Learning to Exploit Temporal Structure for Biomedical Vision–Language Processing

Shruthi Bannur*, Stephanie Hyland*, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay†

Abstract

Self-supervised learning in vision–language processing (VLP) exploits semantic alignment between imaging and text modalities. Prior work in biomedical VLP has mostly relied on the alignment of single image and report pairs even though clinical notes commonly refer to prior images. This does not only introduce poor alignment between the modalities but also a missed opportunity to exploit rich self-supervision through existing temporal content in the data. In this work, we explicitly account for prior images and reports when available during both training and fine-tuning. Our approach, named BioViL-T, uses a CNN–Transformer hybrid multi-image encoder trained jointly with a text model. It is designed to be versatile to arising challenges such as pose variations and missing input images across time. The resulting model excels on downstream tasks both in single- and multi-image setups, achieving state-of-the-art (SOTA) performance on (I) progression classification, (II) phrase grounding, and (III) report generation, whilst offering consistent improvements on disease classification and sentence-similarity tasks. The resulting model excels on downstream tasks both in single- and multi-image setups, achieving state-of-the-art (SOTA) performance on (I) progression classification, (II) phrase grounding, and (III) report generation, whilst offering consistent improvements on disease classification and sentence-similarity tasks. We release a novel multi-modal temporal benchmark dataset, MS-CXR-T, to quantify the quality of vision–language representations in terms of temporal semantics. Our experimental results show the advantages of incorporating prior images and reports to make most use of the data.

1. Introduction

Self-supervision from image–text pairs has enabled the development of flexible general-purpose vision–language models both in the general domain [40, 53, 77] and for specialised domains such as biomedicine and radiology [9, 32, 81]. Vision–language processing (VLP) has shown that cross-modal supervision can provide a richer signal for training both image [19] and text [9] models. However, the success of VLP relies on paired samples sharing semantics, i.e., given an image and text pair, the text should describe the image with minimal extraneous detail [15, 16, 35].

In this regard, VLP in biomedicine and radiology poses a distinctive challenge, as reports routinely include comparisons to prior imaging studies [3, 47, 57]. Without knowl-
edge of this prior image\textsuperscript{1}, temporal information in the text modality, e.g. “Pneumonia is improving”, could pertain to any image containing “Pneumonia”, producing ambiguity during contrastive training (Figure 1). Despite this, the existing VLP work to date considers alignment between only single images and reports [9,32,46,81], going so far as to remove temporal content from reports in training data to prevent ‘hallucinations’ in downstream report generation [54]. However, temporal information can provide complementary self-supervision, solely by exploiting existing structure, and without requiring any additional data.

In this work, we neither ignore nor remove temporal information in the text modality, but explicitly account for it during pre-training. Rather than treating all image–report pairs in the dataset as independent, we exploit temporal correlations by making prior images available for comparison to a given report. To learn from this structure, we develop a temporal VLP pre-training framework named BioViL-T. A core component is its new multi-image encoder that can handle the absence of prior images and potential spatial misalignment between images across time. BioViL-T takes into account prior images where available, removing cross-modal ambiguity as illustrated in Fig. 1. Linking multiple images during pre-training proves beneficial to both image and text models: we report state-of-the-art (SOTA) performance on both temporal image classification and report generation. In the latter case, we show that prefixing the prior report substantially increases performance, again reflecting the value of prior information. We emphasise that the benefit is not restricted to temporal downstream tasks: our approach also achieves SOTA on non-temporal tasks of pneumonia detection [60] and phrase grounding [10], underscoring the value of a cleaner learning signal during VLP without needing to modify or add to the training dataset. Our contributions can be summarised as follows:

- We introduce a novel pre-training framework called BioViL-T. It leverages the temporal relationship of samples to self-supervise VLP models, making commonly used biomedical VLP models (e.g., [9,32,81]) more applicable to a wider range of downstream tasks without compromising performance on existing benchmarks.
- We develop a generic multi-image encoder that handles missing image inputs and incorporates longitudinal information without requiring explicit image registration.
- We achieve SOTA results in chest X-ray (CXR) report generation, temporal image classification, and phrase grounding downstream benchmarks by accounting for prior context in self-supervised training and fine-tuning.
- We release a new multimodal benchmark dataset, MS-CXR-T, curated by an expert radiologist. It enables benchmarking of CXR VLP models in terms of temporal semantics extracted from image and text data.

