Temporal Inversion for Learning Interval Change in Chest X-Rays

Supplementary Material

6. Model

This section provides additional details on implementation, augmentations, hyperparameters, and model selection.

6.1. Implementation Details

For pretraining, we use the AdamW optimizer with a cosine learning-rate schedule and a 100-step warm-up. The learning rate is set to 1×10^{-4} with a batch size of 144, and all parameters are trained in `bfloat16`. Pretraining is conducted on three NVIDIA A6000 GPUs for 30 epochs (approximately 54 GPU-hours).

Fine-tuning also uses AdamW with cosine scheduling and a 5% warm-up. Models are trained end-to-end for 50 epochs using a learning rate of 1×10^{-5} and a batch size of 128. These experiments are performed on a single NVIDIA A6000 GPU. Projection layers for both image and text encoders have dimension 128, and the CXR-BERT text encoder is configured with a maximum token length of 256.

6.2. Augmentation

Image augmentations follow the protocols described in BioViL-T [1] and ALTA [13] and are applied consistently across both pretraining and fine-tuning.

6.3. Hyperparameters

Logit scales τ and τ^{swap} are initialized to $\log 10$, with bias -10 (SigLIP's logit-shift parameter). We set $W = 1$ for the Change-aware Sigmoid loss during pretraining and choose $\lambda = 50$ for fine-tuning to balance the magnitude of the Temporal Consistency Loss relative to the cross-entropy term. The Change-aware Sigmoid loss is activated after 10 pretraining epochs, and TCL is applied after 20 fine-tuning epochs to prevent premature convergence toward predicting `stable` for most pairs.

6.4. Compared Models

We compare only to models with publicly available code to ensure consistent training and reproducible evaluation. The CNN+Transformer baseline follows the BioViL-T [1] implementation with ImageNet initialization. Methods without publicly available code or reproducible training pipelines were excluded.

Temporal CXR baselines remain limited relative to single-image CXR methods. Few existing approaches explicitly model interval change, and the field lacks standardized benchmarks, unified data splits, and publicly released training workflows. To avoid data leakage and ensure fair comparison, we include only models we can reproduce end-to-end and evaluate under our inversion-aware protocol. Where available, we also include models with released pretrained weights, even if full training code is not provided. Recent studies similarly compare against a narrow set of baselines and often rely on reported numbers rather than independent reproduction. Given our focus on temporal robustness—not only raw accuracy—and the need to strictly avoid any MS-CXR-T overlap, controlled and fully reproducible evaluation was essential.

7. Data

This section describes dataset splits, label generation protocols, and ethical considerations, including specific details on the **MS-CXR-T$_{\text{retrieval}}$** benchmark.

7.1. Dataset Splits

We first exclude chest X-ray images without available prior images. The sample counts for each split are presented in Tab. 5. For CheXpert, since the official validation and test splits lack corresponding reports or prior images, we use the CheXpert training split as an external pool for retrieval evaluation. To minimize sampling bias, we repeatedly sample 3,000 image pairs from this pool 10 times and report the mean and 95% confidence interval across these subsamples. For our private dataset, we collected CXR pairs from 2010–2020, filtering for reports and images containing temporal keywords (e.g., "improved", "worsened", "stable"). Labeling was performed by a researcher with four years of experience in chest X-ray interpretation, and these labels were used as the reference for external validation during fine-tuning.

	Train	Validation	Test
MIMIC	183,302	1,330	2,871
CheXpert	-	-	123,374
Private	-	-	2,233

Table 5. Data Distribution for MIMIC, CheXpert, and Private

7.2. Dataset Approvals and Ethics

All LLM-assisted label extraction for MIMIC and the construction of MS-CXR-T$_{\text{retrieval}}$ were conducted in accordance with PhysioNet guidelines for responsible LLM usage (https://physionet.org/news/post/gpt-responsible-use). All preprocessing for MS-CXR-T$_{\text{retrieval}}$ was performed *prior to September 2025*, during a period in which de-

identified report processing on Google Cloud Vertex AI was permitted under the platform's data-handling policies.

The private clinical dataset was collected under institutional IRB approval, and all researchers accessing the data were formally registered, credentialed, and authorized for handling de-identified records.

7.3. Change Label Generation

Change/no-change labels for pretraining were generated using Gemini 2.0 Flash. The model compares each follow-up report with its corresponding prior (when available) and identifies interval changes, including *improved/stable/worsened* and *new/resolved* findings. For compatibility with the Change-aware Sigmoid Loss, we invert the binary output (1→0, 0→1), and exclude uncertain cases ("-1") from training.

