A. Additional Results and Analyses

A.1. Qualitative analysis of generated reports

Table A.1 shows example reports generated with BioViL-T and BioViL models, which are compared to the reference radiologist's reports. In comparison with BioViL which only models the current image, BioViL-T shows the benefit from incorporating prior study information and is able to provide factually more accurate reports especially in terms of describing temporal progression of the findings. This is showcased in the first two examples in the table: In the first row, BioViL-T is able to comment on not only the presence of the pleural effusion but also its improvement while BioViL fails to mention the change. In the second example, BioViL-T is able to correctly identify that there is no relevant change by comparing with the previous study, while BioViL wrongly hallucinates the tube in the current image as a new placement. BioViL-T can also avoid hallucination of the temporal information when there is no prior study. For instance, in the third example, BioViL-T correctly acknowledges that there is no prior image and generates the report based on information from the single current image, while BioViL hallucinates a non-exisistent prior study and wrongly generates temporal descriptions in the report.

A.2. Further analysis on temporal classification

A subset of the *MS-CXR-T* benchmark dataset is re-annotated by an expert radiologist by blinding them to the existing ground-truth labels and displaying only pairs of images obtained from each subject. With the new set of labels, the analysis focuses on measuring the correlation between inter-rater agreement and image model's prediction errors. Figure A.1 shows the dependency between the two where the x-axis corresponds to the cross entropy loss between the *MS-CXR-T* benchmark labels and model predictions. We observe lower model performance on cases with smaller inter-rater reliability for the three classes in the dataset, indicating that the model's prediction errors occur more often for the cases where experts may disagree with each other.

A.3. Self-attention visualisation

In Figure A.2, we show examples of self-attention rollout [1] maps for pleural effusion and consolidation, including radiologist-annotated bounding boxes surrounding the corresponding pathology in each prior and current image.

To model the attention flow through the transformer encoder block, we first average each attention weight matrix across all heads, subsequently we multiply the matrices between every two layers. For every block we add the identity matrix in order to model the residual connections. Last, we only keep the top 10 % of attention weights per block to reduce noise in the final rollout map. In contrast to [21], we do not visualize the rollout map with respect to a [CLS] token. Instead, we choose a reference image patch from the center of the radiologist-annotated bounding boxes, marked with ⋆ in Figure A.2.

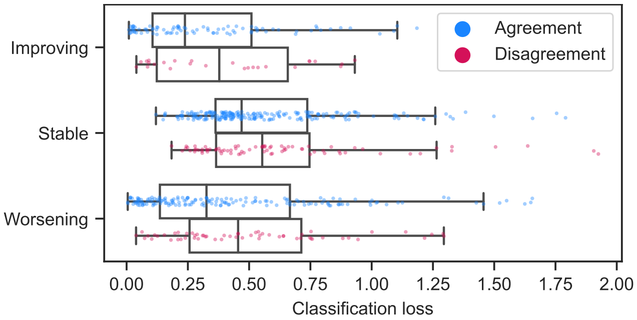


Figure A.1. Cross entropy between model predictions and *MS-CXR-T* temporal classification labels. 'Disagreement' indicates cases for which annotations differed amongst radiologists. Model performance is higher for cases with with low ambiguity ('Agreement').

We find that the rollout maps in Figure A.2 are in good agreement with radiologist-annotated bounding boxes, i.e., the reference patch attends to other patches within the bounding boxes in the prior and current image. In addition, we find that BioViL-T is robust to pose variations, e.g., in Figure A.2 (a) we show that despite the vertical shift between prior and current image, the reference patch attends to the correct image patches in the prior image.

To further assess the robustness of BioViL-T against pose variations between prior and current images, we performed multiple rotations to the prior image within a pair and computed rollout maps from the same reference patch in the current image. Figure A.3 shows that BioViL-T consistently attends to the corresponding anatomical region independently of the spatial transformation applied, demonstrating that registration is not needed.