2. Related work

Vision–language processing Self-supervised VLP can significantly reduce the need for manual labels required for the training of image encoders [19,53]. The availability of large-scale paired image–text datasets has thus led to rapid development of general-purpose VLP models. Objectives include contrastive and discriminative image–text matching [40,53,69] including local variants [32,76], auto-regressive (AR) captioning [4,39,77] and multi-modal masked modelling objectives [13,40,61].

Biomedical vision–language processing Paired medical image–report datasets were originally used for supervised learning via (typically) automated label extraction from clinical reports [33,63,70]. Using such datasets, advances in general-domain self-supervised VLP have been demonstrated to benefit biomedical imaging applications [9,32,81]. Work has incorporated ideas from general-domain VLP such as the original CLIP-style cross-modal contrastive objective [81], multi-modal masking with merged co-attention on image–text representations [46], and adaptations to the data of the domain. For example, a radiology report may have sparse image-specific details, prompting a local modification to the contrastive loss enabling alignment between text tokens and image patches [32]. Domain-specific pre-training of the text model is shown to benefit biomedical VLP [9], and preferential masking of medical terms during masked language modelling (MLM) was explored [75]. Here we use a local loss and domain-specific pre-training of the text model, but did not find a benefit to preferential masking. Similarly, cross-attention [22] is used rather than merged co-attention for image-guided MLM.

Longitudinal modelling of medical images While prior images are used in unimodal supervised longitudinal analysis of medical images [37,58,68,74], temporal information has not directly been employed for self-supervision. The closest work exploits patient metadata to select positive or negative examples in unimodal contrastive learning [67,79].

Existing models typically employ either late fusion of global image representations [58,64,68,74], which can miss fine-grained localised changes [32], or explicit spatial correspondence of features, using fixed spatial grids [48] or object detection [37]. Registering image pairs is commonly used for change detection in other contexts [17,52,59], and has been applied to medical imaging [5,23]. For CXRs however, registration entails the ill-posed problem of aligning 2D projections of 3D geometry, which inevitably results in residual misalignment. Our approach does not rely on bounding boxes or explicit graph construction as it uses

\footnote{In the MIMIC-CXR v2 dataset [36], around 40% of reports explicitly reference a previous image. See Appendix B for details.}
Self-supervision across time  Self-supervision has found applications on densely-sampled time series data (e.g., video) to capture temporal information [30, 55, 78, 80]. Our problem setting involves sparsely and sporadically sampled data where temporal pretext tasks are less applicable [2]. Similarly, it requires text supervision to enable both static and temporal learning, when temporal structure is present.

3. BioViL-T training framework

Our approach comprises a multi-image encoder designed to extract spatio-temporal features from sequences of images (Section 3.1) and a text encoder incorporating optional cross-attention on image features. The models are trained jointly with image-guided MLM and cross-modal global and local contrastive objectives (Section 3.2). The resulting image and text models are later adapted for uni- or multimodal downstream tasks as described in Section 3.3. Implementation details are presented in Appendices E and F.

For a given image and report pair \( (x_{\text{img}}, x_{\text{txt}}) \), the report \( x_{\text{curr}} \) describes the current image content and changes in reference to prior images. Our proposed formulation focuses on a single prior image; however, it can be generalised to multiple prior images depending on the application. Hence, we construct datasets by including the prior image whenever it exists: \( (x_{\text{img}}, x_{\text{prior}}', x_{\text{curr}}') \in \mathcal{D}_m \) or \( (x_{\text{img}}', \emptyset, x_{\text{curr}}') \in \mathcal{D}_s \). With the resulting dataset being a union of single and multi-image examples: \( \mathcal{D} = \mathcal{D}_m \cup \mathcal{D}_s \).

3.1. Extracting spatio-temporal image features

Clinical findings are often observed across different image regions and co-occur simultaneously, which requires dense level visual reasoning across time to capture both static and temporal features. In contrast to late global fusion [64] and bounding-box based approaches [37], BioViL-T leverages local correspondences between image regions across time using transformer self-attention blocks [21]. Thus our method does not require an explicit image registration step between time points.