Sampled-scale validation of LLM choices. To verify that Gemini is not strictly required and to assess the reliability of alternative LLMs, we conducted a small-scale validation on 200 randomly sampled report pairs. For each pair, we validated the LLM-derived change/no-change label against a manually reviewed label provided by a trained annotator.

Agreement rates (with 95% Wilson confidence intervals) were:

- **Gemini 2.0 Flash:** 192/200 (96.0%; 92.3–97.9)
- **Qwen3-14B:** 179/200 (89.5%; 84.5–93.1)

These results show that smaller LLMs can also produce reasonable labels, although Gemini exhibited the highest agreement in this pilot study. The prompt used for label generation is provided below:

> You are given two chest X-ray (CXR) reports: a **previous** CXR report and a **follow-up** CXR report. Your task is to analyze both reports and determine if there are any changes in the follow-up CXR compared to the previous one.
>
> ### Instructions: 1. Compare the findings in both reports. 2. If there are **any new, worsening, or improving conditions**, return '1'. 3. If the reports state **"no interval change"** or findings are **stable**, return '0'. 4. If unsure, return '-1'. 5. Ensure the output is strictly '0', '1', '-1' without additional text.
>
> ### Input Format: - **Previous:** [Insert previous report text] - **Follow-up:** [Insert follow-up report text]
>
> ### Output Format: Return only: - '0' if no changes are detected. - '1' if any changes (new, improved, or worsened findings) are detected. - '-1' if uncertain, or any of the previous report or followup report is not a chest X-ray report.
>
> ### Few-shot Examples:
>
> #### **Example 1: No Changes (Output: 0)** - **Previous:** "Left pleural effusion is noted. The cardiac silhouette is normal. No acute abnormalities." - **Follow-up:** "No interval change." - **Output:** '0'
>
> #### **Example 2: New Finding (Output: 1)** - **Previous:** "The lungs are clear. No pleural effusion or pneumothorax. No focal consolidation. The cardiac silhouette is normal. No acute abnormalities." - **Follow-up:** "A new left lower lobe consolidation is noted, concerning for pneumonia. No pleural effusion or pneumothorax. The cardiac silhouette is normal." - **Output:** '1'
>
> #### **Example 3: Improvement in Findings (Output: 1)** - **Previous:** "Patchy bilateral infiltrates consistent with pneumonia. No pleural effusion. The heart size is within normal limits." - **Follow-up:** "Bilateral infiltrates have significantly improved. No pleural effusion. The heart size remains within normal limits. - **Output:** '1'
>
> #### **Example 4: Stable Findings (Output: 0)** - **Previous:** "Mild left basilar atelectasis. No pneumothorax or pleural effusion. No acute cardiopulmonary abnormalities." - **Follow-up:** "Mild left basilar atelectasis remains unchanged. No pneumothorax or pleural effusion." - **Output:** '0'

7.4. MS-CXR-$T_{\text{retrieval}}$

We describe here the construction process of **MS-CXR-$T_{\text{retrieval}}$**, our benchmark for evaluating temporal reasoning in CXR report retrieval. The creation pipeline consists of three main stages:

Stage 1: Splitting Reports. We first extract the findings section and the corresponding report for each image. Reports are split so that each sentence describes a single radiological finding. This reduces noise in later stages when manipulating progression labels. The prompt used for this sentence-level splitting is as follows:

> Analyze the given radiology report text and split it into sentences, each describing a single radiological finding or view position information. Include both positive and negative findings, as well as view position details, but exclude other non-finding information. Follow these guidelines:
>
> 1. Each sentence should contain only one of the following: (exclude the sentence with view position information) a) A clear radiological finding b) The absence of a specific condition (negative finding)
>
> 2. Treat each negative finding (absence of a condition) as a separate observation and split it into its own sentence.
>
> 3. Keep sentences that describe view position information, but separate them from findings if they appear in the same original sentence.
>
> 4. Exclude sentences that do not describe actual radiological findings or view positions, such as: - Procedural details - General comments about image quality - Patient positioning information (unless it's specifically about the view position)
>
> 5. Maintain the meaning and context of the original findings and view positions while splitting.
>
> 6. Minor sentence structure changes or addition of necessary words are allowed to ensure clarity.
>
> 7. Remove any redundant information and express each finding or view position concisely.
>
> 8. Each split sentence should be understandable independently.
>
> 9. Avoid using lists or enumerations within a single sentence; instead, create separate sentences for each item.
>
> Example of splitting, including view position, and excluding non-findings: Original Report: '1. A single frontal view of the chest is provided. 2. No consolidation, pleural effusion, or pneumothorax is observed in both lungs. 3. The heart size is normal.' Output: 'No consolidation is observed in both lungs. No pleural effusion is observed in both lungs. No pneumothorax is observed in both lungs. The heart size is normal.'
>
> Process the input text according to these guidelines and return the relevant radiological findings and view position information and do not attach any additional text except the split sentences.

