasks. The trend consistency loss further supports stable temporal reasoning by aligning predicted representations across ad-

Method	Temporal Sentence Similarity	Temporal Image Classification				
		Consolidation	Edema	Pleural Effusion	Pneumonia	Pneumothorax
Our	**88.88**	**65.51**	**69.22**	**63.18**	**72.71**	**60.50**
w/o consistency loss	81.57 (−7.31)	53.68 (−11.83)	55.64 (−13.58)	57.91 (−5.27)	65.41 (−7.30)	52.13 (−8.37)
w/o trend loss	75.33 (−13.55)	51.74 (−13.77)	46.62 (−22.60)	56.93 (−6.25)	64.56 (−8.15)	50.75 (−9.75)
w/o global loss	84.44 (−4.44)	56.72 (−8.79)	57.14 (−12.08)	60.58 (−2.60)	66.67 (−6.04)	54.98 (−5.52)

Table 4: Ablation study on temporal tasks on the MS-CXR-T dataset. Accuracy (%) for temporal sentence similarity and temporal image classification. Relative changes from the full model are shown below.

jacent timepoints; ablating it also leads to notable degradation. In contrast, removing the global alignment loss results in a smaller drop, as it primarily governs modality-invariant semantics rather than temporal structure. These results indicate that the trend alignment and consistency losses contribute most directly to temporal understanding, while all three components are necessary for optimal performance.

Sensitivity Analysis

Among TAMM's three loss terms—global alignment, trend alignment, and trend consistency—only the consistency loss is scaled by λ_{cons}, due to its inherently larger magnitude from the L_2-based reconstruction objective. The other two have comparable scales and require no extra weighting. To determine an appropriate value for the loss weight parameter λ_{cons}, we conduct a sensitivity analysis on the MIMIC-CXR validation set during pretraining. This avoids dependence on any downstream-specific supervision and ensures the choice is generalizable before fine-tuning. We experiment with $\lambda_{\text{cons}} \in \{1, 10, 20, 30\}$ and evaluate retrieval precision at top-k (P@5 and P@10) for both text-to-image and image-to-text tasks. Since our retrieval dataset is very large and cross-modal embeddings cannot be precisely aligned, we report P@5 and P@10 rather than P@1 or P@2. As shown in Figure 4, the model achieves the best and most stable performance when λ_{cons} is set to 10 or 20. We choose $\lambda_{\text{cons}} = 10$ as the default setting, balancing overall accuracy and stability.

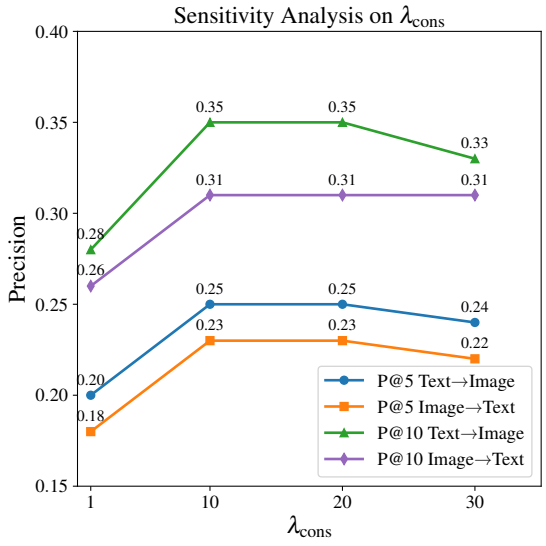

Figure 4: Sensitivity analysis of the loss weight parameter λ_{cons} on retrieval precision metrics (P@5 and P@10) for both text-to-image and image-to-text tasks on the MIMIC-CXR validation set. This evaluation is conducted at the pretraining stage.

Conclusion

In this work, we propose the TAMM framework, a temporal vision-language pretraining approach that leverages trend-aware alignment and consistency-based supervision to enhance multimodal representation learning. Rather than explicitly predicting disease progression labels, TAMM incorporates pseudo-trend labels and rationales generated by LLMs as soft guidance to capture subtle temporal dynamics and semantic correlations across image-text modalities.

This trend-aware supervision leads to improved performance across both *Standard Vision-Language Benchmarks* and *Temporal Trend Understanding Tasks*, such as retrieval, classification, and sentence similarity. Our results demonstrate that LLM-guided reasoning can effectively enhance temporal representation learning, underscoring the value of weakly supervised trend modeling for interpretable and accurate cross-modal alignment.