We propose a hybrid CNN–Transformer encoder model due to its data efficiency and spatial flexibility of cross-attention across time points: \( E_{\text{img}} : \mathbb{R}^{W \times H} \rightarrow \mathbb{R}^{W' \times H' \times D_{\text{img}}} \) (e.g., ResNet-50 [31]) and \( A_{\text{img}} : \mathbb{R}^{T \times L \times D_{\text{img}}} \rightarrow \mathbb{R}^{L \times D_{\text{img}}} \) (e.g., transformer [21]), where \( W, H, \) and \( T \) correspond to spatio-temporal dimensions, \( L = W'H' \) is the number of visual tokens per image, and \( D_{\text{img}} \) is the embedding dimension. Here \( E_{\text{img}} \) serves as a stem network [51] to provide visual token features of individual images. The CNN’s inductive biases [24, 51] ensure data efficiency of our hybrid model, making it ideal for smaller scale biomedical datasets. \( E_{\text{img}} \) is initialised with BioViL weights [9]. The main purpose of \( A_{\text{img}} \) is to capture patch embedding interactions across time when a prior image \( x_{\text{prior}}' \) is available and to aggregate them into a fixed-length token representation. Input visual tokens, \( H_0 = H_{\text{curr}} = E_{\text{img}}(x_{\text{img}}), H_0 = E_{\text{img}}(x_{\text{prior}}') \) are augmented with spatio-temporal positional encodings and flattened across the spatial dimensions. They are then processed by \( K \) transformer encoder [66] layers \( A \) as follows:

\[
\begin{bmatrix}
    H_{k+1}^\text{curr} \\
    H_{k+1}^\text{prior}
\end{bmatrix} = A_k\left(\begin{bmatrix}
    H_k^\text{curr} + S + 1_L \otimes t_{\text{curr}} \\
    H_k^\text{prior} + S + 1_L \otimes t_{\text{prior}}
\end{bmatrix}\right),
\]

for \( k = 1, \ldots, K \), where \( S \in \mathbb{R}^{L \times D_{\text{img}}} \) denotes 2D sinusoidal positional encodings [12] and \( T = [t_{\text{curr}}; t_{\text{prior}}] \in \mathbb{R}^{2 \times D_{\text{img}}} \) is its temporal counterpart, which is learnt (Fig. 2) [4]. The layer-normalised (LN) [6] output of the final transformer encoder block \( P_{\text{diff}} : L(N(H_K^\text{curr})) \) is an ‘aggregated’ representation of patch-level progression information anchored on the current image. Figure 3 shows attention roll-out [1] applied to \( P_{\text{diff}} \) after pre-training, showing how the prior image contributes to the fused representation. Figure A.3 further highlights the robustness to variations in pose underlining that registration is not necessary for this encoder.

Static-temporal feature decomposition When a prior image is available the final image representation \( V := P_{\text{curr}} \oplus P_{\text{diff}} \in \mathbb{R}^{W' \times H' \times 2D_{\text{img}}} \) is formed by concatenating two sets of features (similar to [7]): those from the current image alone (\( P_{\text{curr}} \)) and the temporal features from current and prior images (\( P_{\text{diff}} \)). In this way, self-attention is mainly required to cope with pose variations and patch comparisons across time in extracting temporal content, removing the need for registration or explicit spatial feature alignment. When no prior scan is available (\( x \in D_s \)), \( A_{\text{img}} \) is not used and \( P_{\text{diff}} \) is replaced by a learnable token \( p_{\text{miss}} \in \mathbb{R}^{D_{\text{img}}} \), replicated across the spatial dimensions. Section 4.5 later demonstrates that \( A_{\text{img}} \) highlights the value of feature decomposition for tasks such as phrase grounding which require well-localised features [10].

Hereafter, downstream tasks that require solely single image features, \( P_{\text{curr}} \), are referred to as static tasks, and the ones that benefit from additional progression information, \( P_{\text{diff}} \), as temporal tasks, e.g., report decoding.

3.2. Text-supervision for spatio-temporal learning

Let \( w = (w_1, \ldots, w_M) \) denote a vector of \( M \) tokens of a report \( x_{\text{txt}} \) after tokenisation. We first obtain contextualised token features \( E_{\text{txt}}(w) \in \mathbb{R}^{M \times D_{\text{txt}}} \) by passing a sequence of text tokens \( w = (w_1, \ldots, w_M) \) through a BERT encoder \( E_{\text{txt}}[20] \). The input sequence is prepended with either a [CLS] or [MLM] token associated with a downstream training objective, conditioning the output features
similar to [39, 42]. During training, we do two forward passes through $E_{\text{txt}}$: once with masking at 45% probability (for the MLM objective) and once without masking for contrastive learning, as shown in Figure 2. The text encoder is initialised with the weights of CXR-BERT$^3$ [9] canonical model, trained on domain-specific vocabulary and corpora.