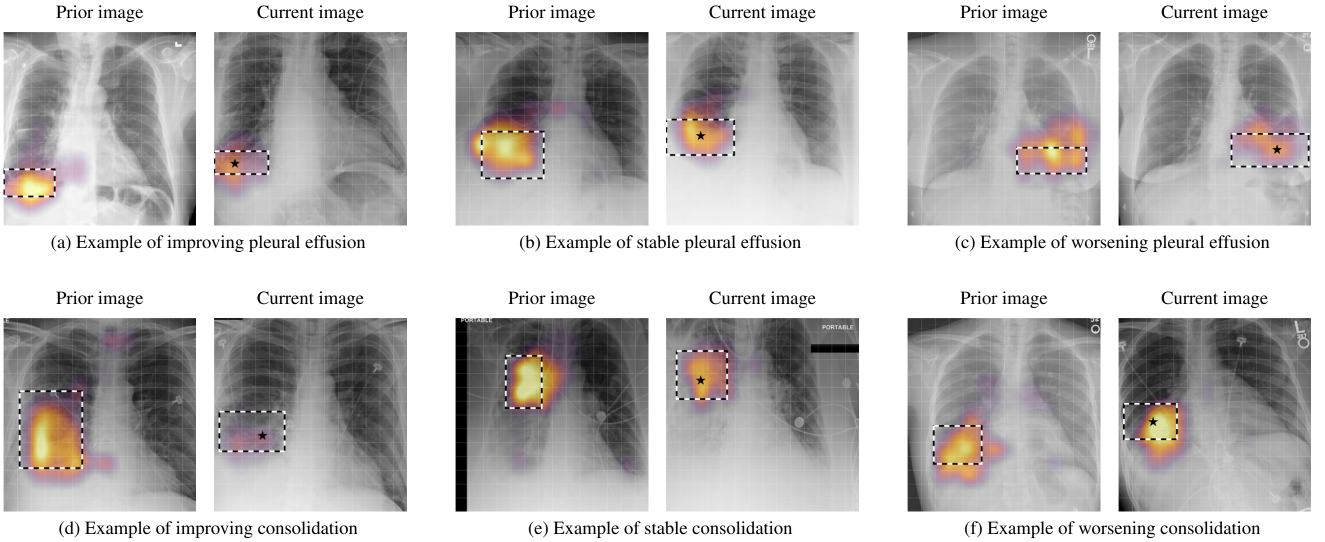
A.4. Data curation of imaging datasets

Large datasets often contain instances that are mislabelled or out of distribution [35]. We used BioViL-T to perform pairwise ranking of instances in MIMIC-CXR (Section 3.3, $\delta = 0.2$) and selected representative examples found in the dataset. Our method is able to select the most appropriate image for a range of different image-acquisition or image-processing issues (Figure A.4).

We found that many lateral acquisitions in the dataset were unexpectedly labelled as frontal (Figure A.4a). Some images contained only noise (Figure A.4b), non-human samples (Figures A.4d and A.4e) or incorrect anatomy (Figure A.4g). Often, acquisitions with an incomplete field of view (FOV) (i.e., the lungs are not completely visible) were repeated (Figure A.4c). Lastly, post-processed images were

Reference (by radiologist)	BioViL	BioViL-T
No evidence of acute cardiopulmonary process. Decreased right pleural effusion.	Small right pleural effusion.	Small right pleural effusion, decreased since _.
No relevant change as compared to the previous image. No evidence of pneumonia or other parenchymal pathology. Normal size of the heart. No pleural effusions.	In comparison with the study of _, there has been placement of a right ij port - a - cath that extends to the lower svc. No evidence of post procedure pneumothorax. The cardiac silhouette is within normal limits and there is no vascular congestion or pleural effusion. There is some asymmetry of opacification at the bases, more prominent on the left. In the appropriate clinical setting, this could possibly represent a developing consolidation.	As compared to the previous radiograph, no relevant change is seen. The lung volumes are normal. Normal size of the cardiac silhouette. Normal hilar and mediastinal structures. No pneumonia, no pulmonary edema, no pleural effusions.
No previous images. The cardiac silhouette is within normal limits and there is no evidence of vascular congestion, pleural effusion, or acute focal pneumonia.	In comparison with the study of _, there is little change and no evidence of acute cardiopulmonary disease. No pneumonia, vascular congestion, or pleural effusion.	No previous images. The cardiac silhouette is within normal limits and there is no vascular congestion, pleural effusion, or acute focal pneumonia.

Table A.1. Comparison between reports generated by radiologists, BioViL using only a single current image and BioViL-T using both the current and previous study. BioViL-T with access to longitudinal information can generate more accurate reports with more precise details on the progression of findings (as in the first and second example) while avoiding hallucination (in the third example). Blue box highlights the correct temporal information and brown box highlights incorrect temporal information including hallucination.

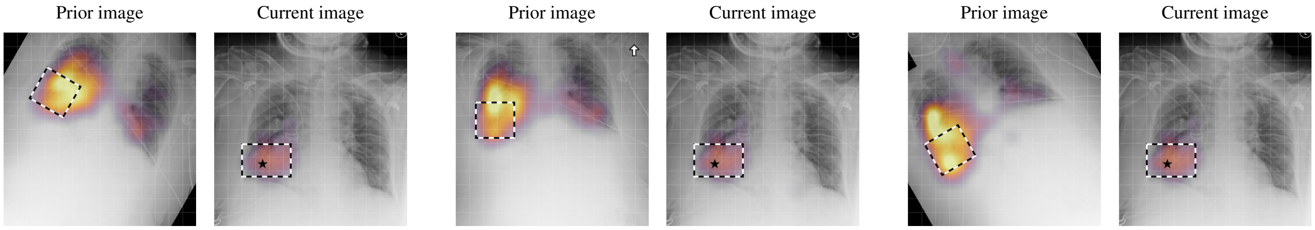


(a) Example of improving pleural effusion
(b) Example of stable pleural effusion
(c) Example of worsening pleural effusion
(d) Example of improving consolidation
(e) Example of stable consolidation
(f) Example of worsening consolidation

Figure A.2. Self-attention rollout maps [1] from the reference patch (marked with ⋆) to the current and prior images, overlaid on example cases of (a) improving, (b) stable and (c) worsening pleural effusion (top row) and consolidation (bottom row). The bounding boxes, annotated by a radiologist, show the area corresponding to the pathology. The centre patch in the bounding box for the current image was selected as reference. The grid (14 × 14) represents the visual tokens processed in the transformer encoder blocks.

detected by the algorithm such as contrast-enhanced scans

(Figure A.4i) that are not often used for diagnostic purposes



Prior image Current image Prior image Current image Prior image Current image

(a) Previous image rotated -30° (b) Original pair (c) Previous image rotated 30°

Figure A.3. Comparison of roll-out maps computed after applying in-plane spatial rotations to the prior image. The reference visual token (⋆) attends to the corresponding anatomical region annotated by an expert independent of the underlying spatial transformation.

in clinical practice.

