CLINICAL TRAJECTORY ALIGNMENT FOR MEDICAL VISION-LANGUAGE PRE-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical vision-language pre-training largely follows a visit-level image-report matching paradigm, aligning paired images and reports at individual visits. While effective for static cross-modal correspondence, this paradigm provides limited supervision for longitudinal clinical change, such as whether abnormalities improve, remain stable, or worsen over time. Learning such change is challenging because temporal semantics are implicit in free-text reports, and different abnormalities within the same patient may evolve asynchronously or even in opposite directions. We propose MedCTA, which reframes medical vision-language pre-training from visit-level cross-modal matching to learning clinical change. Rather than compressing a patient history into a single temporal representation, MedCTA models clinical change at two complementary scopes. At the abnormality scope, clinically grounded queries construct abnormality-conditioned visual and textual trajectories to capture heterogeneous abnormality evolution. At the patient-course scope, global image and report sequences are modeled to capture overall clinical progression beyond any individual abnormality. Structured trend supervision is extracted from longitudinal reports by an offline LLM parser, removing the need for manual temporal annotations. Combined with static image-report alignment, MedCTA learns representations that preserve visit-level cross-modal correspondence while encoding longitudinal change semantics. Experiments on temporal image classification, image-text retrieval, and zero-shot classification show consistent gains over strong medical vision-language baselines.

1 INTRODUCTION

Medical vision–language pre-training (VLP) aims to learn clinically meaningful cross-modal representations from medical images and their corresponding reports. Existing methods based on image-text contrastive learning Miao et al. (2026), masked modeling Cao et al. (2025), and local region-text matching Yan et al. (2025) have achieved notable progress in image-text retrieval, disease classification, and phrase grounding Mei et al. (2024). However, most existing methods largely follow a visit-level image-report matching paradigm, where paired images and reports are aligned at individual visits Bluethgen et al. (2025); Yang et al. (2026). Such supervision is effective for learning static cross-modal correspondence, but provides limited guidance for longitudinal clinical change, such as whether abnormalities improve, remain stable, or worsen over time You et al. (2025); Yang & Shen (2025). As a result, the temporal evolution semantics contained in patient follow-up records are not explicitly modeled, limiting the ability of current models to learn abnormality progression from longitudinal medical data, as illustrated in Figure 1.

Learning longitudinal clinical change is challenging for two main reasons. First, temporal semantics are often implicit in free-text radiology reports rather than available as structured annotations. Reports often describe abnormality changes through expressions such as “increased opacity”, “interval improvement”, or “no significant change”, but these descriptions cannot be directly used as supervision without clinical semantic parsing Cui et al. (2025); Liu et al. (2026). Second, abnormality progression is inherently heterogeneous. Different abnormalities within the same patient may evolve asynchronously or even in opposite directions. Therefore, compressing an entire patient history into a single temporal feature may oversimplify longitudinal abnormality evolution and fail to capture abnormality-specific changes Zhang et al. (2024); Fan et al. (2026).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

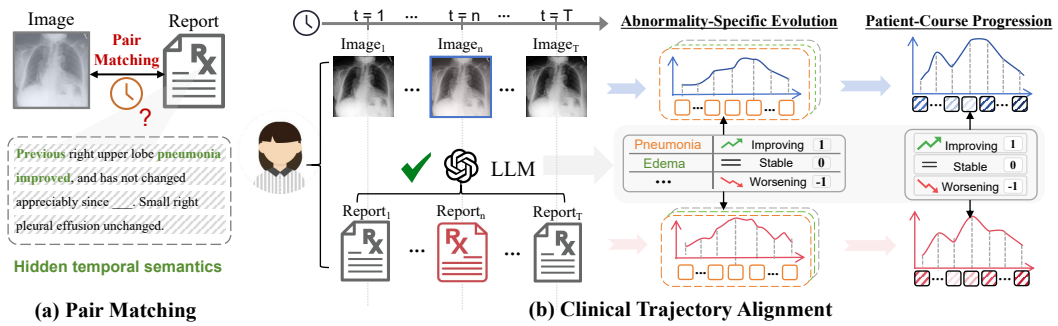


Figure 1: Motivation of MedCTA. (a) Static image–report matching captures visit-level correspondence but overlooks longitudinal clinical change. (b) MedCTA exploits follow-up records with LLM-derived trend supervision to model abnormality-specific and patient-course trajectories.

To address these challenges, we propose MedCTA, a clinical trajectory alignment framework that reframes medical vision–language pre-training from visit-level cross-modal matching to modeling clinical change. Instead of compressing a patient's longitudinal history into a single temporal representation, MedCTA explicitly models longitudinal changes at two complementary scopes: abnormality-specific evolution and patient-course progression. At the abnormality scope, MedCTA constructs abnormality-conditioned visual and textual trajectories across visits using clinically grounded abnormality queries. These trajectories capture the evolution of each abnormality over time and are supervised by abnormality-specific trend labels through Abnormality-Specific Trajectory Alignment. At the patient-course scope, MedCTA further models global image and report sequences across visits through Patient-Course Progression Alignment, thereby capturing the patient's overall clinical progression beyond any individual abnormality. To train these alignment objectives without manual trend annotations, MedCTA adopts an offline large language model to extract structured trend supervision from longitudinal report sequences, including abnormality-specific evolution trends and patient-course progression trends. By jointly optimizing static visit-level image–report alignment, Abnormality-Specific Trajectory Alignment, and Patient-Course Progression Alignment, MedCTA learns cross-modal representations that preserve single-visit image-text correspondence while capturing longitudinal abnormality evolution and patient-course progression.

The main contributions of this paper are summarized as follows:

• We reformulate longitudinal medical VLP as learning from clinical change, extending static visit-level image–report matching to trajectory-level temporal semantic supervision.

• We propose MedCTA, a trajectory alignment framework that models longitudinal change through abnormality-specific trajectories and patient-course progression, thereby capturing heterogeneous abnormality evolution and overall patient-level clinical progression.

• We derive structured trend supervision using an offline LLM parser and demonstrate consistent improvements on temporal classification, image-text retrieval, and zero-shot classification.

2 Related Work

2.1 Medical VLP

Medical VLP learns transferable cross-modal representations from medical images and radiology reports Lai et al. (2026); Nath et al. (2025). Existing methods have made progress through image–text contrastive learning Wang et al. (2022b); Lai et al. (2024), masked modeling Cheng et al. (2023), and fine-grained region–text alignment Yan et al. (2025); Yang et al. (2026), benefiting tasks such as image–text retrieval, disease classification, phrase grounding, and report understanding Li et al. (2026); Chen et al. (2025). However, most methods follow a visit-level image–report matching paradigm, where each pair is treated as an independent static sample. This paradigm learns effective single-visit correspondence but provides limited supervision for longitudinal clinical change, such as improvement, stability, or worsening. Different from prior static medical VLP methods, we reframe

pre-training from visit-level cross-modal matching to clinical change learning from longitudinal patient trajectories.

2.2 Longitudinal Temporal Modeling in Medical VLP

Longitudinal radiology data contain temporal cues about clinical progression across follow-up examinations Yang & Shen (2025); Bannur et al. (2023); Liu et al. (2025). Recent methods have explored such information: ALTA Lian et al. (2025) uses prior X-ray views as temporal context, Med-ST Yang et al. (2024b) introduces cross-modal cycle consistency with historical image–text pairs, and TAMM Bai et al. (2026) leverages LLM-derived supervision from adjacent reports. However, these methods mainly capture visual context, temporal correspondence, or short-range local transitions, and thus may overlook heterogeneous abnormality evolution and global patient-course progression. Our method addresses this limitation by modeling clinical change through abnormality-specific trajectories and patient-course progression.

3 Method

3.1 From Static Pair Matching to Change-Aware Pre-training

Existing medical VLP methods typically align paired images and reports at individual visits. While effective for static cross-modal correspondence, this visit-level matching paradigm provides limited supervision for longitudinal clinical change, such as whether abnormalities improve, remain stable, or worsen over time. MedCTA reframes medical VLP from static pair matching to change-aware pre-training by modeling how clinical states evolve along patient trajectories.