Stage 2: Prior Reference Omission. We remove all references to prior studies from each finding. This neutralizes progression status for unrelated findings and allows our benchmark to focus evaluation on the target finding(s) defined in MS-CXR-T. This omission step also enables evaluation in reversed or swapped settings. The prompt used for removing prior references is:

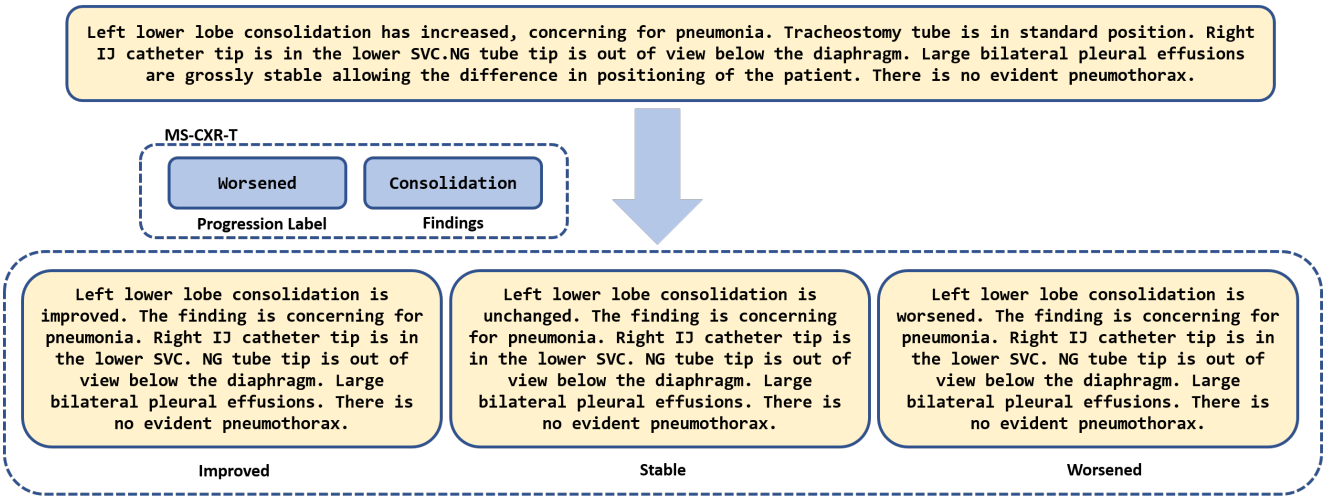

Figure 3. Example workflow for constructing MS-CXR-$T_{\text{retrieval}}$ from the original benchmark.

You are an expert chest X-ray (CXR) radiologist familiar with radiologic reports. Your task is to rewrite the given radiology reports by removing all references to prior reports or comparisons, while preserving the original structure as much as possible. Input: A radiology report for a chest X-ray (CXR). Output: A revised CXR report focusing solely on current medical findings, excluding references to prior reports, comparisons, and irrelevant details. Guidelines: Remove Comparisons: Eliminate any terms or phrases that suggest a comparison, such as "compared to," "in comparison with," "change", "cleared", "constant", "decrease", "elevate", "expand", "improve", "decrease", "increase", "persistent", "reduce", "remove", "resolve", "stable", "worse", "new", etc. Focus on Current Findings: Ensure the report only describes the current state of the patient's lungs and related structures. Preserve Medical Context: Maintain the original medical terminology and descriptions of abnormalities. Retain Negations: Keep any negative statements about the absence of abnormalities.