Both text and image features are later projected into a joint latent space with $\phi_{\text{txt}} : \mathbb{R}^{D_{\text{txt}}} \rightarrow \mathbb{R}^{D}$, and similarly $v_{w,h}^{\text{proj}} = \phi_{\text{img}}(v_{w,h})$ where $\phi_{\text{img}} : \mathbb{R}^{D_{\text{img}}} \rightarrow \mathbb{R}^{D}$, with $\phi$ being a two-layer perceptron in our experiments.

**Contrastive objectives** Let $r := [E_{\text{txt}}(w)]_{[\text{CLS}]}$ denote the global representation of $w$, with $r_{\text{proj}}^{\text{img}} := \phi_{\text{img}}(r)$ its projected version. Given projected patch embeddings $v_{w,h}^{\text{proj}}$, we can compute a global cosine similarity $S_C(\tilde{v}^{\text{proj}}, r^{\text{proj}})$ and a local similarity using weighted pairwise cosine similarities across text tokens and projected patch embeddings [32, 76]. These similarities are used in both global and local contrastive objectives with the InfoNCE loss [49, 53]. The local loss proves crucial both for static phrase-grounding and temporal image classification (see Table 7), highlighting the importance of localised self-supervision.

**Image-guided masked language modelling** Prior work [9, 46] has shown that biomedical visual-language learning benefits from an auxiliary task such as MLM since capturing the joint distribution of tokens can stabilise and improve language understanding during joint learning. Given a batch $B$ of token vectors $w$, it is often defined as the cross-entropy for predicting the randomly sampled masked tokens, $m \in \{1, \ldots, M\}$, $\mathcal{L}_{\text{MLM}} = \frac{-1}{|B|} \sum_{w \in B} \log p_{\theta}(w_m \mid w_{m})$, where $\theta$ are the weights of the text encoder $E_{\text{txt}}$.

In the absence of image information, however, certain masked findings and attributes are not readily predicted, e.g., “[MASK]” is worsening”. As shown in the general domain [13], visual information can help disambiguate such masked predictions and provide additional cross-modal supervision. Thus, we use cross-attention [22,66] to the image features $v_{w,h}^{\text{proj}}$ during this task. Specifically, for our image-guided MLM objective we model $p_{\theta}(w_m \mid w_{m}, v_{w,h}^{\text{proj}})$.

### 3.3. Adaptations to downstream tasks

BioViL-T can be adapted to various downstream tasks. For phrase-grounding and zero-shot inference, we rely on $S_C(v_{w,h}^{\text{proj}}, v_{w,h}^{\text{proj}})$ similar to [9, 32]. For multiple-text prompts, projected text embeddings are marginalised prior to $\ell_2$-normalisation [53]. To enable language decoding, $v_{w,h}^{\text{proj}}$ inputs are cross-attended by text queries $w$, and causal-attention is utilised between text tokens [39,66]. Differing from [9, 32, 81], we show that report generation tasks can greatly benefit from temporal joint latent space.

**Conditioning on prior reports** In contrast to existing work, we incorporate the prior report as a prompt to contextualise the report generation task: $p_{\theta}(w_{\text{curr}} \mid E_{\text{prior}}^{\text{proj}}, w_{\text{prior}}^{\text{proj}})$.
where Φ are the multi-modal encoder–decoder network’s weights, and \( w_{\text{curr}} \), \( w_{\text{prior}} \) denote text tokens for current and prior reports respectively. This is analogous to fine-tuning GPT-3 [11] with prompts and instructions [71], but conditioning on both images and the previous report. A dedicated separation token [SEP] is added into the input sequence \([w_{\text{prior}}, \text{SEP}, w_{\text{curr}}]\).