A.5. Phrase-grounding on external data

We have additionally conducted a robustness analysis on an out-of-distribution dataset. For this purpose, a small set of expert labels (N=137 bounding-box–caption pairs) were collected on Open-Indiana CXR dataset [18] for phrase grounding on the same set of abnormalities as *MS-CXR* benchmark [10]. The dataset differs in terms of text token distribution, demographics, and disease prevalence. The experiment was performed with the same methods and setup described in Section 4.3. The results show that the performance gains due to temporal pre-training is observed to be consistent on external datasets.

Table A.2. Multi-modal phrase-grounding results obtained on a subset of Open-Indiana CXR dataset [18] image-text pairs. “Multi-image” column indicates the input images used at test time. The results are reported in terms of micro-averages owing to the limited number of samples in some classes.

Method	Pre-Train	Multi-Image	Avg. CNR	Avg. mIoU
BioViL [9]	Static	✗	1.19 ± 0.04	0.259 ± 0.003
BioViL-T	Temporal	✗	**1.53** ± **0.05**	**0.289** ± **0.006**

B. Temporal aspects of the MIMIC-CXR v.2 dataset

Subjects in the MIMIC-CXR dataset often have multiple associated studies that happened at different times. A study, sometimes referred to as an ‘exam’ or ‘procedure’, refers to “one or more images taken on a single visit to a medical facility”[8]. To assess pathology progression, radiologists compare images (also referred to as ‘scans’ or ‘series’) from different studies. In the MIMIC-CXR dataset, each study (with one or more images) is accompanied by the report written by the radiologist. Figure B.1 represents the distribution of studies per subject within MIMIC-CXR and the corresponding cumulative distribution function, showing that 67 % of the subjects have at least two different associated studies (and therefore at least two images acquired at different stages of the disease).

Another way to quantify temporal information in MIMIC-CXR is through the progression labels provided by the Chest ImaGenome dataset [72]. These progression labels are extracted from the reports and thus identify the cases when the radiologist explicitly describes changes. We found that in MIMIC, around 40 % of the reports are associated with a progression label from any of the available findings defined by ImaGenome.

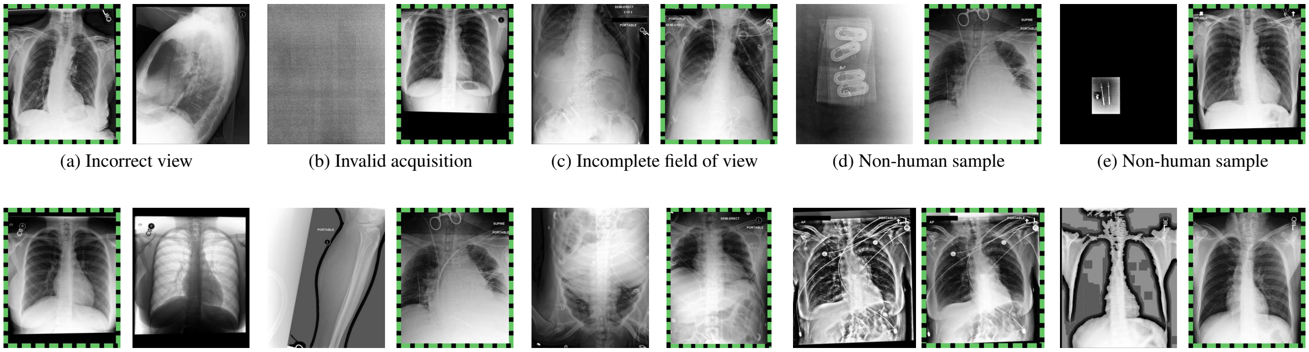
C. *MS-CXR-T* benchmark

C.1. Temporal image classification

The *MS-CXR-T* temporal image classification contains progression labels for five findings (Consolidation, Edema, Pleural Effusion, Pneumonia and Pneumothorax) across three progression classes (`Improving`, `Stable`, and `Worsening`). This benchmark builds on the publicly available Chest ImaGenome gold and Chest ImaGenome silver datasets [72] which provide progression labels automatically derived from radiology reports. We collected a set of studies that are part of the ImaGenome silver dataset, excluding any studies that had been previously verified as part of the ImaGenome gold dataset. Additionally, we excluded studies where there are multiple progression labels for a single pathology (e.g. left pleural effusion has increased, right pleural effusion remains stable). We conducted a review process of the selected candidates, asking a board certified radiologist to either accept or reject the label. To inform their review of the labels, the radiologist was given access to the radiology report for the current image, and the sentence from which the auto generated label had been extracted.