Formally, we consider a longitudinal medical VLP setting, where each patient is associated with a temporally ordered sequence of image–report pairs:

$$\mathcal{D} = \{\mathcal{S}_i\}_{i=1}^{N}, \qquad \mathcal{S}_i = \{(I_{i,t}, R_{i,t}, \tau_{i,t})\}_{t=1}^{T_i}, \tag{1}$$

where $I_{i,t}$ and $R_{i,t}$ denote the medical image and the corresponding radiology report of patient i at the t-th visit, $\tau_{i,t}$ is the examination timestamp, and T_i is the number of visits.

MedCTA retains static image–report alignment to preserve visit-level cross-modal correspondence. For a batch of B image–report pairs, let $\mathbf{v}_b$ and $\mathbf{r}_b$ denote the global image and report representations of the b-th pair, respectively. The static image–report alignment loss is defined as

$$\mathcal{L}_{\text{clip}} = \frac{1}{2}\left(\mathcal{L}_{I\to T} + \mathcal{L}_{T\to I}\right), \quad \mathcal{L}_{I\to T} = -\frac{1}{B}\sum_{b=1}^{B}\log\frac{\exp(\text{sim}(\mathbf{v}_b, \mathbf{r}_b)/\tau)}{\sum_{j=1}^{B}\exp(\text{sim}(\mathbf{v}_b, \mathbf{r}_j)/\tau)}, \tag{2}$$

where $\mathcal{L}_{T\to I}$ is symmetric, $\text{sim}(\cdot,\cdot)$ is cosine similarity, and τ is the temperature.

However, $\mathcal{L}_{\text{clip}}$ only learns visit-level state matching and cannot explicitly supervise longitudinal change. Therefore, MedCTA further introduces two change-aware objectives: Abnormality-Specific Trajectory Alignment for heterogeneous abnormality evolution and Patient-Course Progression Alignment for global clinical progression, as shown in Figure 2.

3.2 Structured Trend Supervision from Longitudinal Reports

The temporal semantics of longitudinal radiology reports are usually implicit in free-text reports and cannot be directly used as structured supervision. To obtain learnable temporal supervision, we use an LLM as an *offline parser* to extract structured trend labels from complete longitudinal report sequences. For patient i, the longitudinal report sequence is

$$\mathcal{R}_i = \{R_{i,1}, R_{i,2}, \ldots, R_{i,T_i}\}. \tag{3}$$

Specifically, MedCTA uses the same offline LLM parser $\mathcal{G}_{\text{LLM}}$ with task-specific prompts to derive two types of trend labels from $\mathcal{R}_i$: abnormality-specific trend labels and patient-course trend labels. Given a fixed abnormality set $\mathcal{A} = \{a_k\}_{k=1}^{K}$, the abnormality-specific trend label is generated as

$$y_{i,k}^{a} = \mathcal{G}_{\text{LLM}}(\mathcal{R}_i, a_k; \pi_{\text{abn}}), \qquad y_{i,k}^{a} \in \{-1, 0, 1\}, \tag{4}$$

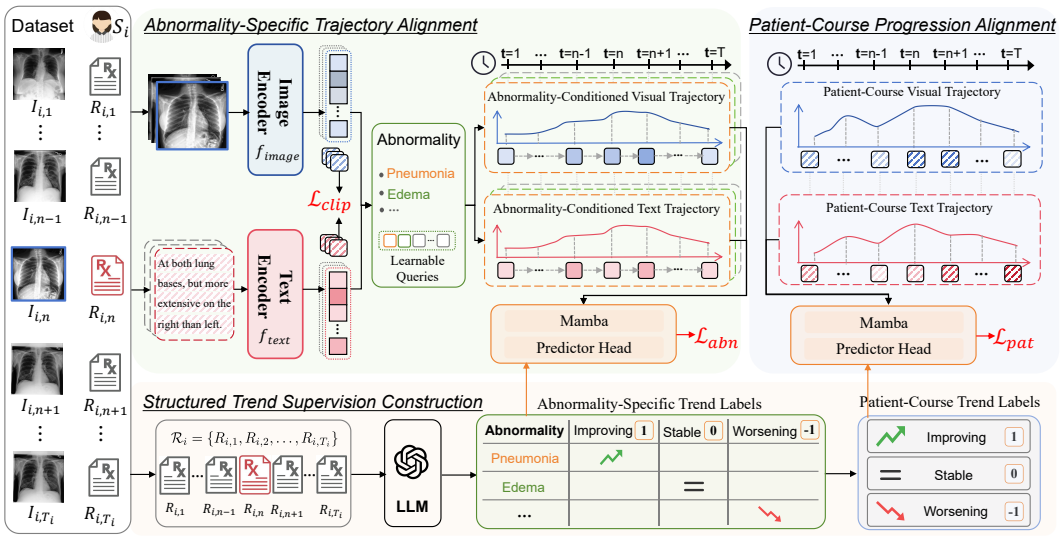


Figure 2: Overview of MedCTA. Given a longitudinal image–report sequence $\mathcal{S}_i$, MedCTA jointly performs static image–report alignment, Abnormality-Specific Trajectory Alignment, and Patient-Course Progression Alignment. Abnormality-specific and patient-course trajectories are temporally encoded by Mamba and constrained by LLM-derived trend supervision, enabling semantic-level trajectory alignment for abnormality evolution and overall clinical progression.

where π_{abn} denotes the abnormality-level prompting instruction. The labels -1, 0, and 1 indicate worsening, stable condition, and improvement, respectively. Since not all abnormalities appear in each patient sequence, we use a validity mask $m_{i,k} \in \{0, 1\}$, where $m_{i,k} = 1$ if abnormality a_k has an available trend label for patient i, and $m_{i,k} = 0$ otherwise.

In parallel, the patient-course trend label is generated by the same LLM parser with a patient-level prompt:

$$y_i^p = \mathcal{G}_{\text{LLM}}(\mathcal{R}_i; \pi_{\text{pat}}), \qquad y_i^p \in \{-1, 0, 1\}, \tag{5}$$

where π_{pat} denotes the patient-course prompting instruction. This label provides sequence-level supervision for patient-course progression modeling.

3.3 Abnormality-Conditioned Trajectory Construction

To model the longitudinal evolution of individual abnormalities, abnormality queries serve as semantic selectors that identify abnormality-relevant visual and textual representations at each visit. These visit-wise representations are then organized across time to form abnormality-conditioned visual and textual trajectories. Given a predefined abnormality set $\mathcal{A} = \{a_k\}_{k=1}^K$, we associate each abnormality a_k with a learnable query $\mathbf{q}_k \in \mathbb{R}^d$, and collect all queries as $\mathbf{Q} = [\mathbf{q}_1; \mathbf{q}_2; \cdots; \mathbf{q}_K] \in \mathbb{R}^{K \times d}$. To inject clinical semantic priors, we initialize each query using the global [CLS] embedding of an abnormality-specific prompt P_k, e.g., "a clinical finding of a_k", encoded by the text encoder Alsentzer et al. (2019):

$$\mathbf{q}_k^{(0)} = f_{\text{text}}^{\text{CLS}}(P_k), \qquad k = 1, \ldots, K. \tag{6}$$

The initialized query $\mathbf{q}_k^{(0)}$ is then treated as the initial value of the learnable parameter $\mathbf{q}_k$ and is updated during pre-training.

To adapt the query to different modalities, we project it into image and text spaces:

$$\mathbf{q}_k^I = \mathbf{W}_I \mathbf{q}_k, \qquad \mathbf{q}_k^T = \mathbf{W}_T \mathbf{q}_k, \tag{7}$$

where $\mathbf{W}_I$ and $\mathbf{W}_T$ are modality-specific projection matrices.