Example: Original Report: The left apex has not been included on this radiograph. The ET tube terminates 3.9 cm above the carina. The NG tube terminates in the stomach. Surgical clips and a faint metallic coil project over the chest. A left PICC terminates in the mid SVC. EKG leads overlie the chest wall. The lung volumes are low. There are persistent bilateral mid and lower zone hazy opacities. There are persistent bilateral hilar and perihilar linear opacities. No significant interval change is observed in the lung opacities. Bilateral pleural effusions are present. The right pleural effusion is greater than the left. No pneumothorax is observed on the right. No cardiomegaly is present. No interval change is observed in the mediastinal silhouette. No significant interval change is observed in the bony thorax. Output: The left apex has not been included on this radiograph. The ET tube terminates 3.9 cm above the carina. The NG tube terminates in the stomach. Surgical clips and a faint metallic coil project over the chest. A left PICC terminates in the mid SVC. EKG leads overlie the chest wall. The lung volumes are low. There are persistent bilateral mid and lower zone hazy opacities. There are bilateral hilar and perihilar linear opacities. Bilateral pleural effusions are present. The right pleural effusion is greater than the left. No pneumothorax is observed on the right. No cardiomegaly is present.

Stage 3: Creating Progression-Specific Reports. For each predefined finding of interest, we generate three reports—one each for "improved," "stable," and "worsened" progression. If a specific finding does not exist in the original report, we synthesize the corresponding sentence. The prompt used to generate these progression-specific sentences is:

Role: You are an expert assistant specialized in processing medical text, specifically Chest X-Ray (CXR) reports.

Task: Given a CXR report text and a specific clinical 'finding', perform the following steps:

1. **Preprocessing:** * Review the input CXR report. * Remove any sentences that *solely* describe the view position (e.g., "PA and lateral views were obtained.", "AP portable view.", "Single frontal view provided."). Do *not* remove view information if it's integrated into a sentence describing a finding (though this is less common). * Retain all sentences describing clinical observations, findings, comparisons, and impressions. Avoid removing other general "unnecessary" words; focus primarily on removing dedicated view position sentences.

2. **Modification based on Finding:** * Identify if the preprocessed report text contains mentions of the provided 'finding'. * **If the finding is mentioned:** * Locate the primary sentence(s) describing the status or appearance of the 'finding'. * Create three versions of the preprocessed report: * **Improved:** Modify the relevant sentence(s) minimally to indicate the finding has 'improved', 'decreased', 'resolved', or similar positive change. * **Stable:** Modify the relevant sentence(s) minimally to indicate the finding is 'stable', 'unchanged', or 'similar'. If the original text already implies stability, ensure this version reflects that clearly. * **Worsened:** Modify the relevant sentence(s) minimally to indicate the finding has 'worsened', 'increased', become 'more severe', or similar negative change. * Make *only the minimal changes* necessary to the specific part about the finding's status. Keep the rest of the report text identical to the preprocessed version. * **If the finding is NOT mentioned:** * Create three versions of the report by appending a new, concise sentence to the end of the preprocessed report text: * **Improved:** Append a sentence like: "The [finding] shows improvement." or "[Finding] is improved." * **Stable:** Append a sentence like: "The [finding] appears stable." or "[Finding] is stable." * **Worsened:** Append a sentence like: "The [finding] has worsened." or "There is worsening of the [finding]." or "[Finding] has increased." * Use the original preprocessed report text for the beginning of each version.

3. **Output:** * Format the final output as a single JSON object string. * The JSON object must have exactly three keys: '"improved"', '"stable"', and '"worsened"'. * The value for each key should be the full text of the corresponding modified report generated in Step 2. Ensure the output is valid JSON.

Input Format Reminder: The user will provide input in the following format: 'finding': [The specific clinical finding] 'report': [The full text of the CXR report]

Example 1 (Illustrative - do not repeat in output): 'finding': pleural effusion 'report': "When compared to the prior study, the left-sided pleural effusion appears stable. Left consolidation appears relatively stable. No pneumothoraces are seen. The rest of the support lines and tubes are unchanged in position. PA and lateral views were obtained."

Expected Output Example 1 (Illustrative - do not repeat in output): "'json { "improved": "When compared to the prior study, the left-sided pleural effusion is improved. Left consolidation appears relatively stable. No pneumothoraces are seen. The rest of the support lines and tubes are unchanged in position.", "stable": "When compared to the

prior study, the left-sided pleural effusion appears stable. Left consolidation appears relatively stable. No pneumothoraces are seen. The rest of the support lines and tubes are unchanged in position.", "worsened": "When compared to the prior study, the left-sided pleural effusion is worsened. Left consolidation appears relatively stable. No pneumothoraces are seen. The rest of the support lines and tubes are unchanged in position." }

Example 2 (Illustrative - do not repeat in output): 'finding': pneumothorax 'report': "Interval improved aeration is noted at both lung bases. Residual patchy and linear left lower lobe atelectasis remains. A small left pleural effusion is present. Single frontal view."