After collecting our curated labels and labels from the ImaGenome gold dataset, we matched the report-based labels to specific image pairs, performing a second data curation step to create the image dataset. To ensure the diagnostic quality of all images in the dataset, if a study had multiple frontal scans we performed a quality control step asking a radiologist to select the best image for each study. Fig. F.1 shows examples from the benchmark across differ-

[8] Adapted from `https://ncithesaurus.nci.nih.gov/`



(a) Incorrect view (b) Invalid acquisition (c) Incomplete field of view (d) Non-human sample (e) Non-human sample

(f) Inverted intensities (g) Non-chest sample (h) Image orientation (i) Post-processed image (j) Processing artefacts

Figure A.4. Pairwise ranking of images performed by the proposed data curation method (see Section 3.3) on images from the MIMIC-CXR v2 dataset. Images highlighted with dashed green rectangles are automatically selected by our method and used for training to improve model's downstream performance. The rejected image samples may not be appropriate for training due to image acquisition or image processing issues as shown in each subfigure above.

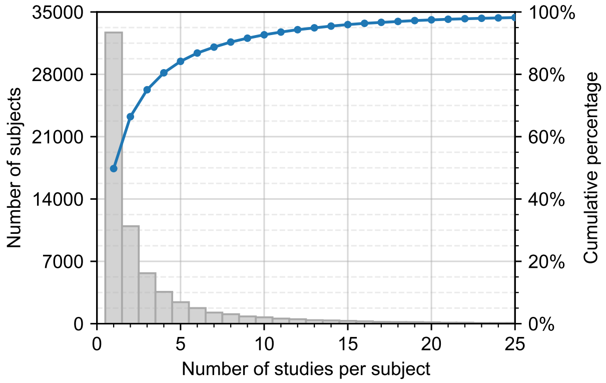


Figure B.1. Number of studies per subject in the MIMIC-CXR dataset. A study, sometimes referred to as an 'exam' or 'procedure', refers to "one or more images taken on a single visit to a medical facility" (adapted from https://ncithesaurus.nci.nih.gov/). Note that 67 % of subjects have at least two studies that happened at different times.

ent pathologies and progression labels.

The class distribution for the image classification task in *MS-CXR-T* is shown in Tab. C.1. As seen in the table, the class distribution of the dataset skews towards the `stable` and `worsening` classes. This could be explained as patients are more likely to get a chest X-ray scan when their condition is stable or deteriorating as opposed to when there is an improvement in patient condition.

C.2. Temporal sentence similarity

In this section, we describe the process of creating the *MS-CXR-T* temporal sentence similarity benchmark, which consists of pairs of paraphrase or contradiction sentences in terms of disease progression. We create this dataset using two different methods, `RadGraph` where paraphrase and contradiction sentence pairs are discovered by analysing graph representations of sentences and `Swaps` where paraphrases and contradictions are created by swapping out temporal keywords in the sentence.

To create this dataset, we first collected a set of sentences from the MIMIC dataset, using the Stanza constituency parser [82] to extract individual sentences from reports. Using the CheXbert labeller [63], we filtered this set to sentences that described one of seven pathologies - Atelectasis, Consolidation, Edema, Lung Opacity, Pleural Effusion, Pneumonia or Pneumothorax. We then filtered to sentences which contained at least one mention of a temporal keyword. Using this sentence pool, paraphrase and contradiction pairs were constructed in two ways. (I) We paired sentences from the sentence pool by matching on RadGraph

Table C.1. *MS-CXR-T* temporal image classification benchmark: Showing the distribution of multi-image studies across different clinical findings, distribution of classes {`Improving`, `Stable`, `Worsening`} per finding, and number of subjects.

Findings	**# of annotation pairs**	**Class distribution**	**# of subjects**
Consolidation	201	14% / 42% / 44%	187
Edema	266	31% / 26% / 43%	241
Pleural effusion	411	19% / 49% / 32%	370
Pneumonia	237	8% / 25% / 67%	218
Pneumothorax	211	15% / 55% / 30%	148
Total	1326	18% / 40% / 42%	800

Table C.2. *MS-CXR-T* temporal sentence similarity benchmark: Number of paraphrase and contradiction examples in the full dataset and across the `RadGraph` and `Swaps` subsets.

Subset	**# of paraphrase pairs**	**# of contradiction pairs**	**Total**
Radgraph	42	75	117
Swaps	99	145	244
Total	141	220	361

Table C.3. Examples of paraphrase and contradiction sentence pairs from the *MS-CXR-T* temporal sentence similarity benchmark. The examples are selected from the `RadGraph` and `Swaps` subsets (see Appendix C.2).

	Label	Sentence 1	Sentence 2
Swaps	Paraphrase	"Unchanged small-to-moderate right pleural effusion."	"Stable small-to-moderate right pleural effusion."
	Contradiction	"Interval worsening of the right-sided pneumothorax."	"Interval resolution of the right-sided pneumothorax."
RadGraph	Paraphrase	"There has also been a slight increase in left basal consolidation."	"There is slight interval progression of left basal consolidation."
	Contradiction	"Right mid and lower lung consolidations are unchanged."	"There has been worsening of the consolidation involving the right mid and lower lung fields."

[34] entities, relaxing the matching constraint only for temporal keywords and possible mentions of pathologies. (II) We swapped out temporal keywords in a sentence to create sentence pairs, choosing swap candidates from the top 5 masked token predictions from CXR-BERT-Specialized [9] provided they were temporal keywords. After creating candidate sentence pairs, we manually filtered out sentence pairs with ambiguous differences in terms of disease progression. A board certified radiologist then annotated each sentence pair as either paraphrase or contradiction. Sentences were filtered out in the annotation process if (I) they were not clear paraphrases or contradictions (II) the sentences differed in meaning and this difference was not related to any temporal information (III) they were not grammatically correct. The distribution of sentence pairs across the paraphrase and contradiction classes are described in Table C.2, see Table C.3 for examples from the benchmark.