For each visit $(I_{i,t}, R_{i,t})$, the image encoder f_{image} Dosovitskiy et al. (2021) and text encoder f_{text} Alsentzer et al. (2019) produce global features for visit-level alignment and local features for

abnormality-specific extraction:

$$f_{\text{image}} : I_{i,t} \rightarrow (\mathbf{g}^I_{i,t}, \mathbf{H}^I_{i,t}), \quad f_{\text{text}} : R_{i,t} \rightarrow (\mathbf{g}^T_{i,t}, \mathbf{H}^T_{i,t}). \tag{8}$$

where $\mathbf{g}^I_{i,t}$ and $\mathbf{g}^T_{i,t}$ denote the global image and text representations, respectively, while $\mathbf{H}^I_{i,t}$ and $\mathbf{H}^T_{i,t}$ denote the patch-level image features and token-level report features, respectively.

The modality-specific queries attend to local features to obtain abnormality-specific visual and textual states:

$$\mathbf{e}^I_{i,t,k} = \text{Attn}(\mathbf{q}^I_k, \mathbf{H}^I_{i,t}, \mathbf{H}^I_{i,t}), \qquad \mathbf{e}^T_{i,t,k} = \text{Attn}(\mathbf{q}^T_k, \mathbf{H}^T_{i,t}, \mathbf{H}^T_{i,t}), \tag{9}$$

where $\mathbf{e}^I_{i,t,k}$ and $\mathbf{e}^T_{i,t,k}$ denote the image-side and text-side abnormality-specific embeddings of a_k at visit t, respectively. In this way, the image and text representations are organized into abnormality-conditioned semantic spaces.

Given the visit-wise abnormality-conditioned representations, MedCTA constructs visual and textual trajectories for each abnormality by organizing these representations across visits. For patient i and abnormality a_k, the resulting image-side and text-side abnormality trajectories are defined as:

$$\mathcal{E}^I_{i,k} = \{\mathbf{e}^I_{i,1,k}, \mathbf{e}^I_{i,2,k}, \ldots, \mathbf{e}^I_{i,T_i,k}\}, \qquad \mathcal{E}^T_{i,k} = \{\mathbf{e}^T_{i,1,k}, \mathbf{e}^T_{i,2,k}, \ldots, \mathbf{e}^T_{i,T_i,k}\}. \tag{10}$$

3.4 Abnormality-Specific Trajectory Alignment

Given the abnormality-conditioned trajectories, we use modality-specific Mamba-based temporal encoders Gu & Dao (2023) to encode the image-side and text-side evolution of each abnormality:

$$\mathbf{z}^{I,a}_{i,k} = g^{I,a}_{\text{Mamba}}(\mathcal{E}^I_{i,k}), \qquad \mathbf{z}^{T,a}_{i,k} = g^{T,a}_{\text{Mamba}}(\mathcal{E}^T_{i,k}), \tag{11}$$

where $\mathbf{z}^{I,a}_{i,k}$ and $\mathbf{z}^{T,a}_{i,k}$ denote the temporally encoded image-side and text-side trajectory representations of abnormality a_k for patient i, respectively. Padding masks are applied during temporal encoding to handle variable-length visit sequences across patients.

Instead of directly matching image and text trajectory embeddings, MedCTA aligns them in a shared clinical-change space supervised by the LLM-derived abnormality trend label $y^a_{i,k}$. The image-side and text-side trend predictions are computed as

$$\hat{\mathbf{p}}^{I,a}_{i,k} = \text{softmax}(h^a_I(\mathbf{z}^{I,a}_{i,k})), \qquad \hat{\mathbf{p}}^{T,a}_{i,k} = \text{softmax}(h^a_T(\mathbf{z}^{T,a}_{i,k})), \tag{12}$$

where $h^a_I(\cdot)$ and $h^a_T(\cdot)$ are image-side and text-side abnormality-specific trend classification heads.

The abnormality-specific trajectory alignment loss is defined as

$$\mathcal{L}_{\text{abn}} = \frac{1}{2\sum_{i=1}^{N}\sum_{k=1}^{K} m_{i,k}} \sum_{i=1}^{N}\sum_{k=1}^{K} m_{i,k} \left[\text{CE}\left(\hat{\mathbf{p}}^{I,a}_{i,k}, y^a_{i,k}\right) + \text{CE}\left(\hat{\mathbf{p}}^{T,a}_{i,k}, y^a_{i,k}\right)\right], \tag{13}$$

where $\text{CE}(\cdot,\cdot)$ denotes the cross-entropy loss, and $m_{i,k}$ is a validity mask indicating whether abnormality a_k has an available trend label for patient i. The label $y^a_{i,k}$ is mapped to a three-class trend category, corresponding to worsening, stable, and improving conditions. This objective encourages image and text trajectories of the same abnormality to be consistent at the trend-semantic level.

3.5 Patient-Course Progression Alignment

While abnormality-specific trajectories capture fine-grained evolution, they may not fully reflect the overall patient course. MedCTA therefore models patient-course progression from global longitudinal image and text representations.

For patient i, we organize the visit-level global representations from Eq. (8) into image and text sequences:

$$\mathcal{U}^I_i = \{\mathbf{g}^I_{i,t}\}_{t=1}^{T_i}, \qquad \mathcal{U}^T_i = \{\mathbf{g}^T_{i,t}\}_{t=1}^{T_i}. \tag{14}$$

To capture long-range patient-course progression, we perform Mamba-based temporal modeling Gu & Dao (2023) over the global image and text sequences:

$$\mathbf{z}^{I,p}_i = g^{I,p}_{\text{Mamba}}(\mathcal{U}^I_i), \qquad \mathbf{z}^{T,p}_i = g^{T,p}_{\text{Mamba}}(\mathcal{U}^T_i). \tag{15}$$

where $\mathbf{z}_i^{I,p}$ and $\mathbf{z}_i^{T,p}$ denote the image-side and text-side patient-course progression representations.

The learned patient-course representations are supervised by the LLM-derived patient-course trend label y_i^p. The image-side and text-side predictions are computed as

$$\hat{\mathbf{p}}_i^{I,p} = \mathrm{softmax}(h_I^p(\mathbf{z}_i^{I,p})), \qquad \hat{\mathbf{p}}_i^{T,p} = \mathrm{softmax}(h_T^p(\mathbf{z}_i^{T,p})), \tag{16}$$

where $h_I^p(\cdot)$ and $h_T^p(\cdot)$ are patient-course trend classification heads. The patient-course progression alignment loss is defined as

$$\mathcal{L}_{\mathrm{pat}} = \frac{1}{2N} \sum_{i=1}^{N} \left[\mathrm{CE}\left(\hat{\mathbf{p}}_i^{I,p}, y_i^p\right) + \mathrm{CE}\left(\hat{\mathbf{p}}_i^{T,p}, y_i^p\right) \right], \tag{17}$$

where y_i^p is mapped to a three-class label corresponding to overall worsening, stable condition, and improvement. This loss encourages the model to capture the overall clinical progression of each patient across longitudinal visits, thereby enabling patient-course progression alignment.

3.6 Joint Optimization with Static Image–Report Alignment

MedCTA is trained by jointly optimizing static image–report alignment, abnormality-specific trajectory alignment, and patient-course progression alignment:

$$\mathcal{L} = \mathcal{L}_{\mathrm{clip}} + \mathcal{L}_{\mathrm{abn}} + \mathcal{L}_{\mathrm{pat}}, \tag{18}$$

where $\mathcal{L}_{\mathrm{clip}}$ preserves visit-level cross-modal correspondence, $\mathcal{L}_{\mathrm{abn}}$ supervises abnormality-specific evolution, and $\mathcal{L}_{\mathrm{pat}}$ models overall patient-course progression.

4 Experiments

4.1 Experimental Setting

We evaluate MedCTA on both longitudinal temporal reasoning and cross-modal understanding tasks. Below, we briefly describe the pre-training data, downstream benchmarks, and baselines. Additional implementation and evaluation details are provided in Appendix A.

Pre-training Data. We pre-train MedCTA on MIMIC-CXR Johnson et al. (2019) and reorganize image–report pairs into patient-level longitudinal sequences ordered by study time. Following prior temporal VLP settings Yang et al. (2024b), each sequence contains up to four consecutive visits.