**Curation of imaging datasets** CXR datasets [36] often contain multiple image acquisitions \( Z = \{x_1^{\text{img}}, \ldots, x_N^{\text{img}}\} \) in a single visit due to data quality issues such as a limited field-of-view or scanning the wrong body part (Figure A.4). Unlike [9, 32, 81], we conduct curation to choose higher quality images among the potential candidates instead of performing a random selection. For this step, a separate BioViL-T is trained on ‘clean’ studies with single acquisitions and later used in a zero-shot setting to detect out-of-distribution samples [26, 27] arising from the re-imaging process. The candidate \( \hat{z} \) is selected as follows: 
\[
\hat{z} = \arg \max_{z \in Z} S_C(\psi^{\text{proj}}_z, \hat{r}^{\text{proj}})
\]  
subject to \( |s_z - S_{\text{proj}}^{\text{mean}}| > \delta \) for a margin \( \delta \). This approach is applied to enhance the quality of the temporal classification dataset given its limited size.

### 4. Datasets & experiments

Here, we demonstrate BioViL-T’s data efficiency and adaptability to a wide range of applications, and show how the model achieves SOTA performance on various downstream tasks by learning from data instances linked across time, making effective use of domain priors and the available training data. Specifically, our model is evaluated on a diverse set of downstream tasks including zero- and few-shot and temporal image classification, report generation, phrase-grounding [10], and sentence similarity.

**MS-CXR-T benchmark** We release a new multi-modal benchmark dataset\(^4\), MS-CXR-T, to evaluate chest X-ray VLP models on two distinct temporal tasks: image classification and sentence similarity. The former comprises multi-image and ground-truth label pairs \((N = 1326)\) across 5 findings, with classes corresponding to 3 states of disease progression for each finding: \{Improving, Stable, Worsening\}. The latter quantifies the temporal-semantic similarity of text embeddings extracted from pairs of sentences \((N = 361)\). The pairs can be either paraphrases or contradictions in terms of disease progression. The data for both tasks was manually annotated and reviewed by a board certified radiologist. Appendix C provides further details on its data distribution and annotation protocol.

**Datasets** For pre-training, we use the MIMIC-CXR v2 [28, 36] chest X-ray dataset, which contains longitudinal imaging studies with corresponding radiological reports, see Fig. B.1 for the distribution of studies. We only use frontal view scans and discard samples where reports do not contain an IMPRESSION section. From this data, we gather 174.1k and 4.9k text-image pairs for training and validation respectively, with a majority of pairs including a prior image: \(|D_{\text{train}}^{\text{txt}}| = 118.8k, |D_{\text{train}}^{\text{img}}| = 55.3k\). The text consists of the IMPRESSION section and, for MLM additionally the FINDINGS section if available. Note that no manual labels are used during pre-training and no additional data is used for the methods that leverage the link between current and prior images. For early stopping we track the validation loss, see Appendix E for implementation details.

Downstream evaluations are performed on a disjoint held-out test set shared across all tasks, \(|D_{\text{test}}| = 2971\). For report generation, we extend this test set with samples from healthy subjects \((N = 815)\) to match the prevalence of pathological studies used in prior work [14, 25, 45]. For fine-tuning on temporal image classification, we use labels from the Chest ImaGenome dataset [72] as in [37] (statistics in Table F.2). In detail, we use the following benchmark datasets: (I) MS-CXR [10] for phrase grounding, (II) the RSNA Pneumonia dataset [60, 70] to test zero-shot and fine-tuned classification, (III) MS-CXR-T for temporal sentence similarity and temporal image classification.

**Comparison approaches** We compare our approach to other domain-specific SOTA pre-training frameworks [9, 32] specifically on phrase-grounding and zero-shot predictive performance. The non-temporal BioViL framework [9] is most similar to our approach and provides insight into non-temporal pre-training. We additionally compare to internal ablations such as removing the past report during report generation and masking prior images during phrase grounding. For SOTA performance comparison, various AR and nearest-neighbour (NN) based language decoding approaches are used as baselines: IFCC [45], R2Gen [14], CXR-RePaiR-2 [25], and CXR-RePaiR-Select [25].