D. Temporal entity matching

To quantify how well the generated report describes progression-related information, we propose a new metric, namely temporal entity matching (TEM) score.

D.1. Metric Formulation

We first extract entities (tagged as "observation" or "observation_modifier") from the text by running the named entity recognition model in the Stanza library [82]. Within the extracted entities, we manually curated a list of temporal entities that indicate progression (Appendix D.2). The list is reviewed by an expert radiologist. Given extracted temporal entities E in N pairs of reference and generated reports, we calculate global precision (p_E) and global recall (r_E), which are later used to compute the TEM score. It is defined as the harmonic mean of precision and recall (also known as the F1 score).

$$p_E = \frac{\sum_{i=1}^{N} |E^i_{gen} \cap E^i_{ref}|}{\sum_{i=1}^{N} |E^i_{gen}|} \tag{3}$$

$$r_E = \frac{\sum_{i=1}^{N} |E^i_{gen} \cap E^i_{ref}|}{\sum_{i=1}^{N} |E^i_{ref}|} \tag{4}$$

D.2. List of temporal keywords

The list of temporal keywords used to compute the TEM score are as follows: {bigger, change, cleared, constant, decrease, decreased, decreasing, elevated, elevation, enlarged, enlargement, enlarging, expanded, greater, growing, improved, improvement, improving, increase, increased, increasing, larger, new, persistence, persistent, persisting, progression, progressive, reduced, removal, resolution, resolved, resolving, smaller, stability, stable, stably, unchanged, unfolded, worse, worsen, worsened, worsening, unaltered}.

E. Architecture and implementation details

E.1. Hyper-parameters

The models are trained in a distributed setting across 8 GPU cards. For pre-training, we use a batch size of 240 (30 * 8 GPUs) and the AdamW optimizer [43]. We use a linear learning rate scheduler with a warm-up proportion of 0.03 and base learning rate of 2×10^{-5}. We train for a maximum of 50 epochs and use validation set loss for checkpoint selection. The overall loss is a sum of components with weighting factors: global contrastive (1.0), local contrastive (0.5), and image-guided MLM (1.0) respectively, see Sec. 3.1 for further details on their formulation.

Following [9] we use sentence permutation as text-based data augmentation. Similarly, spelling errors in the reports are corrected prior to tokenisation of the text data[9]. For image augmentations, note that we apply the same augmentation to current and prior images to prevent severe misalignment. We resize the shorter edge to 512 and centre-crop to (448, 448). We apply random affine transformations (rotation up to 30° and shear up to 15°) and colour jitter (brightness and contrast).

E.2. Training infrastructure

We train with distributed data processing (DDP) on eight NVIDIA Tesla V100s with 32GB of memory each. To handle inconsistently-present prior images with DDP, we define

[9] https://github.com/farrell236/mimic-cxr/blob/master/txt/section_parser.py

a custom batch sampler. This sampler is a mixture of two samplers, in proportion to their dataset coverage: a sampler which produces batches with *only* multi-image examples – $(\mathbf{x}_{\text{img}}^{\text{curr}}, \mathbf{x}_{\text{img}}^{\text{prior}}, \mathbf{x}_{\text{txt}}^{\text{curr}}) \in \mathcal{D}_m$ and one with only single-image examples – $(\mathbf{x}_{\text{img}}^{\text{curr}}, \varnothing, \mathbf{x}_{\text{txt}}^{\text{curr}}) \in \mathcal{D}_s$. Each GPU then processes a batch which is entirely single or multi-image, avoiding branching logic within the forward pass and enabling an efficient single pass through the CNN to process all input images (current or prior) by concatenating them along the batch dimension.

We confirmed that although the custom sampler theoretically impacts the order in which the dataset is traversed, it has a negligible effect on training metrics relative to fully random sampling. Since we train on eight GPUs and collect negatives across all GPUs during contrastive training, each update involves on average a representative mixture of both single-image and multi-image samples.

Finally, following [9] we use the DICOM images from MIMIC-CXR to avoid JPEG compression artefacts.

F. Adaptation and experimentation details

F.1. Fine-tuning BioViL-T for report generation

During fine-tuning of BioViL-T for report generation, we minimise the cross entropy loss to maximise the log likelihood of the report in an autoregressive manner given the input images. The model is initialised from the pretrained weights of the image encoder and the text encoder. Similar to the cross-modal masked language modelling task, we additionally train a linear projection layer to map the projected patch embeddings to the same hidden dimension of the text encoder, and we train cross-attention layers in each transformer block. The difference from the masked language modelling task is that we change the bidirectional self-attention to unidirectional causal attention that can only access the past tokens. If trained with prior report, we pass the prior report as prefix to condition the generation of the current report (the current and prior report are separated by `[SEP]`), and we only back-propagate the gradients from the loss on the tokens in the current report.

For all experiments, we train the model for 100 epochs and we chose the best checkpoint according to metrics on the validation set. We performed grid search for learning rate in $[10^{-5}, 2 \times 10^{-5}, 5 \times 10^{-5}]$ and found 2×10^{-5} to be optimal. We ran each experiment with 3 random seeds and report mean and standard deviation.