Downstream Tasks. We evaluate MedCTA on temporal reasoning and cross-modal understanding tasks. Temporal reasoning includes temporal image classification, temporal sentence similarity classification, and dynamic phrase grounding on Chest ImaGenome Wu et al. (2021) and MS-CXR-T Bannur et al. (2023). Cross-modal understanding includes image–text retrieval on MIMIC-5×200 Wang et al. (2022b) and zero-shot classification on COVIDx Wang et al. (2020) and RSNA Pneumonia Shih et al. (2019).

Baselines. We compare MedCTA with representative medical vision–language methods under the same evaluation settings. The baselines include non-temporal VLP methods, i.e., MGCA Wang et al. (2022a), MedCLIP Wang et al. (2022b), CARZero Lai et al. (2024), PRIOR Cheng et al. (2023), and MAVL Phan et al. (2024), as well as temporal medical VLP methods, i.e., Med-ST Yang et al. (2024b), ALTA Lian et al. (2025), and TAMM Bai et al. (2026). When official code is available, we reproduce results using the released implementations; otherwise, we follow the original protocols with the same backbone and training settings for fair comparison.

4.2 Experimental Results

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

Table 1: Temporal image classification results. Accuracy (%) is reported across five abnormality categories. Best results are shown in bold, and second-best results are underlined.

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
TAMM	<u>65.51 ± 0.83</u>	<u>69.22 ± 0.01</u>	<u>63.18 ± 0.45</u>	<u>72.71 ± 0.77</u>	<u>60.50 ± 0.01</u>
Ours	**67.14 ± 0.82**	**70.30 ± 0.16**	**65.59 ± 1.03**	**72.85 ± 0.35**	**61.66 ± 0.05**

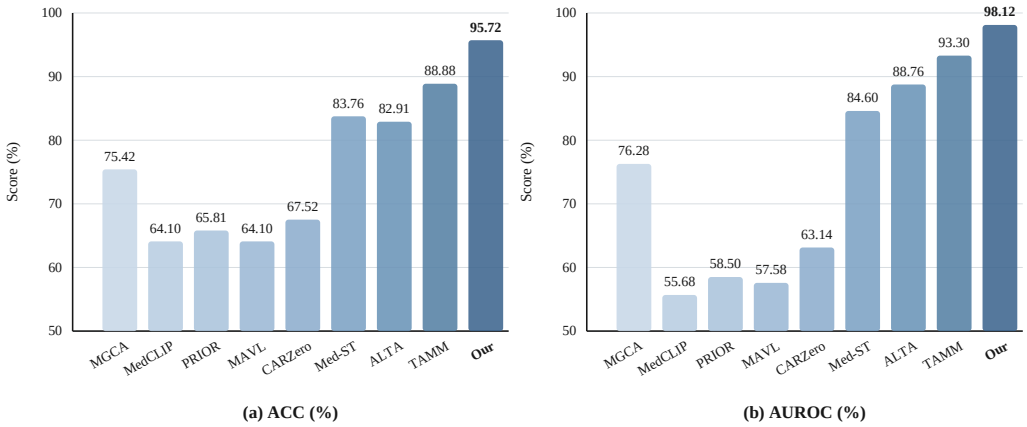


Figure 3: Temporal sentence similarity classification results on MS-CXR-T: (a) ACC and (b) AUROC, both reported as percentages.

4.2.1 Temporal Trend Understanding Tasks

Temporal Image Classification. To evaluate temporal progression modeling from longitudinal chest X-rays, we fine-tune on Chest ImaGenome and evaluate temporal image classification on MS-CXR-T over five abnormalities: Consolidation, Edema, Pleural Effusion, Pneumonia, and Pneumothorax. The task predicts one of three progression categories, i.e., `improving`, `stable`, and `worsening`, for each abnormality. As shown in Table 1, MedCTA achieves superior performance across all diseases. This indicates that explicitly modeling longitudinal clinical change at both abnormality and patient-course levels provides more discriminative representations than visit-level image–report matching alone, especially for distinguishing improvement, stability, and worsening.

Temporal Sentence Similarity Classification. To evaluate the temporal language understanding ability of MedCTA, we conduct zero-shot sentence similarity classification on the MS-CXR-T dataset. In particular, this experiment examines the text encoder of each model to determine how well it captures semantic shifts and temporal progression in textual descriptions. Fig. 3 shows that MedCTA achieves the highest ACC of 95.72% and AUROC of 98.12%. This result indicates that MedCTA learns more temporally aware textual representations, enabling better recognition of progression-related semantics such as improvement, stability, and worsening in reports.

Dynamic Phrase Grounding. We evaluate dynamic phrase grounding using Chest ImaGenome bounding-box annotations, aiming to localize temporal progression descriptions across two consecutive chest X-rays. Specifically, we compute cosine similarity between patch-level image embeddings and temporal description embeddings, and upsample the patch-wise similarity scores to generate grounding heatmaps. Fig. 4 presents representative examples from three progression categories, i.e., `improving`, `stable`, and `worsening`, with the left and right images denoting the prior and current studies, respectively. The results show that MedCTA can better localize regions as-

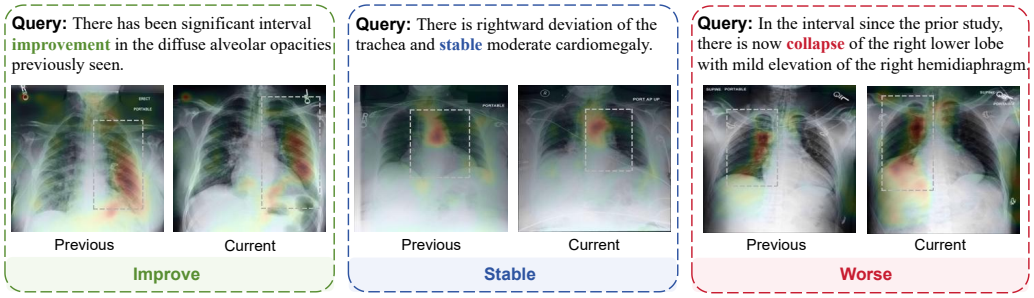


Figure 4: Qualitative dynamic phrase grounding results on Chest ImaGenome for improving, stable, and worsening cases. The left and right images denote the prior and current studies, respectively, with highlighted regions indicating visual evidence related to temporal progression.

sociated with longitudinal changes by organizing abnormality-relevant visual representations across visits.

4.2.2 Standard Vision-Language Benchmarks

Cross-modal Retrieval on MIMIC-5×200. We evaluate text-to-image and image-to-text retrieval on MIMIC-5×200. As shown in Table 2, MedCTA achieves the best results on seven of eight metrics and ranks second on text-to-image P@2. Appendix A.5 provides qualitative image-to-text and text-to-image retrieval examples, with green and red boxes indicating correct and incorrect retrievals, respectively. These results demonstrate that combining static alignment with abnormality-conditioned trajectories and patient-course progression improves cross-modal representation learning.

Table 2: Cross-modal retrieval results on the MIMIC-CXR 5×200 benchmark. Performance is reported as precision (%) at top-k (P@k). Best results are shown in bold, and second-best results are underlined.