For the temporal classification task, we compare against a baseline exploiting the BioViL image encoder [9], and an
Table 1. Results for report generation task: Predictions are evaluated in terms of lexical (BLEU-4, ROUGE) and factuality metrics (CHEXBERT, TEM). Approaches are grouped into two broad categories: nearest-neighbour (NN) and auto-regressive (AR). BioViL-T pre-training consistently yields improved decoding. Further, the consistent performance gains of using prior image and report demonstrate the importance of such domain priors. ‘PI / PR’ indicate usage of prior image and report, respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>Pre-training</th>
<th>PI / PR</th>
<th>BLEU-4</th>
<th>ROUGE</th>
<th>CHEXBERT</th>
<th>TEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>CXR-RePaR-S [25]</td>
<td>BioViL</td>
<td>✓ / ✓</td>
<td>5.1</td>
<td>14.3</td>
<td>28.1</td>
<td>12.5</td>
</tr>
<tr>
<td>Proposed (NN)</td>
<td>BioViL-T</td>
<td>✓ / ✓</td>
<td>4.5</td>
<td>20.5</td>
<td>29.0</td>
<td>13.0</td>
</tr>
<tr>
<td>Baseline (AR) [9]</td>
<td>BioViL</td>
<td>/</td>
<td>7.5</td>
<td>27.0</td>
<td>25.3</td>
<td>13.8</td>
</tr>
<tr>
<td>Proposed</td>
<td>BioViL-T</td>
<td>✓ / ✓</td>
<td>9.2</td>
<td>23.0</td>
<td>31.7</td>
<td>17.3</td>
</tr>
</tbody>
</table>

Table 2. Temporal image classification results (repeated for 4 random seeds) on the MS-CXR-T benchmark for fully-supervised and zero-/few-shot (Z&F) learning settings, in terms of macro-accuracy across the three classes for each finding. Affine registration is performed for the baseline method (denoted with suffix ‘w/reg’), to partially address the pose variations across scans.

<table>
<thead>
<tr>
<th>Method</th>
<th>Pre-train Consolation</th>
<th>PL effusion</th>
<th>Paucimientos</th>
<th>Paucimientos Edema</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioViL-T [9]</td>
<td>Temporal</td>
<td>53.6 ± 1.9</td>
<td>59.7 ± 2.1</td>
<td>56.9 ± 3.9</td>
</tr>
<tr>
<td>[2] Bioll [9]</td>
<td>Temporal</td>
<td>59.7 ± 2.4</td>
<td>62.4 ± 1.4</td>
<td>60.1 ± 2.1</td>
</tr>
<tr>
<td>CNN + Transformer [25]</td>
<td>Temporal</td>
<td>44.0 ± 2.0</td>
<td>61.3 ± 1.6</td>
<td>45.1 ± 3.5</td>
</tr>
<tr>
<td>[3] CheXNet [37]</td>
<td>Temporal</td>
<td>56.1 ± 1.5</td>
<td>62.3 ± 1.1</td>
<td>59.4 ± 1.0</td>
</tr>
<tr>
<td>BioViL [9]</td>
<td>Static</td>
<td>56.1 ± 1.5</td>
<td>63.6 ± 0.9</td>
<td>60.2 ± 0.7</td>
</tr>
<tr>
<td>BioViL-T (10%)</td>
<td>Temporal</td>
<td>59.9 ± 1.7</td>
<td>65.5 ± 0.3</td>
<td>61.5 ± 2.2</td>
</tr>
<tr>
<td>BioViL-T (10%)</td>
<td>Temporal</td>
<td>61.1 ± 2.4</td>
<td>67.0 ± 0.8</td>
<td>61.9 ± 1.9</td>
</tr>
</tbody>
</table>

4.1. Temporal pre-training yields data efficiency

The sections ‘NN’ and ‘Z&F’ on Tables 1 and 2 report zero- and few-shot performance on tasks benefitting from temporal information: temporal image classification and report generation. Here we measure the quality of the learnt joint latent space and the extent to which BioViL-T enables efficient use of raw data. For zero-shot classification we prompt the AR fine-tuned model with prefix: ‘[FINDING] is’ and compare the next-token probability of words meaning ‘improving’, ‘stable’, and ‘worsening’ (Appendix F.4).

Without using any labelled data, Table 2 shows that the proposed AR-based approach already yields performance superior to prior fully-supervised work [37] on temporal image classification. With only 10% of labels, classification fine-tuning provides a further boost, indicating that BioViL-T produces a multi-image encoder readily adapted to temporal tasks. Similarly, in a zero-shot report-retrieval setting, the findings show that compared to temporally-agnostic pre-training, BioViL-T leveraging prior images improves across all metrics. Consistent with prior work [25], the retrieved reports already preserve factuality with high CheXbert scores, more-so than the other metrics which measure fine-grained specifics of phrasing. This demonstrates that the latent space captures the high-level semantics of the clinical features. Fine-grained phrasing however will be substantially improved by AR fine-tuning.