Expected Output Example 2 (Illustrative - do not repeat in output): "'json { "improved": "Interval improved aeration is noted at both lung bases. Residual patchy and linear left lower lobe atelectasis remains. A small left pleural effusion is present. The Pneumothorax is improved.", "stable": "Interval improved aeration is noted at both lung bases. Residual patchy and linear left lower lobe atelectasis remains. A small left pleural effusion is present. The Pneumothorax is stable.", "worsened": "Interval improved aeration is noted at both lung bases. Residual patchy and linear left lower lobe atelectasis remains. A small left pleural effusion is present. The Pneumothorax has worsened." }

Now process the following input:

8. Binary Interval-Change Dataset

We construct binary interval-change labels (`change` vs. `no change`) from radiology report impressions across four datasets. Cases are labeled as `no change` if the impression contains the phrase *"no interval change."* Cases are labeled as `change` if the impression contains any of the following progression-related keywords:

- **Worsening:** aggravated, exacerbated, increase, worsen, progression, enlarged
- **Improving:** improve, decrease, diminished, reduce, regress, resolve, disappear
- **New/Developing:** new, newly, developed, developing, recur, recurrence
- **Interval changes:** interval decrease, interval increase, interval improvement, interval worsening

Reports matching neither criterion are excluded. Label counts per dataset are summarized in Tab. 6.

Table 6. Binary interval-change label distribution. Label 0 denotes `no change` and Label 1 denotes `change`.

Dataset	Label 0	Label 1	Total
CheXpert	49,591	53,505	103,096
MIMIC	22,527	45,370	67,897
RexGradient	8,902	10,921	19,823
Private	5,000	5,000	10,000

CheXpert and MIMIC exhibit moderate class imbalance toward `change`, reflecting the higher prevalence of progression-related language in follow-up reports. The private hospital cohort (SNU2) is balanced by randomly sampling 5,000 cases per class from the full dataset.

