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†

Microsoft Health Futures

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. 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-

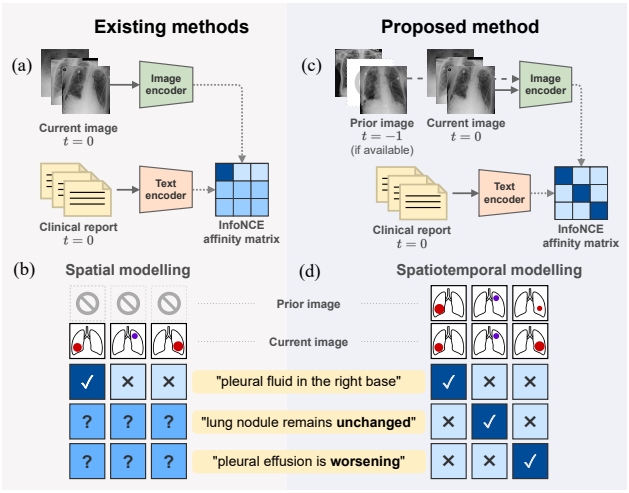

Figure 1. (a) Existing visual–language pre-training approaches [9, 32, 81] often use only a single image for contrastive learning (e.g., InfoNCE [49]). (b) In such settings, discarding the temporal connectivity of images limits the alignment of image–text pairs as shown with the affinity matrix, leading to suboptimal pre-training and missed opportunity to create additional model supervision for free. (c, d) Our approach exploits this domain knowledge by learning to incorporate a series of images and correlate them to reports, leading to pre-trained models that can generalise to a wider range of downstream tasks whilst achieving SOTA performance.

*These authors contributed equally.
†Corresponding author: ozan.oktay@microsoft.com

edge of this prior image[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.

[1] In the MIMIC-CXR v2 dataset [36], around *40% of reports* explicitly reference a previous image. See Appendix B for details.

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

self-attention of visual tokens across time to handle any spatial misalignment.

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 multi-modal downstream tasks as described in Section 3.3. Implementation details are presented in Appendices E and F.

For a given image and report pair $(\mathbf{x}_{\text{img}}^{\text{curr}}, \mathbf{x}_{\text{txt}}^{\text{curr}})$, the report $\mathbf{x}_{\text{txt}}^{\text{curr}}$ 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[2]: $(\mathbf{x}_{\text{img}}^{\text{curr}}, \mathbf{x}_{\text{img}}^{\text{prior}}, \mathbf{x}_{\text{txt}}^{\text{curr}}) \in \mathcal{D}_m$ or $(\mathbf{x}_{\text{img}}^{\text{curr}}, \varnothing, \mathbf{x}_{\text{txt}}^{\text{curr}}) \in \mathcal{D}_s$ 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 spatiotemporal dimensions, $L = W'H'$ is the number of visual tokens per image, and D_{img} is the embedding dimension. Here E_{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_{img} is initialised with BioViL weights [9]. The main purpose of A_{img} is to capture patch embedding interactions across time when a prior image $\mathbf{x}_{\text{img}}^{\text{prior}}$ is available and to aggregate them into a fixed-length token representation. Input visual tokens, $\mathbf{H}_0^{\text{curr}} = \mathbf{P}^{\text{curr}} := E_{\text{img}}(\mathbf{x}_{\text{img}}^{\text{curr}})$, $\mathbf{H}_0^{\text{prior}} := E_{\text{img}}(\mathbf{x}_{\text{img}}^{\text{prior}})$ 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:

$$\begin{bmatrix} \mathbf{H}_k^{\text{curr}} \\ \mathbf{H}_k^{\text{prior}} \end{bmatrix} = A_k\left(\begin{bmatrix} \mathbf{H}_{k-1}^{\text{curr}} + \mathbf{S} + \mathbf{1}_L \otimes \mathbf{t}^{\text{curr}} \\ \mathbf{H}_{k-1}^{\text{prior}} + \mathbf{S} + \mathbf{1}_L \otimes \mathbf{t}^{\text{prior}} \end{bmatrix}\right), \quad (1)$$