4.2. Achieving SOTA performance with BioViL-T

A wide range of downstream tasks benefit substantially from temporally-aware pre-training.

Through downstream adaptations and fine-tuning our model, we report SOTA performance on report generation and temporal image classification tasks. For the former, using both prior images and reports during fine-tuning substantially improves across metrics (Table 1). In particular, TEM metric results show that temporal context is key for accurately describing change in the generated report while avoiding hallucinations (see Table A.1 for examples). Comparing to published results on a comparable test split and...
Table 4. Image classification results on RSNA Pneumonia Detection Benchmark [60] for train and test splits of 70% – 30% respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>% of Labels</th>
<th>Supervision</th>
<th>Acc.</th>
<th>F1</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLoRIA [32]</td>
<td>✗</td>
<td>Zero-shot</td>
<td>0.70</td>
<td>0.58</td>
<td></td>
</tr>
<tr>
<td>BioViL [9]</td>
<td>✗</td>
<td>Zero-shot</td>
<td>0.732</td>
<td>0.665</td>
<td>0.831</td>
</tr>
<tr>
<td>BioViL-T</td>
<td>✗</td>
<td>Zero-shot</td>
<td>0.805</td>
<td>0.706</td>
<td>0.871</td>
</tr>
<tr>
<td>BioViL [9] 1%</td>
<td>✓</td>
<td>Few-shot</td>
<td>0.805</td>
<td>0.723</td>
<td>0.881</td>
</tr>
<tr>
<td>BioViL-T 1%</td>
<td>✓</td>
<td>Few-shot</td>
<td>0.814</td>
<td>0.730</td>
<td>0.890</td>
</tr>
</tbody>
</table>

metrics (Sec. 4.1), we conclude that BioViL-T with fine-tuning achieves SOTA on report generation, producing reports that are lexically on par with prior work but substantially more factually accurate. Note that we do ‘vanilla’ AR fine-tuning to focus on the impact of the pre-trained encoders, so application-specific supervision [45] could be used in conjunction to further boost performance.

In temporal image classification (Tab. 2), BioViL-T pre-training outperforms the non-temporal baseline (BioViL) and improves on previously-reported results [37] by up to 20 percentage points (pp). Furthermore, baseline methods that rely on image registration (BioViL w/reg), underperform compared to the proposed approach. Further analysis reveals that errors tend to be in cases with disagreement between radiologists (Appendix A.2). We also note that pre-training is critical for a hybrid CNN-transformer model on this task, likely due to the small labelled dataset. Lastly, cur- ration of temporal training data is observed to improve the classification results by .08 pp aggregated across the find- ings, see Appendix A.4 for details.

4.3. Static tasks benefit from temporal learning

**BioViL-T broadens the range of applicable downstream tasks whilst contributing to performance on static tasks.**

In this section, we demonstrate that performance improvements afforded by BioViL-T are not restricted to tem- poral tasks – static tasks also benefit. Table 4 reports results on zero- and few-shot pneumonia classification from single images [60], where BioViL-T establishes a new SOTA compared to prior work [9, 32].

We see a similar trend on the MS-CXR phrase grounding benchmark (Tab. 5). This task can be solved with single images, however we show that the inclusion of the prior image (where available) does not impair the performance of BioViL-T. Feature decomposition effectively preserves localised information from the current image.

4.4. Towards better sentence embedding quality

**Language models acquire increased temporal sensitivity.**

We hypothesise that text encoders learn temporal semantics through supervision from longitudinal image series. To verify this, RadNLI [45] and MS-CXR-T datasets are used in a zero-shot binary classification setting. Cosine similarity of sentence pair embeddings [56] are treated as class-logits to label each pair either as paraphrase or contradiction. See Appendix F6 for further details.

Our text model is benchmarked against SOTA domain-specific BERT models. Table 6 shows that the proposed framework greatly increases the sensitivity of sentence embeddings to temporal content whilst better capturing the static content (RadNLI). Note that CXR-BERT-Specialised [9] is learnt through single-images starting from the same canonical model, illustrating the substantial increase in tem- poral and static sensitivity due to BioViL-T pre-training.