In addition to the metrics we reported in the main text, we also evaluate the generated reports by named entity metric (NEM). This metric was defined in [45] to measure the accuracy of reporting clinically relevant entities in the generated reports (Similar to how TEM is computed to measure the match of temporal entities in our study). Following [45], we extract entities (tagged as “observation” or “observation_ modifier”) from the text by running the named entity recognition model in the Stanza library [82]. The results are presented in Tab. F.1.

Table F.1. Results for report generation task: Predictions are evaluated on NEM. The approaches are grouped into two broad categories: NN (Nearest Neighbour) and AR (Auto-Regressive). BioViL-T pre-training consistently yields superior decoding performance. Further, the use of prior image and report consistently yield performance gains demonstrating the importance of such domain priors.

	Method	Pre-training	Prior Img/Report	NEM
NN	CXR-RePaiR-2 [25]	BioViL	✗ / ✗	13.36
	Baseline (NN) [9]	BioViL	✗ / ✗	16.25
	Proposed (NN)	BioViL-T	✓/ ✗	17.55
AR	Baseline (AR) [9]	BioViL	✗ / ✗	24.27 ± 0.22
	Proposed	BioViL-T	✓/ ✗	25.50 ± 0.04
	Proposed	BioViL-T	✓/ ✓	**26.95 ± 0.17**

F.2. Nearest-neighbour-based report retrieval

The joint latent space learnt by BioViL-T can also be used to directly perform report retrieval without requiring task-specific model fine-tuning. Given the test image, we retrieve its semantically closest report from the training set in the joint latent space. Specifically, we encode each test image with the image model in BioViL-T and collect its projected image embeddings, and similarly we encode all the reports in the training data with their projected text embeddings. For each test study, we compute cosine similarity between the test image embedding and all the text embeddings from the training set in the joint latent space, and we retrieve the closest text embedding and use its corresponding report as the prediction. To evaluate the retrieval performance, we use the same decoding metrics on the retrieved reports and report results in the top section of Table 1. In a separate set of experiments, we also tried performing nearest neighbour search only within the image embedding space by retrieving the report associated with the closet image embedding, but this yielded sub-optimal performance compared with using the joint latent space.

F.3. Fine-tuning for temporal image classification

In this section, we describe the training dataset and fine-tuning procedure for the fully supervised and few-shot settings of the temporal image classification task. For this task, we finetune BioViL-T on a subset of the Chest ImaGenome silver dataset [72] to predict progression labels for 5 different pathologies. To create our training dataset, we filter out image pairs from this dataset where there are multiple directions of progression of a single pathology in the image-pair. We additionally perform an automatic data curation step to choose higher quality image pairs when possible, as described in 3.3. Table F.2 shows the number of training samples and label distribution for the training dataset.

Table F.2. Statistics of the training dataset used for downstream fine-tuning on temporal image classification.

Findings	**# labelled pairs**	**Class distribution**	**# of subjects**
Consolidation	7012	15% / 42% / 43%	3308
Edema	14170	28% / 33% / 39%	4813
Pleural effusion	26320	16% / 53% / 31%	6838
Pneumonia	8471	12% / 29% / 59%	4197
Pneumothorax	3795	21% / 57% / 22%	1161

For the fully supervised setting, we add a multilayer classification head to the BioViL-T image encoder and fine-tune the model independently for each pathology. We use weighted cross entropy loss with a batch size of 128 and the AdamW optimizer [43]. During parameter optimisation, positional encodings and missing-image embeddings are exempt from weight decay penalty as in [73]. We train for 30 epochs, with a linear learning rate schedule, a warmup proportion of 0.03 and a base learning rate of 1×10^{-5}. For data augmentation, we first resize the shorter edge of the image to 512 and centre crop to (448, 448). We apply random horizontal flips, random cropping, random affine transformations (rotation up to 30°, shear up to 15°), colour transforms (brightness and contrast) and Gaussian noise.

For the few-shot setting we tune only a single-layer linear head on the BioViL-T image encoder and freeze the rest of the encoder. We initialise the weight matrix of the linear head with values from encoded text prompts [9] for each of the three progression classes, and the bias matrix is initialised with zeros. To train, we again use weighted cross entropy loss, with a batch size of 32 and the AdamW optimizer. We use a learning rate of 1×10^{-3} and train for 40 epochs. For data augmentation, we resize the shorter edge of the image to 448 and center crop to (448, 488). We apply random horizontal flips, random affine transformations (rotation up to 45° and shear up to 25°), colour transforms (brightness and contrast). As in the pre-training step, we always synchronise image data augmentations to apply the identical transforms to the current and prior images.

F.4. Auto-regressive prompting for zero-shot temporal image classification

Following the GPT-3 style language prompting [11], we prompt the fine-tuned AR language decoding model with the template: "`[FINDING]` is" and infer the next token to perform temporal classification for each of the five findings. The mapping from the predicted next token to the three progression classes is characterised by a short list of tokens provided in Table F.3. After computing the posterior for each token in the list, the obtained values are normalised across the three classes, and the class with the highest score is selected as the prediction. The corresponding results are reported in Table 2.

Table F.3. Prompting the AR language decoding model for zero-shot image classification. The list above shows the mapping from decoded tokens to progression classes.