9. Experiment

9.1. Zero-Shot Prompt Design

For each finding and progression class, we design 12–17 distinct prompts to capture the diverse phrasing typically found in radiology reports. Using multiple prompts per class helps reduce score variance, as relying on a single template can lead to unstable results. During zero-shot classification, we compute the cosine similarity between the image representation and each prompt corresponding to a specific progression label, and average these scores to obtain the final prediction.

```
all_prompt={
   'pneumothorax':
      {'improving':["Improved right pneumothorax.",
         "Decreased size of pneumothorax compared to
             prior.",
         "Interval improvement in pneumothorax.",
         "Partial resolution of left pneumothorax.",
         "Pneumothorax has decreased in size.",
         "Improvement in previously noted
             pneumothorax.",
         "Marked reduction in size of pneumothorax.",
         "Smaller right apical pneumothorax noted
             today.",
         "Improved pneumothorax, no acute findings.",
         "Reduction in pneumothorax volume.",
         "Improved appearance of left apical
             pneumothorax.",
         "Pneumothorax is resolving.",
         "Pneumothorax shows interval decrease."],
      'stable' : [
         "Stable small right pneumothorax.",
         "No increase in size of pneumothorax.",
         "Pneumothorax appears unchanged from prior.",
         "Small left pneumothorax, no acute findings.",
         "No evidence of expanding pneumothorax.",
         "Pneumothorax is stable with no tension
             physiology.",
         "Minimal pneumothorax, no intervention
             needed.",
         "Persistent small pneumothorax without
             progression.",
         "Pneumothorax noted, patient remains stable
             clinically.",
         "No signs of worsening pneumothorax.",
         "Pneumothorax is stable in appearance and
             size.",
         "No interval change in pneumothorax.",
         "Left apical pneumothorax stable compared to
             prior.",
         "Pneumothorax remains small and non-tension."],
      'worsening' : [
         "Worsening right pneumothorax.",
         "Increased size of pneumothorax compared to
             prior.",
         "Interval increase in pneumothorax.",
         "Progression of left pneumothorax.",
         "Pneumothorax has enlarged.",
         "Worsening left apical pneumothorax.",
         "Marked increase in pneumothorax size.",
         "Pneumothorax increasing, consider
             intervention.",
         "Expansion of previously noted pneumothorax.",
         "Pneumothorax now involves greater lung
             volume.",
         "New increase in size of right pneumothorax.",
         "Pneumothorax shows interval worsening.",
         "Pneumothorax progressing compared to previous
             imaging.",
         "Enlarging pneumothorax noted on follow-up."]
      },
   'pleural_effusion':{
```

```
    "improving": [
        "Improved right pleural effusion.",
        "Decreased size of pleural effusion compared
            to prior.",
        "Interval improvement in pleural effusion.",
        "Partial resolution of left pleural effusion.",
        "Reduction in pleural effusion volume.",
        "Pleural effusion has decreased in size.",
        "Marked reduction in right pleural effusion.",
        "Pleural effusion is resolving.",
        "Improved appearance of left pleural
            effusion.",
        "Improvement in previously noted pleural
            effusion.",
        "Less fluid seen in pleural space than
            before.",
        "Pleural effusion shows interval decrease.",
        "Decreased right basilar pleural effusion.",
    ],
    "stable": [
        "Stable small right pleural effusion.",
        "No increase in size of pleural effusion.",
        "Pleural effusion appears unchanged from
            prior.",
        "Small left pleural effusion, no acute
            findings.",
        "No evidence of expanding pleural effusion.",
        "Pleural effusion is stable in appearance and
            size.",
        "Minimal pleural effusion, no intervention
            needed.",
        "Persistent small pleural effusion without
            progression.",
        "Pleural effusion noted, patient remains
            stable clinically.",
        "No signs of worsening pleural effusion.",
        "No interval change in pleural effusion.",
        "Left basilar pleural effusion stable compared
            to yesterday.",
        "Pleural effusion remains small and
            unchanged.",
        "Stable bilateral pleural effusions."
    ],
    "worsening": [
        "Worsening right pleural effusion.",
        "Increased size of pleural effusion compared
            to prior.",
        "Interval increase in pleural effusion.",
        "Progression of left pleural effusion.",
        "Pleural effusion has enlarged.",
        "Worsening left basilar pleural effusion.",
        "Marked increase in pleural effusion size.",
        "Expansion of previously noted pleural
            effusion.",
        "Pleural effusion increasing, consider
            intervention.",
        "New increase in size of pleural effusion.",
        "Pleural effusion now causes greater lung
            compression.",
        "Pleural effusion shows interval worsening.",
        "Pleural fluid accumulation appears
            progressive.",
        "Enlarging pleural effusion noted on
            follow-up."
    ]
    },
'consolidation':{
    "improving": [
        "Improved right lower lobe consolidation.",
        "Decreased area of consolidation.",
        "Interval improvement in consolidation.",
        "Consolidation has partially resolved.",
        "Marked reduction in pulmonary consolidation.",
        "Clearing of previously noted consolidation.",
        "Consolidation is less extensive than prior.",
        "Improved left basilar consolidation.",
        "Reduction in airspace consolidation.",
        "Consolidation resolving on follow-up
            imaging.",
        "Less dense consolidation compared to prior.",
        "Airspace opacity improving.",
        "Consolidation has diminished since prior
            study.",
        "Patchy consolidation appears improved.",