	Text → Image				Image → Text			
Method	**P@1**	**P@2**	**P@5**	**P@10**	**P@1**	**P@2**	**P@5**	**P@10**
MGCA	74.50	76.30	71.16	60.85	63.62	63.88	64.11	61.95
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	75.25	76.18	71.36	<u>65.04</u>	64.50	63.97	62.48	59.00
ALTA	69.80	65.85	52.60	37.72	56.80	55.40	46.80	46.80
TAMM	<u>78.69</u>	**79.26**	<u>73.07</u>	64.03	<u>66.44</u>	<u>67.63</u>	<u>66.05</u>	<u>62.50</u>
Ours	**80.24**	<u>79.08</u>	**75.68**	**65.82**	**69.13**	**68.63**	**67.23**	**63.01**

Zero-shot Classification Tasks. We evaluate the transferability of MedCTA on RSNA Pneumonia and COVIDx under a zero-shot setting. As shown in Table 3, MedCTA achieves the highest ACC and F1 on both datasets: 86.05% and 79.25% on RSNA Pneumonia, and 95.66% and 88.48% on COVIDx, respectively. It also achieves the best AUROC on RSNA Pneumonia (89.82%) and the second-best on COVIDx (96.92%). This indicates that modeling longitudinal clinical change enhances visual representation learning. By integrating static image–report alignment with abnormality-conditioned trajectories and patient-course progression, MedCTA learns more transferable visual features for zero-shot disease classification.

4.3 Ablation Study

Contribution of Different Components. We ablate $\mathcal{L}_{\text{clip}}$, $\mathcal{L}_{\text{abn}}$, and $\mathcal{L}_{\text{pat}}$ to assess the contribution of each objective. As shown in Table 4, removing any objective consistently reduces performance on temporal sentence similarity and temporal image classification, indicating their complementary

Table 3: Zero-shot classification results on the RSNA Pneumonia and COVIDx datasets. Performance is reported in terms of ACC (%), F1 (%), and AUROC (%). Best results are shown in bold, and second-best results are underlined.

	RSNA Pneumonia			COVIDx		
Method	**ACC**	**F1**	**AUROC**	**ACC**	**F1**	**AUROC**
MGCA	85.06 ± 0.05	76.56 ± 0.10	89.34 ± 0.14	92.13 ± 0.10	80.10 ± 0.65	88.16 ± 1.81
MedCLIP	84.17 ± 0.16	77.38 ± 0.35	89.01 ± 0.05	90.19 ± 0.09	75.60 ± 2.10	94.63 ± 0.17
PRIOR	81.30 ± 0.35	72.02 ± 1.09	82.80 ± 2.03	90.01 ± 0.19	87.97 ± 0.68	96.09 ± 0.15
MAVL	79.99 ± 0.83	73.14 ± 0.48	83.56 ± 0.49	88.77 ± 0.92	85.09 ± 1.71	95.52 ± 0.25
CARZero	84.28 ± 0.07	76.94 ± 0.40	88.68 ± 0.11	91.92 ± 0.20	85.83 ± 0.98	$\mathbf{97.64 \pm 0.60}$
Med-ST	85.15 ± 0.04	76.96 ± 0.25	88.99 ± 0.11	92.27 ± 0.17	80.22 ± 0.35	91.44 ± 0.57
ALTA	84.24 ± 0.20	76.54 ± 0.61	88.50 ± 0.26	91.53 ± 0.19	77.62 ± 1.26	94.55 ± 1.46
TAMM	85.48 ± 0.05	78.21 ± 0.14	89.36 ± 0.24	93.03 ± 0.06	87.32 ± 0.38	95.50 ± 0.58
Ours	$\mathbf{86.05 \pm 0.03}$	$\mathbf{79.25 \pm 0.08}$	$\mathbf{89.82 \pm 0.27}$	$\mathbf{95.66 \pm 0.12}$	$\mathbf{88.48 \pm 0.16}$	96.92 ± 0.61

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.

Method	**Temporal Sentence Similarity**	**Temporal Image Classification**				
		Consolidation	Edema	Pleural Effusion	Pneumonia	Pneumothorax
Ours	**95.72**	**67.14**	**70.30**	**65.59**	**72.85**	**61.66**
w/o clip loss	92.24 (−3.48)	59.58 (−7.56)	59.51 (−10.79)	63.44 (−2.15)	67.50 (−5.35)	56.75 (−4.91)
w/o abnormality loss	87.47 (−8.25)	55.22 (−11.92)	51.46 (−18.84)	58.41 (−7.18)	65.24 (−7.61)	53.33 (−8.33)
w/o patient loss	90.55 (−5.17)	56.98 (−10.16)	57.66 (−12.64)	59.58 (−6.01)	65.80 (−7.05)	54.18 (−7.48)

effects. The full model performs best, demonstrating the effectiveness of combining static alignment with dual-scope clinical change supervision. More detailed ablations are reported in Appendix B.1.

Effect of Query Initialization. We compare random and clinically initialized abnormality queries in Appendix B.2. Clinical initialization consistently performs better, indicating that medical prompts provide useful semantic priors for constructing abnormality-specific trajectories.

Effect of Different LLM Parsers. MedCTA uses Qwen2-7B Yang et al. (2024a) as the offline parser for abnormality-specific and patient-course trend supervision. To assess parser robustness, we replace it with Llama 3-8B Grattafiori et al. (2024). Appendix B.3 shows stable image-text retrieval performance, suggesting that MedCTA's gains come from structured temporal supervision that transforms implicit longitudinal semantics into learnable clinical change signals, rather than from a specific LLM backbone.

5 Conclusion

We introduced MedCTA, a clinical trajectory alignment framework that extends medical vision–language pre-training from visit-level image–report matching to longitudinal clinical change learning. MedCTA models change at two complementary scopes: abnormality-specific trajectories capture heterogeneous abnormality evolution, while patient-course progression captures overall clinical status. With structured trend supervision extracted from longitudinal reports by an offline LLM parser, MedCTA enables trajectory-level learning without manual temporal annotations while preserving static image–report alignment. Experiments on temporal image classification, image–text retrieval, and zero-shot classification show consistent gains over strong medical VLP baselines, demonstrating the value of modeling clinical change through both abnormality-specific evolution and patient-course progression.

AI USE STATEMENT

We employed LLMs for two purposes. First, they supported manuscript preparation, including translation, linguistic refinement, structural organization, and LaTeX typesetting. Second, an LLM served as an offline annotation tool that derived abnormality-specific and patient-course trend labels from longitudinal radiology reports; the corresponding extraction protocol and quality-control procedures are detailed in the manuscript. No LLM was involved in implementing the proposed method or developing mathematical derivations or proofs. The authors inspected and validated all AI-assisted material and extracted annotations and assume full responsibility for the final manuscript, scientific claims, and reported experimental results.

REFERENCES

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pp. 72–78, 2019.

Liang Bai, Zhi Wang, Huimin Yan, and Xian Yang. Medical vision–language pretraining with llm-guided temporal supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 19666–19674, 2026.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15016–15027, 2023.

Christian Bluethgen, Pierre Chambon, Jean-Benoit Delbrouck, Rogier Van Der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P Langlotz, and Akshay S Chaudhari. A vision–language foundation model for the generation of realistic chest x-ray images. *Nature Biomedical Engineering*, 9(4):494–506, 2025.

Weiwei Cao, Jianpeng Zhang, Zhongyi Shui, Sinuo Wang, Zeli Chen, Xi Li, Le Lu, Xianghua Ye, Qi Zhang, Tingbo Liang, et al. Boosting vision semantic density with anatomy normality modeling for medical vision-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23041–23050, 2025.

Yanyuan Chen, Dexuan Xu, Yu Huang, Songkun Zhan, Hanpin Wang, Dongxue Chen, Xueping Wang, Meikang Qiu, and Hang Li. Mimo: A medical vision language model with visual referring multimodal input and pixel grounding multimodal output. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24732–24741, 2025.

Pujin Cheng, Li Lin, Junyan Lyu, Yijin Huang, Wenhan Luo, and Xiaoying Tang. Prior: Prototype representation joint learning from medical images and reports. In *2023 IEEE/CVF International Conference on Computer Vision*, pp. 21304–21314. IEEE, 2023.

Hejie Cui, Alyssa Unell, Bowen Chen, Jason Alan Fries, Emily Alsentzer, Sanmi Koyejo, and Nigam H Shah. Timer: Temporal instruction modeling and evaluation for longitudinal clinical records. *npj Digital Medicine*, 8(1):577, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations*, 2021.