Target class	**Tokens**
Improving	better, cleared, decreased, decreasing, improved, improving, reduced, resolved, resolving, smaller
Stable	constant, stable, unchanged
Worsening	bigger, developing, enlarged, enlarging, greater, growing, increased, increasing, larger, new, progressing, progressive, worse, worsened, worsening

F.5. Further analysis of image-guided MLM

In Section 4.6 we used a simplified notation for the computation of $\Delta^{\text{prior}}_{\text{img}}(m)$ for ease of exposition – here we provide further detail. Recall that $\mathbf{w} = (w_1, \ldots, w_M)$ is a sequence of tokens and $\mathbf{w}_{\setminus m}$ is that sequence with token m masked. Let $p_\theta(\mathbf{w}_m \,|\, \mathbf{w}_{\setminus m}, \mathbf{x}^{\text{curr}}_{\text{img}}, \mathbf{x}^{\text{prior}}_{\text{img}})$ be the text model's predicted probability of token m given $\mathbf{x}^{\text{curr}}_{\text{img}}, \mathbf{x}^{\text{prior}}_{\text{img}}$, and $\mathbf{w}_{\setminus m}$ (θ are the weights of the model). Then, $l(w, p_\theta(\mathbf{w}_m \,|\, \mathbf{w}_{\setminus m}, \mathbf{x}^{\text{curr}}_{\text{img}}, \mathbf{x}^{\text{prior}}_{\text{img}}))$ is the cross-entropy loss of predicting token m given those inputs.

It is possible for different sentences in a report to refer to the same image finding. Since we mask single tokens at a time, to prevent information leakage from other sentences we consider each sentence in a report independently. Suppose report $\mathbf{x}^{\text{curr}}_{\text{txt}}$ consists of S sentences, so we have $\mathbf{x}^{\text{curr}}_{\text{txt}} = [\mathbf{w}^1, \texttt{[SEP]}, \ldots, \texttt{[SEP]}, \mathbf{w}^S]$, where $\mathbf{w}^s$ is the tokens of sentence s and `[SEP]` separates sentences.

For a given sample $(\mathbf{x}^{\text{curr}}_{\text{img}}, \mathbf{x}^{\text{prior}}_{\text{img}}, \mathbf{x}^{\text{curr}}_{\text{txt}}) \in \mathcal{D}_m$ in the test set indexed by i, we define

$$\begin{aligned}\delta_i(m) = \sum_{s \in S} [& l(m, p_\theta(\mathbf{w}^s_m \,|\, \mathbf{w}^s_{\setminus m}, \mathbf{x}^{\text{curr}}_{\text{img}}, \varnothing)) \\ & - l(m, p_\theta(\mathbf{w}^s_m \,|\, \mathbf{w}^s_{\setminus m}, \mathbf{x}^{\text{curr}}_{\text{img}}, \mathbf{x}^{\text{prior}}_{\text{img}}))]\end{aligned}$$

This is the MLM loss for predicting m given each *sentence* in the report with and without the prior image. Note that if m does not appear in a given sentence, its contribution to the sum is zero. The overall $\Delta^{\text{prior}}_{\text{img}}(m)$ is computed across all samples:

$$\Delta^{\text{prior}}_{\text{img}} = \frac{1}{N_m} \left(\sum_{i \in \mathcal{D}^{\text{test}}_m} \delta_i(m) \right) \tag{5}$$

where N_m is the number of *sentences* in reports in $\mathcal{D}^{\text{test}}_m$ in which token m appears. This estimate is subject to high variance when N_m is small. Hence, for Figure 4 we filter

Category	**Description**	**Examples**
Progression	Pertaining to change or progression	*bigger, cleared, new*
Support devices	Tubes, lines and implants	*nasogastric, pacemaker, cannula*
‘Other’	No clear category	*can, relevant, overall*
Stop word	‘Insignificant’ words	*the, no, of*
Positional	Localisation (not anatomical)	*right, lower, bilateral*
Meta	Pertaining to the report itself or practice of radiology	*evidence, radiograph, study*
Anatomy	Anatomical locations	*pulmonary, chest, mediastinal*
Descriptive	Qualitative appearance of a finding	*layering, focal, patchy*
Size or degree	Quantifying extent or severity	*extensive, moderate, severe*
Finding	Radiographic finding or pathology	*edema, penumonia, pneumothorax*
Uncertain	Expression of certainty or doubt	*may, possible, concerning*

Table F.4. Semantic categories used in Figure 4.

to tokens m with $N_m \geq 10$. We collected 931 tokens with $N_m \geq 10$ from the validation set for manual annotation by a board-certified radiologist. The categories, shown in Figure 4 and described in Table F.4 are specific to the radiology domain.