        "Fading consolidation with treatment."
    ],
    "stable": [
        "Stable consolidation in right lower lobe.",
        "No significant change in consolidation.",
        "Persistent left basilar consolidation.",
        "Consolidation appears unchanged from prior.",
        "Airspace opacity remains stable.",
        "No interval change in consolidation.",
        "Chronic consolidation with no acute
            findings.",
        "Stable patchy consolidation.",
        "Consolidation noted without progression.",
        "No new consolidation identified.",
        "Findings consistent with stable
            consolidation.",
        "Consolidation remains unchanged.",
        "Stable appearance of parenchymal
            consolidation.",
        "Consolidation is chronic and stable.",
        "No worsening of consolidation."
    ],
    "worsening": [
        "Worsening right upper lobe consolidation.",
        "Increased area of consolidation.",
        "Consolidation more extensive than prior.",
        "Interval progression of consolidation.",
        "New or expanding consolidation noted.",
        "Consolidation has worsened.",
        "Marked increase in pulmonary consolidation.",
        "Confluent consolidation involving multiple
            lobes.",
        "Airspace consolidation increasing.",
        "Progressive dense consolidation.",
        "Patchy consolidation more pronounced.",
        "Worsening consolidation despite treatment.",
        "New left basilar consolidation with
            progression.",
        "Expanding area of alveolar consolidation.",
        "Increased opacification consistent with
            worsening consolidation."
    ]
    },
'edema': {
    "improving": [
        "Improved pulmonary edema.",
        "Decreased pulmonary vascular congestion.",
        "Edema appears less prominent than prior.",
        "Interval improvement in pulmonary edema.",
        "Partial resolution of interstitial edema.",
        "Reduction in alveolar edema.",
        "Pulmonary edema has decreased in extent.",
        "Marked reduction in pulmonary edema.",
        "Clearing of previously seen pulmonary edema.",
        "Improved vascular congestion.",
        "Improved interstitial markings.",
        "Pulmonary edema resolving with treatment.",
        "Decreased perihilar opacities.",
        "Edema improving compared to previous study.",
        "Less pulmonary edema seen on current film."
    ],
    "stable": [
        "Stable pulmonary edema.",
        "No significant change in pulmonary edema.",
        "Pulmonary edema appears unchanged from
            prior.",
        "Edema remains stable in extent.",
        "Persistent mild pulmonary edema.",
        "Pulmonary vascular congestion unchanged.",
        "Interstitial markings stable.",
        "No interval change in pulmonary edema.",
        "Pulmonary edema noted without progression.",
        "No worsening of pulmonary edema.",
        "Edema appears chronic and stable.",
```

```
        "Stable vascular congestion.",
        "No new signs of fluid overload.",
        "Pulmonary edema similar to previous exam.",
        "Mild pulmonary edema, no acute change."
    ],
    "worsening": [
        "Worsening pulmonary edema.",
        "Increased pulmonary vascular congestion.",
        "Edema appears more prominent than prior.",
        "Interval increase in pulmonary edema.",
        "Progressive alveolar edema.",
        "Pulmonary edema has worsened.",
        "Marked increase in pulmonary edema.",
        "Expansion of interstitial edema.",
        "New or worsening bilateral pulmonary edema.",
        "Pulmonary edema now more confluent.",
        "Increasing perihilar opacities.",
        "Worsening interstitial markings.",
        "Pulmonary congestion progressing.",
        "Increased fluid overload signs on imaging.",
        "Diffuse worsening of pulmonary edema pattern."
    ]},
'pneumonia':{
    "improving": [
        "Improved right lower lobe pneumonia.",
        "Decreased consolidation in left lung.",
        "Pneumonia shows interval improvement.",
        "Clearing of previously seen infiltrates.",
        "Reduction in airspace opacity.",
        "Partial resolution of pneumonia.",
        "Consolidation is less extensive than prior.",
        "Improved left basilar pneumonia.",
        "Decreased right middle lobe opacities.",
        "Pneumonia improving with antibiotic therapy.",
        "Pulmonary infiltrates have diminished.",
        "Improved patchy opacities.",
        "Interval decrease in parenchymal opacities.",
        "Airspace disease appears less prominent.",
        "Pneumonia resolving compared to previous
            imaging."
    ],
    "stable": [
        "Stable right lower lobe pneumonia.",
        "Pneumonia appears unchanged from prior.",
        "Persistent left basilar consolidation.",
        "No interval change in airspace disease.",
        "Patchy infiltrates stable in appearance.",
        "Consolidation remains without significant
            change.",
        "No progression of pneumonia.",
        "Airspace opacity unchanged.",
        "Stable pneumonia on follow-up imaging.",
        "No new consolidation identified.",
        "Chronic-appearing infiltrates, no acute
            change.",
        "Stable bilateral patchy opacities.",
        "No worsening of pneumonia noted.",
        "Findings consistent with prior pneumonia,
            stable.",
        "Stable airspace disease, no interval change."
    ],
    "worsening": [
        "Worsening right upper lobe pneumonia.",
        "Increased consolidation in left lower lobe.",
        "Pneumonia appears more extensive than prior.",
        "Interval progression of pneumonia.",
        "New or worsening bilateral infiltrates.",
        "Airspace opacities have increased.",
        "Expansion of previously seen pneumonia.",
        "Pneumonia worsening despite treatment.",
        "More confluent consolidation noted today.",
        "Increasing parenchymal opacities.",
        "Marked progression of airspace disease.",
        "Increased patchy opacities compared to
            prior.",
        "Pulmonary infiltrates have progressed.",
        "Worsening left lower lobe pneumonia.",
        "New consolidation suggestive of worsening
            pneumonia."
    ]
    }
}
```