for $k = 1, \ldots, K$, where $\mathbf{S} \in \mathbb{R}^{L \times D_{\text{img}}}$ denotes 2D sinusoidal positional encodings [12] and $\mathbf{T} = [\mathbf{t}^{\text{curr}}; \mathbf{t}^{\text{prior}}] \in \mathbb{R}^{2 \times D_{\text{img}}}$ is its temporal counterpart, which is learnt (Fig. 2) [4]. The layer-normalised (LN) [6] output of the final transformer encoder block $\mathbf{P}^{\text{diff}} := \text{LN}(\mathbf{H}_K^{\text{curr}})$ is an 'aggregated' representation of patch-level progression information anchored on the current image. Figure 3 shows attention roll-out [1] applied to $\mathbf{P}^{\text{diff}}$ 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 $\mathbf{V} := \mathbf{P}^{\text{curr}} \oplus \mathbf{P}^{\text{diff}} \in \mathbb{R}^{W' \times H' \times 2D_{\text{img}}}$ is formed by concatenating two sets of features (similar to [7]): those from the current image alone ($\mathbf{P}^{\text{curr}}$) and the temporal features from current and prior images ($\mathbf{P}^{\text{diff}}$). 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 ($\mathbf{x} \in \mathcal{D}_s$), A_{img} is not used and $\mathbf{P}^{\text{diff}}$ is replaced by a learnable token $\mathbf{p}^{\text{miss}} \in \mathbb{R}^{D_{\text{img}}}$, replicated across the spatial dimensions. Section 4.5 later demonstrates that A_{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, $\mathbf{P}^{\text{curr}}$, are referred to as *static tasks*, and the ones that benefit from additional progression information, $\mathbf{P}^{\text{diff}}$, as *temporal tasks*, e.g., report decoding.

3.2. Text-supervision for spatio-temporal learning

Let $\mathbf{w} = (w_1, \ldots, w_M)$ denote a vector of M tokens of a report $\mathbf{x}_{\text{txt}}$ after tokenisation. We first obtain contextualised token features $E_{\text{txt}}(\mathbf{w}) \in \mathbb{R}^{M \times D_{\text{txt}}}$ by passing a sequence of text tokens $\mathbf{w} = (w_1, \ldots, w_M)$ through a BERT encoder E_{txt} [20]. The input sequence is prepended with either a `[CLS]` or `[MLM]` token associated with a downstream training objective, conditioning the output features

[2] The prior *report* is not included during pre-training as it may further reference an earlier study, reintroducing temporal ambiguity.

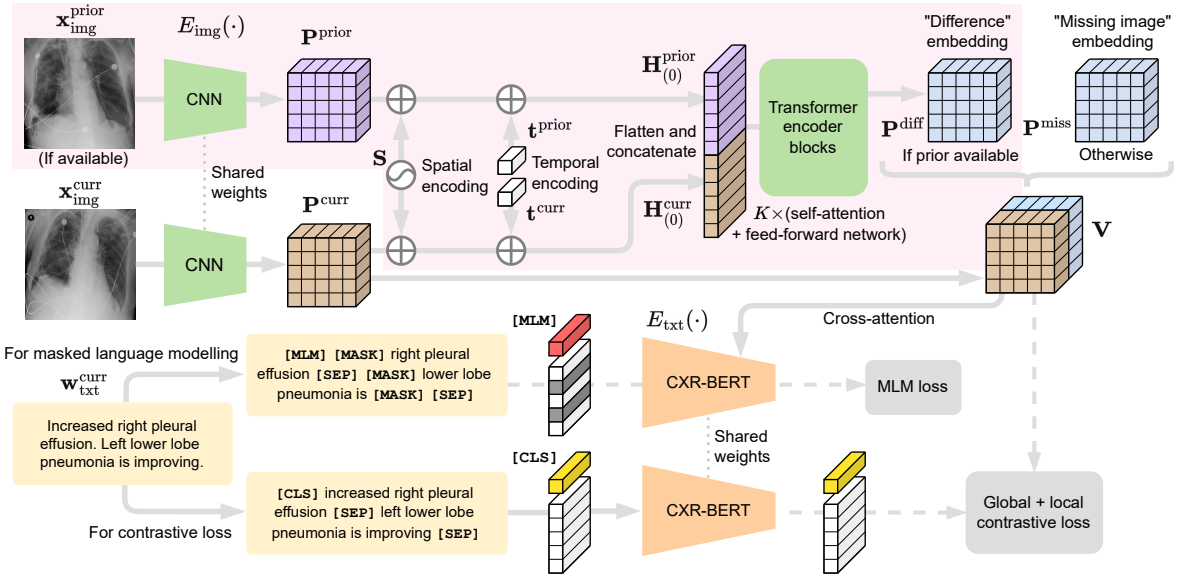

Figure 2. The proposed self-supervised VLP training framework BioViL-T: Image representations $\mathbf{V}$ are extracted from single and multiple input scans (whenever available) using a hybrid CNN and transformer encoder [24, 51]. This design choice is to increase the data-efficiency and enable the fusion of temporal content without requiring image registration. They are later matched with their corresponding text representations obtained with CXR-BERT [9] using local [32] and global InfoNCE [49] training objectives. As an additional model supervision, multi-modal fused representations, obtained with cross-attention, are used for image-guided masked language modelling.

similar to [39, 42]. During training, we do two forward passes through E_{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 $\mathbf{v}_{w,h}^{\text{proj}} := \phi_{\text{img}}(\mathbf{v}_{w,h})$ where $\phi_{\text{img}} : \mathbb{R}^{D_{\text{img}}} \rightarrow \mathbb{R}^{D}$, with ϕ being a two-layer perceptron in our experiments.