Wei Fan, Jingru Fei, Jindong Han, Jie Lian, Hangting Ye, Xiaozhuang Song, Xin Lv, Kun Yi, and Min Li. Medvia: Empowering medical time series classification with vision augmentation and multimodal fusion. *Information Fusion*, 127:103659, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

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

Haoran Lai, Qingsong Yao, Zihang Jiang, Rongsheng Wang, Zhiyang He, Xiaodong Tao, and S. Kevin Zhou. Carzero: Cross-attention alignment for radiology zero-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11137–11146, 2024.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *IEEE Transactions on Medical Imaging*, 45(6):2727–2737, 2026.

Tianbin Li, Yanzhou Su, Wei Li, Bin Fu, Zhe Chen, Ziyan Huang, Guoan Wang, Chenglong Ma, Ying Chen, Ming Hu, et al. Gmai-vl & gmai-vl-5.5 m: A large vision-language model and a comprehensive multimodal dataset towards general medical ai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23177–23185, 2026.

Chenyu Lian, Hong-Yu Zhou, Dongyun Liang, Jing Qin, and Liansheng Wang. Efficient medical vision-language alignment through adapting masked vision models. *IEEE Transactions on Medical Imaging*, 44(11):4499–4510, 2025.

Kang Liu, Zhuoqi Ma, Xiaolu Kang, Yunan Li, Kun Xie, Zhicheng Jiao, and Qiguang Miao. Enhanced contrastive learning with multi-view longitudinal data for chest x-ray report generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10348–10359, 2025.

Kang Liu, Zhuoqi Ma, Zikang Fang, Yunan Li, Kun Xie, and Qiguang Miao. Priorrg: Prior-guided contrastive pre-training and coarse-to-fine decoding for chest x-ray report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7206–7214, 2026.

Xin Mei, Rui Mao, Xiaoyan Cai, Libin Yang, and Erik Cambria. Medical report generation via multimodal spatio-temporal fusion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 4699–4708, 2024.

Pingyi Miao, Xianlai Chen, Kai Sun, Yunbo Wang, Shuang Zhao, and Ying An. Biodpp: Dynamic prompt policy learning for biomedical vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8052–8060, 2026.

Vishwesh Nath, Wenqi Li, Dong Yang, Andriy Myronenko, Mingxin Zheng, Yao Lu, Zhijian Liu, Hongxu Yin, Yee Man Law, Yucheng Tang, et al. Vila-m3: Enhancing vision-language models with medical expert knowledge. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14788–14798, 2025.

Vu Minh Hieu Phan, Yutong Xie, Yuankai Qi, Lingqiao Liu, Liyang Liu, Bowen Zhang, Zhibin Liao, Qi Wu, Minh-Son To, and Johan W Verjans. Decomposing disease descriptions for enhanced pathology detection: A multi-aspect vision-language pre-training framework. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11492–11501. IEEE, 2024.

George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019.

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in neural information processing systems*, 35:33536–33549, 2022a.

Linda Wang, Zhong Qiu Lin, and Alexander Wong. 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, 2020.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 3876–3887, 2022b.

Joy Wu, Nkechinyere Agu, Ismini Lourentzou, Arjun Sharma, Joseph Paguio, Jasper Seth Yao, Edward Christopher Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. Chest imagenome dataset. *Physio Net*, 2021.

Huimin Yan, Xian Yang, Liang Bai, and Jiye Liang. Local alignment for medical vision-language pre-training. *IEEE Transactions on Image Processing*, 34:7321–7334, 2025.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024a.

Hao Yang, Hong-Yu Zhou, Jiarun Liu, Weijian Huang, Cheng Li, Zhihuan Li, Yuanxu Gao, Qiegen Liu, Yong Liang, Qi Yang, et al. A multimodal vision–language model for generalizable annotation-free pathology localization. *Nature Biomedical Engineering*, 10:1595–1609, 2026.

Jinxia Yang, Bing Su, Xin Zhao, and Ji-Rong Wen. Unlocking the power of spatial and temporal information in medical multimodal pre-training. In *Forty-first International Conference on Machine Learning*, pp. 56382–56396, 2024b.

Zhuoyi Yang and Liyue Shen. Tempa-vlp: Temporal-aware vision-language pretraining for longitudinal exploration in chest x-ray image. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4625–4634. IEEE, 2025.

Xin You, Runze Yang, Chuyan Zhang, Zhongliang Jiang, Jie Yang, and Nassir Navab. Fb-diff: Fourier basis-guided diffusion for temporal interpolation of 4d medical imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 28010–28020, 2025.

Ke Zhang, Hanliang Jiang, Jian Zhang, Qingming Huang, Jianping Fan, Jun Yu, and Weidong Han. Semi-supervised medical report generation via graph-guided hybrid feature consistency. *IEEE Transactions on Multimedia*, 26:904–915, 2024.

Appendix

A Additional Experimental Details

A.1 Pretraining Data

We pretrain MedCTA on MIMIC-CXR Johnson et al. (2019), which contains chest X-ray images and corresponding radiology reports. For longitudinal modeling, we reorganize the data into patient-level sequences ordered by study time. Following prior temporal medical VLP settings Yang et al. (2024b), each sequence includes up to four consecutive image–report pairs, while shorter sequences are preserved. The dataset is publicly available at `https://physionet.org/content/mimic-cxr-jpg/2.1.0/`.

A.2 Downstream Tasks

We evaluate MedCTA on downstream benchmarks disjoint from the pre-training data, covering both longitudinal temporal reasoning and cross-modal understanding.

Longitudinal temporal reasoning: We evaluate three tasks. *Temporal image classification* is fine-tuned on Chest ImaGenome Wu et al. (2021) and tested on MS-CXR-T Bannur et al. (2023), reporting macro-averaged accuracy over progression types. *Temporal sentence similarity classification* is evaluated as a zero-shot binary classification on MS-CXR-T sentence pairs, reporting accuracy and AUROC. *Dynamic phrase grounding* is assessed through qualitative visualization on Chest ImaGenome, where we examine whether the model localizes image regions relevant to temporal progression descriptions across consecutive chest X-rays.

Cross-modal understanding: We evaluate *cross-modal retrieval* on MIMIC-5×200 Wang et al. (2022b), reporting P@1, P@2, P@5, and P@10 for both image-to-text and text-to-image retrieval. We also assess *zero-shot classification* on COVIDx Wang et al. (2020) and RSNA Pneumonia Shih et al. (2019) using accuracy, F1, and AUROC.

A.3 Baselines

We compare MedCTA with recent representative medical vision–language methods, grouped according to whether they explicitly model temporal information. *Non-temporal vision–language baselines* focus on static image–text alignment: MGCA Wang et al. (2022a) aligns radiographs and reports at multiple granularities; MedCLIP Wang et al. (2022b) improves contrastive learning with semantic-aware matching; CARZero Lai et al. (2024) leverages LLM-based prompts and cross-attention for zero-shot classification; PRIOR Cheng et al. (2023) combines cross-modal reconstruction with global-local supervision; and MAVL Phan et al. (2024) uses LLM-guided disease decomposition to enhance attribute-level grounding. In contrast, *temporal baselines* explicitly incorporate longitudinal context: Med-ST Yang et al. (2024b) introduces temporal supervision through historical image–text pairs and cycle consistency, while ALTA Lian et al. (2025) models temporal views with masked pretraining and efficient adaptation. TAMM Bai et al. (2026) leverages LLM-derived trend labels and rationales from adjacent reports to inject temporal supervision into medical vision-language pretraining.

A.4 Implementation Details

We initialize MedCTA with pretrained MGCA weights to improve optimization stability and accelerate convergence. The model is trained for 8 epochs on 4 NVIDIA A40 GPUs. We use the AdamW optimizer with a weight decay of 1×10^{-6} and an initial learning rate of 1×10^{-5}. A cosine annealing schedule with 40% warm-up is adopted, gradually decaying 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 evaluation, we either use the pretrained model in a zero-shot manner or fine-tune a lightweight task-specific head, depending on the benchmark. For cross-modal retrieval, temporal sentence similarity classification, and dynamic phrase grounding, we directly evaluate the pretrained model without additional training, and thus do not report standard deviations. In contrast, zero-shot classification and temporal image classification involve training a lightweight classifier on top of pretrained representations. For these tasks, we run experiments with three different random seeds and report the mean and standard deviation to reflect training variability. This evaluation protocol follows prior work for fair comparison.