Importantly, the trend labels and rationales are not treated as hard targets but are integrated into contrastive and reconstruction losses as soft constraints. This design mitigates sensitivity to noisy annotations and encourages the model to focus on broader temporal patterns. As a result, TAMM remains robust and generalizable even under imperfect supervision. Further analysis of noisy label robustness is presented in *Appendix: Analysis of Temporal Supervision with Noisy Labels*.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.62432006, 62276159) and the Fundamental Research Program of Shanxi Province (No.202303021223004).

References

Alsentzer, E.; Murphy, J. R.; Boag, W.; Weng, W.-H.; Jin, D.; Naumann, T.; and McDermott, M. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*.

Bannur, S.; Hyland, S.; Liu, Q.; Perez-Garcia, F.; Ilse, M.; Castro, D. C.; Boecking, B.; Sharma, H.; Bouzid, K.; Thieme, A.; et al. 2023. Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15016–15027.

Boecking, B.; Usuyama, N.; Bannur, S.; Castro, D. C.; Schwaighofer, A.; Hyland, S.; Wetscherek, M.; Naumann, T.; Nori, A.; Alvarez-Valle, J.; et al. 2022. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, 1–21. Springer.

Chen, W.; Shen, L.; Lin, J.; Luo, J.; Li, X.; and Yuan, Y. 2023. Fine-grained image-text alignment in medical imaging enables explainable cyclic image-report generation. *arXiv preprint arXiv:2312.08078*.

Cheng, P.; Lin, L.; Lyu, J.; Huang, Y.; Luo, W.; and Tang, X. 2023. PRIOR: Prototype Representation Joint Learning from Medical Images and Reports. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 21304–21314.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gai, X.; Zhou, C.; Liu, J.; Feng, Y.; Wu, J.; and Liu, Z. 2024. Medthink: Explaining medical visual question answering via multimodal decision-making rationale. *arXiv preprint arXiv:2404.12372*.

Huang, S.-C.; Shen, L.; Lungren, M. P.; and Yeung, S. 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*, 3942–3951.

Huang, W.; Li, C.; Zhou, H.-Y.; Yang, H.; Liu, J.; Liang, Y.; Zheng, H.; Zhang, S.; and Wang, S. 2024. Enhancing representation in radiography-reports foundation model: A granular alignment algorithm using masked contrastive learning. *Nature Communications*, 15(1): 7620.

Johnson, A. E.; Pollard, T. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Peng, Y.; Lu, Z.; Mark, R. G.; Berkowitz, S. J.; and Horng, S. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Lai, H.; Yao, Q.; Jiang, Z.; Wang, R.; He, Z.; Tao, X.; and Zhou, S. K. 2024. Carzero: Cross-attention alignment for radiology zero-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11137–11146.

Li, M.; Meng, M.; Fulham, M.; Feng, D. D.; Bi, L.; and Kim, J. 2025. Enhancing medical vision-language contrastive learning via inter-matching relation modelling. *IEEE Transactions on Medical Imaging*.

Lian, C.; Zhou, H.-Y.; Liang, D.; Qin, J.; and Wang, L. 2025. Efficient Medical Vision-Language Alignment Through Adapting Masked Vision Models. *IEEE Transactions on Medical Imaging*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Moon, J. H.; Lee, H.; Shin, W.; Kim, Y.-H.; and Choi, E. 2022. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics*, 26(12): 6070–6080.

Moor, M.; Huang, Q.; Wu, S.; Yasunaga, M.; Dalmia, Y.; Leskovec, J.; Zakka, C.; Reis, E. P.; and Rajpurkar, P. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, 353–367. PMLR.

Phan, V. M. H.; Xie, Y.; Qi, Y.; Liu, L.; Liu, L.; Zhang, B.; Liao, Z.; Wu, Q.; To, M.; and Verjans, J. W. 2024. Decomposing Disease Descriptions for Enhanced Pathology Detection: A Multi-Aspect Vision-Language Pre-Training Framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11492–11501. IEEE.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Shaib, C.; Li, M. L.; Joseph, S.; Marshall, I. J.; Li, J. J.; and Wallace, B. C. 2023. Summarizing, Simplifying, and Synthesizing Medical Evidence Using GPT-3 (with Varying Success). *arXiv preprint arXiv:2305.06299*.