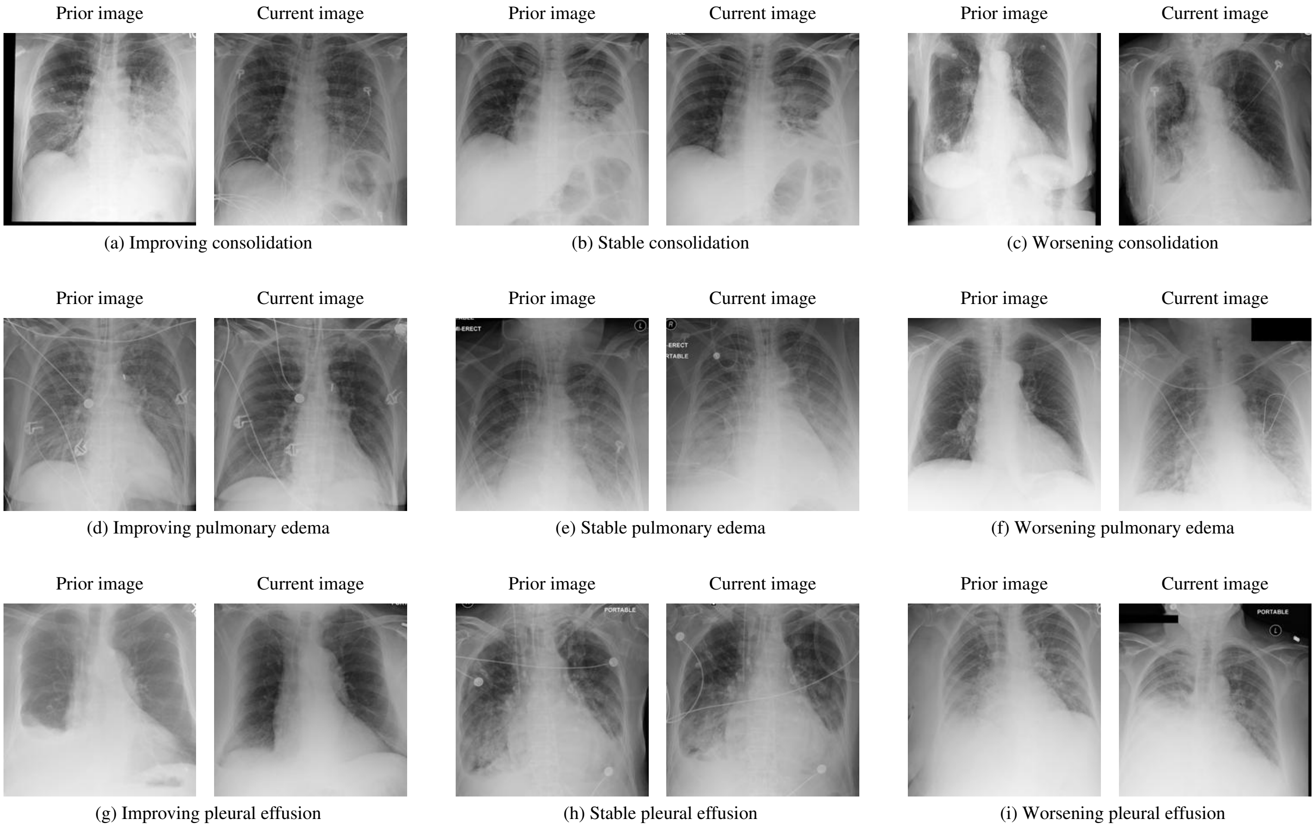
F.6. Sentence similarity experiment

The text models are evaluated in isolation to observe if their encoding is sensitive to key clinical observations. To achieve this, we assess the quality of sentence representations obtained from our text model by examining how well the contradiction and paraphrase pairs can be separated in the embedding space. Unlike the traditional NLI task where a model needs to be fine-tuned, here the models are probed in a zero-shot setting and the BERT output token embeddings are utilised. To do so, we encode the sentences from RadNLI and *MS-CXR-T* sentence similarity datasets with the `[CLS]` token from CXR-BERT-Specialised [9] and BioViL-T. For PubMedBERT [29] and CXR-BERT-General [9] which did not directly optimise the `[CLS]` token during pretraining, we follow [56] to average the token output embeddings to represent each sentence.

Cosine similarity is computed between the representations of each sentence pair in the dataset [56] and is used as logits for the binary classification between paraphrase and contradiction. Note that for RadNLI, we use the subset of ‘entailment’ and ‘contradiction’ pairs and discard the ’neutral’ pairs to unify the task across the two datasets. Given the similarities for each sentence pair, we report ROC-AUC and binary-accuracy. For the latter, a threshold value for each method is derived by setting aside a validation set. For this, we perform ten-fold cross validation and tune the threshold with step size of 0.005 on the validation set.

F.7. Image registration algorithm

In Section 4.2, image registration is applied to pairs of images as a preprocessing step to enable a fair comparison for the baseline approaches (e.g., BioViL [9]). We performed bidirectional multi-scale registration between image pairs optimising an affine transformation (4 degrees of freedom), using mutual information (MI) [65] with 128 bins as the similarity criterion. In more detail, the spatial transformation is characterised by four parameters: two for translation, one for isotropic scaling, and one for rotation. The optimisation is repeated five times with different random seeds for initialisation, and the run with the highest MI is selected to determine the final spatial alignment. To better identify the correspondences between the scans, bilateral filtering is applied to each image before registration to remove detailed texture whilst preserving edge information [38]. Our implementation is based on the SimpleITK library [44].



Prior image Current image

(a) Improving consolidation

Prior image Current image

(b) Stable consolidation

Prior image Current image

(c) Worsening consolidation

Prior image Current image

(d) Improving pulmonary edema

Prior image Current image

(e) Stable pulmonary edema

Prior image Current image

(f) Worsening pulmonary edema

Prior image Current image

(g) Improving pleural effusion

Prior image Current image

(h) Stable pleural effusion

Prior image Current image

(i) Worsening pleural effusion

Figure F.1. Examples of image pairs in our *MS-CXR-T* benchmark.