A.5 Qualitative Cross-Modal Retrieval Results

Figure 5 presents qualitative image-to-text and text-to-image retrieval examples on MIMIC-5×200, comparing MedCTA with Med-ST, ALTA, and TAMM. For each method, we display the top two retrieved results for the same query. Green and red boxes indicate correct and incorrect abnormality-category matches, respectively. For image-to-text retrieval, MedCTA returns two descriptions of left basilar atelectasis, whereas the baselines retrieve at least one description of another abnormality. For text-to-image retrieval, both MedCTA and TAMM return two correct matches for the cardiomegaly query. These examples illustrate MedCTA's cross-modal semantic alignment.

B Ablation experimental results

In this section, we provide additional ablation results to further analyze the contribution of each component in MedCTA, the effect of abnormality query initialization, and the robustness to different LLM-based trend supervision extractors. We further perturb visit order to assess whether MedCTA captures order-dependent temporal dynamics.

B.1 Contribution of Different Components

To further validate the effectiveness of the three training objectives, we report additional ablation results on image–text retrieval. Specifically, we remove the static image–report alignment loss $\mathcal{L}_{\text{clip}}$,

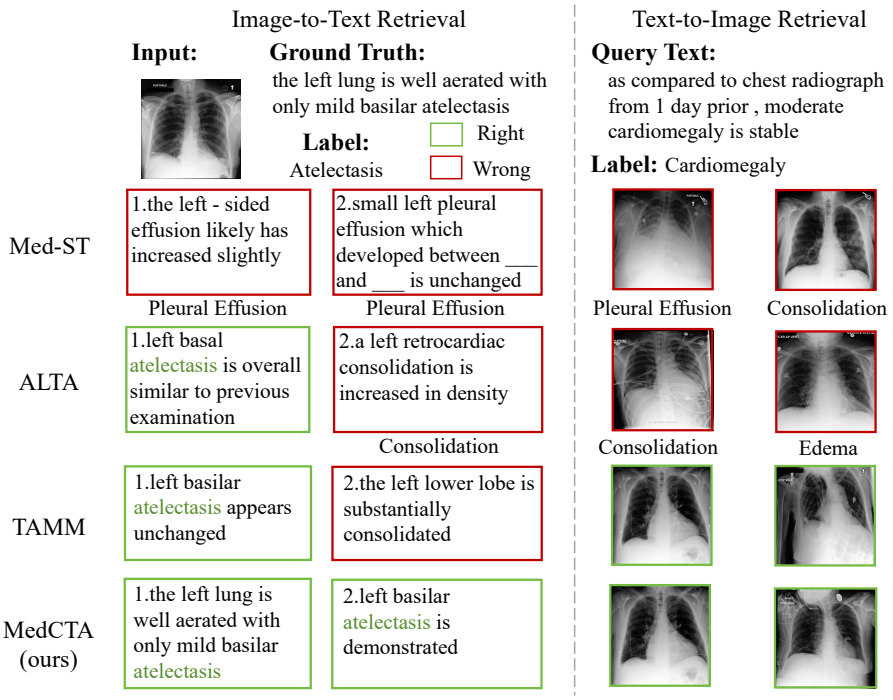


Figure 5: Qualitative examples of image-to-text and text-to-image retrieval on MIMIC-5×200. For each method, we show the top two retrieved samples given an input chest X-ray or a query sentence. Green and red boxes denote correct and incorrect retrievals, respectively, with abnormality categories annotated below the incorrect samples.

Table 5: Additional ablation results on image–text retrieval on the MIMIC-CXR 5×200 benchmark. Performance is reported as precision (%) at top-k (P@k). Relative changes from the full model are shown below.

Method	**Text → Image**				**Image → Text**			
	P@1	**P@2**	**P@5**	**P@10**	**P@1**	**P@2**	**P@5**	**P@10**
MedCTA	**80.24**	**79.08**	**75.68**	**65.82**	**69.13**	**68.63**	**67.23**	**63.01**
w/o $\mathcal{L}_{\text{clip}}$	58.05 (−22.19)	55.35 (−23.73)	54.32 (−21.36)	46.06 (−19.76)	53.22 (−15.91)	53.19 (−15.44)	53.98 (−13.25)	48.72 (−14.29)
w/o $\mathcal{L}_{\text{abn}}$	77.67 (−2.57)	76.59 (−2.49)	73.45 (−2.23)	63.93 (−1.89)	67.51 (−1.62)	66.36 (−2.27)	64.75 (−2.48)	61.28 (−1.73)
w/o $\mathcal{L}_{\text{pat}}$	78.90 (−1.34)	78.02 (−1.06)	74.01 (−1.67)	65.00 (−0.82)	68.21 (−0.92)	68.67 (+0.04)	66.51 (−0.72)	61.80 (−1.21)

the abnormality-specific trend loss $\mathcal{L}_{\text{abn}}$, and the patient-course trend loss $\mathcal{L}_{\text{pat}}$ from the full objective, respectively. As shown in Table 5, removing each objective generally degrades retrieval performance across most metrics, indicating that the three objectives provide complementary supervision. Removing $\mathcal{L}_{\text{clip}}$ causes the largest drop, suggesting that visit-level image–report alignment remains essential for cross-modal retrieval. Removing $\mathcal{L}_{\text{abn}}$ also reduces performance, reflecting the benefit of modeling fine-grained abnormality evolution. Removing $\mathcal{L}_{\text{pat}}$ leads to smaller but mostly consistent declines, suggesting that patient-course progression modeling further improves the learned representations. These retrieval results are broadly consistent with the temporal reasoning results in the main paper and support the complementary roles of the three objectives.

B.2 Effect of Query Initialization

We further investigate whether the abnormality queries benefit from clinical semantic initialization. We compare two strategies: random initialization and clinical semantic initialization using

Table 6: Effect of abnormality query initialization on image–text retrieval on the MIMIC-CXR 5×200 benchmark. Performance is reported as precision at top-k (P@k).

Method	Text → Image				Image → Text			
	P@1	P@2	P@5	P@10	P@1	P@2	P@5	P@10
Random init	78.37	77.69	72.94	65.32	68.27	65.42	68.15	61.82
Clinical init	80.24	79.08	75.68	65.82	69.13	68.63	67.23	63.01

abnormality-specific medical prompts. As shown in Table 6, clinical semantic initialization improves most retrieval metrics compared with random initialization. This suggests that abnormality prompts provide useful semantic priors, helping the queries select abnormality-relevant visual and textual evidence and construct more effective abnormality-specific progression trajectories.

B.3 Effect of Different LLM Parsers

MedCTA uses an offline LLM parser to extract structured trend supervision from longitudinal reports. To examine whether the performance depends on a specific LLM backbone, we replace Qwen2-7B Yang et al. (2024a) with Llama 3-8B Grattafiori et al. (2024) while keeping the remaining training and evaluation settings unchanged. As shown in Table 7, MedCTA maintains stable performance under different LLM settings on image–report retrieval. These results indicate that MedCTA does not rely on a particular LLM backbone. Instead, its gains mainly come from the structured temporal supervision that converts implicit longitudinal report semantics into learnable clinical change signals.

Table 7: Robustness to different LLM parsers on image–text retrieval on the MIMIC-CXR 5×200 benchmark. Performance is reported as precision at top-k (P@k).

Method	Text → Image				Image → Text			
	P@1	P@2	P@5	P@10	P@1	P@2	P@5	P@10
Qwen2-7B	80.24	79.08	75.68	65.82	69.13	68.63	67.23	63.01
Llama 3-8B	80.11	78.92	76.89	66.09	69.56	68.54	66.85	63.14

B.4 Effect of Visit-Order Perturbation

To investigate whether the Mamba-based temporal encoder benefits from order-dependent temporal dynamics, rather than merely aggregating visit features through order-invariant pooling, we conduct a visit-order perturbation ablation. Specifically, we randomly shuffle the visit order within each patient sequence while preserving its set of visits and sequence-level trend label. We then repeat pre-training under the same experimental settings as the chronological-order baseline.