Shih, G.; Wu, C.-C.; Halabi, S. S.; Kohli, M. D.; Prevedello, L. M.; Cook, T. S.; Sharma, A.; Amorosa, J. K.; Arteaga, V.; Galperin-Aizenberg, M.; et al. 2019. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1): e180041.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2022. Large Language Models Encode Clinical Knowledge. *arXiv preprint arXiv:2212.13138*.

Sultan, R. I.; Zhu, H.; Li, C.; and Zhu, D. 2025. BiPVL-Seg: Bidirectional Progressive Vision-Language Fusion with Global-Local Alignment for Medical Image Segmentation. *arXiv preprint arXiv:2503.23534*.

Wan, Z.; Liu, C.; Zhang, M.; Fu, J.; Wang, B.; Cheng, S.; Ma, L.; Quilodrán-Casas, C.; and Arcucci, R. 2023. Med-UNIC: Unifying cross-lingual medical vision-language pre-training by diminishing bias. *arXiv preprint arXiv:2305.19894*.

Wang, F.; Zhou, Y.; Wang, S.; Vardhanabhuti, V.; and Yu, L. 2022a. Multi-granularity cross-modal alignment for gener-

alized medical visual representation learning. In *Advances in Neural Information Processing Systems*.

Wang, H.; Liu, C.; Xi, N.; Qiang, Z.; Zhao, S.; Qin, B.; and Liu, T. 2023. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*.

Wang, L.; Lin, Z. Q.; and Wong, A. 2020. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific reports*, 10(1): 19549.

Wang, X.; Luo, J.; Wang, J.; Zhong, Y.; Zhang, X.; Wang, Y.; Bhatia, P.; Xiao, C.; and Ma, F. 2024. Unity in diversity: Collaborative pre-training across multimodal medical sources. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2024, 3644.

Wang, Y.; Zhao, Y.; and Petzold, L. 2023. Are large language models ready for healthcare? a comparative study on clinical language understanding. In *Machine learning for healthcare conference*, 804–823. PMLR.

Wang, Z.; Wu, Z.; Agarwal, D.; and Sun, J. 2022b. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, 3876.

Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF international conference on computer vision*, 21372–21383.

Wu, J.; Agu, N.; Lourentzou, I.; Sharma, A.; Paguio, J.; Yao, J. S.; Dee, E. C.; Mitchell, W.; Kashyap, S.; Giovannini, A.; Celi, L. A.; Syeda-Mahmood, T.; and Moradi, M. 2021. Chest ImaGenome Dataset (Version 1.0.0). https://physionet.org/content/chest-imagenome/1.0.0/. PhysioNet.

Yan, H.; Yang, X.; Bai, L.; Li, J.; and Liang, J. 2025. Multi-Grained Vision-and-Language Model for Medical Image and Text Alignment. *IEEE Transactions on Multimedia*.

Yang, J.; Su, B.; Zhao, W. X.; and Wen, J.-R. 2024. Unlocking the power of spatial and temporal information in medical multimodal pre-training. *arXiv preprint arXiv:2405.19654*.

Yunxiang, L.; Zihan, L.; Kai, Z.; Ruilong, D.; and You, Z. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*.

Zhang, K.; Yang, Y.; Yu, J.; Jiang, H.; Fan, J.; Huang, Q.; and Han, W. 2023a. Multi-task paired masking with alignment modeling for medical vision-language pre-training. *IEEE Transactions on Multimedia*, 26: 4706–4721.

Zhang, X.; Wu, C.; Zhang, Y.; Xie, W.; and Wang, Y. 2023b. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14(1): 4542.

Zhang, Z.; Yu, Y.; Chen, Y.; Yang, X.; and Yeo, S. Y. 2025. Medunifier: Unifying vision-and-language pre-training on medical data with vision generation task using discrete visual representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 29744–29755.

Zhou, H.; Chen, X.; Zhang, Y.; Luo, R.; Wang, L.; and Yu, Y. 2022. Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. *Nature Machine Intelligence*, 4(1): 32–40.

Zhou, H.; Lian, C.; Wang, L.; and Yu, Y. 2023. Advancing radiograph representation learning with masked record modeling. *arXiv preprint arXiv:2301.13155*.