### 4.5. Ablation experiments

In Table 7 we report extensive ablations across the multi- image encoder architecture, pre-training choices, and AR fine-tuning for report generation.

**Image encoder** Table 7 shows that decomposition of static and progression features is essential to ensure good performance on single-image tasks, such as phrase grounding. For temporal representations, on the other hand, positional encodings (T) are essential to disambiguate the order of scans, i.e., permutation variance across time.

**Model pre-training** The corresponding results are shown in the middle section of Table 7. The local contrastive loss proves crucial to ensure meaningful language supervision during pre-training, followed by the image-guided MLM objective. Lastly, use of the FINDINGS section results in only minor performance gains as the key findings are al- ready captured in the IMPRESSION section.

**Report generation** The importance of prior image and report is demonstrated by the substantial drop in the “no prior image and report” ablation, confirming our hypothesis that temporal context is crucial for improving report quality. While both inputs are crucial for optimal performance,
the prior report is more so because it summarises the image and provides a clearer signal. The prior image however cannot be dismissed entirely as it provides granular details which may not always be documented in a report. Finally, we found the separation token is crucial in differentiating between the predicted tokens for the current report and tokens from the prior report.

4.6. Which tokens require a prior image in MLM?

We leverage the MLM objective in an inference setting to analyse the influence of prior images in predicting masked tokens. Inspired by the $\Delta$ image loss of [8], we define $\Delta^\text{prior}_\text{img}$ as the change in loss by conditioning the estimation with a prior image for a given token $w$ as follows:

$$\Delta^\text{prior}_\text{img}(w) = l(w, x^\text{curr}_\text{img}, \emptyset) - l(w, x^\text{curr}_\text{img}, x^\text{prior}_\text{img})$$

where $l(w, x^\text{curr}_\text{img}, x^\text{prior}_\text{img})$ is the cross-entropy of predicting the masked token $w$ given visual features (MLM loss for a single token), averaged over sentences in which $w$ appears. $\Delta^\text{prior}_\text{img}$ is a measure of how much that token benefits from access to the prior image, as well as an assessment of the contribution of the prior image to the image representation. In Figure 4 we show the distribution of $\Delta^\text{prior}_\text{img}$ as a function of token category (e.g., Anatomy, Positional; see F.5 for annotation details). For Progression-type terms in particular, the model heavily relies on the prior image for image-guided MLM. We further observe that this effect is specific to temporal tokens; as expected, those from other semantic categories do not consistently rely on the prior image.

5. Conclusion

In this paper, we introduced BioViL-T, a vision-language pre-training framework enabling alignment between text and multiple images. BioViL-T makes use of a novel multi-image encoder and explicitly decomposes static-temporal features to augment the current image representation with information from prior images. This enables the grounding of temporal references in the text. To our knowledge, this is the first method capable of leveraging the temporal content commonly present in biomedical text. It addresses an important limitation in existing VLP approaches, which simply discard such context. Also, incorporating such multi-modal temporal content provides strong learning signals to the model, resulting in richer representations and improved downstream performance.

We demonstrate the value of this paradigm through extensive experiments: BioViL-T excels on both static and temporal tasks, establishing new SOTA on report generation, temporal image classification, few/zero-shot pneumonia detection, and phrase grounding. Furthermore, we release a new multi-modal benchmark (MS-CXR-T) to measure the quality of image and text representations in terms of temporal semantics, enabling more diverse evaluation of biomedical VLP models. The corresponding model weights and code are publicly available.

Further exploration and evaluation are required on diverse datasets to characterise what kinds of tasks would benefit from a temporal modelling approach, and specifically from the proposed methodology.

Acknowledgements: We would like to thank Hannah Richardson, Hoifung Poon, Melanie Bernhardt, Melissa Bristow and Naoto Usuyama for their valuable feedback.

---

6Models can be found at: https://aka.ms/biovil-t-model

7Code can be found at: https://aka.ms/biovil-t-code
References


[23] Stanley Durrleman, Xavier Pennec, Alain Trouvé, José Braga, Guido Gerig, and Nicholas Ayache. Toward a com-


[35] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In International Conference on Machine Learning, pages 4904–4916. PMLR, 2021. 1, 13