As shown in Tables 8 and 9, shuffling visits reduces text-to-image retrieval P@1 from 80.24% to 75.99% and image-to-text retrieval P@1 from 69.13% to 65.88%. On temporal sentence similarity classification, ACC decreases from 95.72% to 89.73%, while AUROC decreases from 98.12% to 93.20%. These results show that preserving the visit set alone is insufficient to maintain performance when chronological order is disrupted. They support the interpretation that the Mamba-based temporal encoder exploits chronological structure under relatively coarse sequence-level trend supervision, rather than functioning solely as an order-invariant pooling operator.

C Reliability of LLM-Generated Trend Supervision

MedCTA uses an offline LLM parser to construct structured trend supervision from longitudinal radiology reports. Since the resulting labels may contain errors, we examine their reliability through temporal-reversal consistency filtering and controlled label-noise experiments. We also present a qualitative failure case illustrating the challenges of assigning a single trend label to a non-monotonic patient course.

Table 8: Cross-modal retrieval results under visit-order perturbation on the MIMIC-CXR 5×200 benchmark. Performance is reported as precision (%) at top-k. The chronological-order baseline is denoted as MedCTA.

	Text → Image				Image → Text			
Method	P@1	P@2	P@5	P@10	P@1	P@2	P@5	P@10
MedCTA	**80.24**	**79.08**	**75.68**	**65.82**	**69.13**	**68.63**	**67.23**	**63.01**
Randomized order	75.99	74.65	72.58	61.56	65.88	65.59	65.16	62.49

Table 9: Temporal sentence similarity classification results under visit-order perturbation on MS-CXR-T. Performance is reported as ACC (%) and AUROC (%). The chronological-order baseline is denoted as MedCTA.

Method	ACC (%)	AUROC (%)
MedCTA	**95.72**	**98.12**
Randomized order	89.73	93.20

C.1 Temporal-Reversal Consistency Filtering

To reduce label noise, we apply a temporal-reversal consistency filtering strategy. Given a chronologically ordered report sequence $\mathcal{R}_i = (R_{i,1}, \ldots, R_{i,T_i})$, we construct its reversed sequence $\mathcal{R}_i^{\text{rev}} = (R_{i,T_i}, \ldots, R_{i,1})$. The LLM parser predicts a trend label for each ordering using the corresponding trend-labeling prompt.

For labels encoded as -1, 0, and 1 for worsening, stable, and improving, respectively, the consistency criterion is

$$\hat{y}_i^{\text{rev}} = -\hat{y}_i^{\text{fwd}}, \tag{19}$$

where $\hat{y}_i^{\text{fwd}}$ and $\hat{y}_i^{\text{rev}}$ denote predictions from the chronological and reversed sequences. Thus, an improving prediction should become worsening after reversal, a worsening prediction should become improving, and a stable prediction should remain stable. This procedure removes approximately 6% of patient sequences with directionally inconsistent trend labels. The retained sequences constitute the final pre-training dataset used throughout this study. All reported MedCTA results are obtained using models pre-trained on this filtered dataset.

C.2 Robustness to Synthetic Trend-Label Noise

We evaluate the sensitivity of MedCTA to erroneous trend supervision by randomly replacing 1%, 5%, and 10% of the valid abnormality-specific and patient-course labels with incorrect classes. The 0% setting denotes the baseline without additional synthetic label corruption.

Tables 10 and 11 report temporal image classification and cross-modal retrieval performance, respectively. Performance decreases gradually as the proportion of corrupted labels increases. At 10% added label noise, mean accuracy across the five temporal image classification tasks decreases from 67.51% to 65.45%, a reduction of 2.06 percentage points. The mean across the eight retrieval metrics decreases from 71.10% to 70.48%, a reduction of 0.62 percentage points. These results suggest that the model exhibits a certain degree of robustness to mild random label noise.

C.3 Qualitative Failure Analysis

Our manual evaluation identified failure cases in which the LLM-generated abnormality-specific trend supervision could provide misleading labels, particularly for non-monotonic disease trajectories. A representative example is shown Table 12. During the earlier visits, the patient exhibited enlarging bilateral pleural effusions and pulmonary edema. At the final visit, however, the report described improved lung volumes, more uniformly distributed mild pulmonary edema, small bilateral pleural effusions, and improved cardiomegaly. The LLM nevertheless labeled the sequence as *worsening*, whereas the human annotators classified the overall trajectory as *improving*. This case

Table 10: Temporal image classification under synthetic trend-label noise. Accuracy (%) is reported for five abnormality categories at different levels of randomly injected label noise. Average denotes the unweighted mean accuracy across the five categories.

Added noise	Consolidation	Edema	Pl. Effusion	Pneumonia	Pneumothorax	Average
0% (MedCTA)	67.14	70.30	65.59	72.85	61.66	**67.51**
1%	67.11	70.22	65.49	72.79	61.54	67.43
5%	66.23	69.68	64.76	71.84	60.52	66.61
10%	65.33	68.37	63.80	70.54	59.23	65.45

Table 11: Cross-modal retrieval under synthetic trend-label noise on the MIMIC-CXR 5×200 benchmark. Performance is reported as precision (%) at top-k. Average denotes the unweighted mean of P@1, P@2, P@5, and P@10 across both text-to-image and image-to-text retrieval.

	Text → Image				Image → Text				
Added noise	P@1	P@2	P@5	P@10	P@1	P@2	P@5	P@10	Average
0% (MedCTA)	80.24	79.08	75.68	65.82	69.13	68.63	67.23	63.01	**71.10**
1%	80.15	79.03	75.61	65.72	69.05	68.60	67.17	62.86	71.02
5%	79.97	78.67	75.63	64.84	68.85	68.17	67.11	62.57	70.73
10%	79.41	78.63	75.30	64.76	68.59	67.92	66.90	62.35	70.48

suggests that the parser may overemphasize repeated evidence of deterioration in earlier reports while underweighting subsequent recovery. This case demonstrates the difficulty of compressing a complex, non-monotonic clinical course into a single ternary trend label. In particular, deterioration described across multiple earlier reports may dominate the final prediction, even when the latest examination indicates improvement.

Table 12: A representative failure case of LLM-generated patient-course trend supervision. Reports are shown in chronological order with their original wording preserved.

Visit	**Radiology report**
Visit 1	**Worsening moderate bilateral pleural effusions.** A nasogastric tube terminates within the stomach. Moderate cardiomegaly is unchanged. **Moderate bilateral pleural effusions have enlarged since the previous examination.** Underlying bibasilar atelectasis appears unchanged. There is no pneumothorax.
Visit 2	There is marked cardiomegaly with upper-zone redistribution and interstitial and pulmonary edema. There is more confluent opacity at the right and left lung bases, consistent with collapse and/or consolidation. **No gross effusion is observed, although small effusions may be present.** The mediastinal contours and hila appear prominent on this examination.
Visit 3	**Worsening bilateral pleural effusions with unchanged moderate edema.** Superimposed consolidation, particularly at the lung bases, cannot be excluded. The heart is mildly enlarged. **There is worsening of the bilateral pleural effusions, as well as central vascular congestion.** Moderate pulmonary edema is again observed. There is no large pneumothorax, although the apices are obscured by the patient's head.
Visit 4 (final)	AP chest radiograph compared with the previous examination. A new endotracheal tube is in the standard position. The nasogastric tube passes below the diaphragm and out of view. **Lung volumes have improved, and mild pulmonary edema is more uniformly distributed. Small bilateral pleural effusions are presumed. Mild-to-moderate cardiomegaly has improved,** possibly because of positive-pressure ventilatory support. No pneumothorax is present.
LLM label: Worsening	**Human label:** Improving