9.2. Ablation for Hyperparameters

We ablate the pretraining weight W for the Change-aware Sigmoid loss and the fine-tuning weight λ for the Temporal Consistency Loss (TCL). Table 7 reports average macro-accuracy on MS-CXR-$T_{\text{retrieval}}$ when varying W. Although $W = 0.5$ achieves the highest Standard accuracy, $W = 1$ provides the best overall trade-off, yielding the strongest Reversed, Combined, and Consistency scores. Setting $W = 2$ places too much emphasis on the inverted pairs and degrades performance across all protocols.

Table 8 summarizes the effect of the fine-tuning weight λ on supervised MS-CXR-T classification. Introducing BiCE alone ($\lambda = 0$) already results in a substantial improvement over the baseline across all evaluation protocols, demonstrating that enforcing label inversion provides a strong directional signal. Adding TCL with a small weight ($\lambda = 1$) produces performance similar to $\lambda = 0$, reflecting the scale difference between the cross-entropy and TCL objectives. Increasing the TCL weight to $\lambda = 50$ yields the best balanced performance, with the high Reversed, Combined, and Consistency scores and only a minor reduction relative to the BiCE-only setting in the Standard metric. At $\lambda = 100$, directional metrics remain strong, but Standard and Combined accuracies begin to drop, indicating that overly large TCL weights may over-regularize the model. Based on these trends, we adopt $\lambda = 50$ for all main experiments.

Table 7. Effect of pretraining weight W on **MS-CXR-$T_{\text{retrieval}}$**. We report average macro-accuracy (%) across all findings for each evaluation protocol.

W	Average Accuracy (%)			
	Standard	Reversed	Combined	Consistency
0	50.2	50.9	55.4	32.7
0.5	**54.9**	53.2	59.1	36.6
1	54.1	**54.0**	**59.2**	**37.9**
2	49.7	51.2	54.3	33.1

Also, in practice, we do not apply the inversion-aware objectives from the very beginning of training. If BiCE or TCL is introduced too early, the model can quickly converge to a degenerate solution that predicts `stable` for most pairs, which locally minimizes the bidirectional losses but is not clinically meaningful. To avoid this shortcut, we first warm up the model with standard objectives and then enable the Change-aware Sigmoid loss after 10 pretraining epochs, and TCL after 20 epochs of fine-tuning. This staging allows the backbone to learn non-trivial temporal struc-

Table 8. Effect of fine-tuning weight λ on **MS-CXR-T** supervised classification. We report average macro-accuracy (%) across all findings for each evaluation protocol.

λ	Average Accuracy (%)			
	Standard	Reversed	Combined	Consistency
Base	61.1	53.3	59.6	39.5
0	**65.2**	62.0	63.4	53.1
1	65.0	61.2	63.1	53.3
50	64.1	63.7	**63.6**	**57.3**
100	62.7	**63.8**	61.4	55.6

ture before inversion-aware regularization is applied.

10. Additional Information

10.1. Clinical Utility of Reversed and Combined Evaluations

Radiologists typically assess temporal progression only in the forward (standard) direction. The *Reversed* and *Combined* evaluations are therefore introduced as analytical tools to rigorously validate the reliability of forward predictions.

The *Reversed* evaluation tests whether a model truly understands temporal progression. For instance, a model reaching 70% accuracy in the forward direction but only 40% in the reversed direction likely exploits spurious cues rather than genuine temporal reasoning. Such discrepancies may undermine the trustworthiness of forward predictions.

The *Combined* evaluation complements this by enforcing consistency across both directions. By aggregating results from forward and reversed pairs, it exposes and helps mitigate directional biases. Together, these settings provide crucial analytical validation, ensuring that predictions in standard clinical scenarios are both reliable and interpretable.

Implications for Temporal Inversion. While reversible scenarios predominate across these findings, temporal inversion is not intended to model exact biological recovery. Instead, it provides a controlled perturbation that tests whether models capture the *direction* of temporal change. Radiologic follow-up primarily evaluates changes in lesion size, burden, or conspicuity—features that exhibit approximate reversibility and therefore lend themselves to inversion-based analysis. Our experiments Our experiments indicate that including inverted pairs improves directional sensitivity without compromising forward-order predictions, supporting temporal inversion as a practical stress test for assessing order-aware interval-change modeling.

10.2. Use of Large Language Models

We disclose the use of large language models (LLMs) in this work. LLMs were used in three ways: (i) to refine the clarity and presentation of writing; (ii) to assist in generating change/no-change labels from radiology reports (see Sec. 7.3); and (iii) to construct the MS-CXR-$T_{\text{retrieval}}$ benchmark by modifying radiology reports under controlled prompts (see Sec. 7.4). All LLM usage was limited to these supporting roles and did not alter the core experimental results or conclusions.