Contrastive objectives Let $\mathbf{r} := [E_{\text{txt}}(\mathbf{w})]_{\texttt{[CLS]}}$ denote the global representation of $\mathbf{w}$, with $\mathbf{r}^{\text{proj}} := \phi_{\text{txt}}(\mathbf{r})$ its projected version. Given projected patch embeddings $\mathbf{v}_{w,h}^{\text{proj}}$, we can compute a global cosine similarity $S_C(\bar{\mathbf{v}}^{\text{proj}}, \mathbf{r}^{\text{proj}})$ 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 $\mathcal{B}$ of token vectors $\mathbf{w}$, it is often defined as the cross-entropy for predicting the randomly sampled masked tokens, $m \subset \{1, \ldots, M\}$, $\mathcal{L}_{\text{MLM}} = -\frac{1}{|\mathcal{B}|} \sum_{\mathbf{w} \in \mathcal{B}} \log p_\theta(\mathbf{w}_m \,|\, \mathbf{w}_{\setminus m})$, where θ are the weights of the text encoder E_{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 $\mathbf{v}_{w,h}^{\text{proj}}$ during this task. Specifically, for our image-guided MLM objective we model $p_\theta(\mathbf{w}_m \,|\, \mathbf{w}_{\setminus m}, \mathbf{v}_{w,h}^{\text{proj}})$.

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 $S_C(\mathbf{r}^{\text{proj}}, \mathbf{v}_{w,h}^{\text{proj}})$ similar to [9, 32]. For multiple-text prompts, projected text embeddings are marginalised prior to ℓ_2-normalisation [53]. To enable language decoding, $\mathbf{v}_{w,h}^{\text{proj}}$ inputs are cross-attended by text queries $\mathbf{w}$, 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: $p_\Phi(\mathbf{w}_{\text{txt}}^{\text{curr}} | \mathbf{w}_{\text{txt}}^{\text{prior}}, \mathbf{v}_{w,h}^{\text{proj}})$,

[3] https://huggingface.co/microsoft/BiomedVLP-CXR-BERT-general

where Φ are the multi-modal encoder–decoder network's weights, and $\mathbf{w}_{\text{txt}}^{\text{curr}}$, $\mathbf{w}_{\text{txt}}^{\text{prior}}$ 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 $[\mathbf{w}_{\text{txt}}^{\text{prior}}, \texttt{[SEP]}, \mathbf{w}_{\text{txt}}^{\text{curr}}]$.

Curation of imaging datasets CXR datasets [36] often contain multiple image acquisitions $\mathcal{Z} = \{\mathbf{x}_1^{\text{img}}, \ldots, \mathbf{x}_Z^{\text{img}}\}$ 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: $\hat{z} = \arg\max_{z \in \mathcal{Z}} S_C(\bar{\mathbf{v}}_z^{\text{proj}}, \mathbf{r}^{\text{proj}})$ s.t. $|s_{\hat{z}} - s_{\mathcal{Z}\setminus\hat{z}}| > \delta$ for a margin δ. 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 static 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: $|\mathcal{D}_m^{\text{train}}| = 118.8k$, $|\mathcal{D}_s^{\text{train}}| = 55.3k$. 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, $|\mathcal{D}^{\text{test}}| = 2971$. 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.

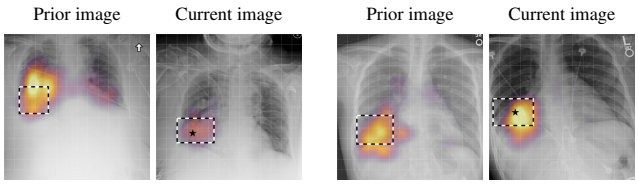

Figure 3. Attention rollout maps [1] from the reference patch (marked with $\star$) to the current and prior images. The bounding boxes, annotated by a radiologist, show the extent of consolidation. Note that the reference patch attends to its anatomical neighbourhood in the prior image despite the misalignment between prior and current images. The grid (14×14) represents the patch tokens processed in the transformer encoder blocks.

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

[4] *MS-CXR-T* benchmark dataset can be accessed through PhysioNet: https://aka.ms/ms-cxr-t

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.

	Method	Pre-training	PI / PR	BLEU-4	ROUGE	CHEXBERT	TEM
NN	CXR-RePaiR-2 [25]	BioViL	✗ / ✗	2.1	14.3	28.1	12.5
	Baseline (NN) [9]	BioViL	✗ / ✗	3.7	20.0	28.3	11.1
	Proposed (NN)	BioViL-T	✓ / ✗	4.5	20.5	29.0	13.0
AR	Baseline (AR) [9]	BioViL	✗ / ✗	7.5 ± 0.1	27.9 ± 0.1	29.3 ± 0.3	13.8 ± 0.1
	Proposed	BioViL-T	✓ / ✗	8.2 ± 0.1	28.7 ± 0.1	30.2 ± 0.7	16.0 ± 0.3
	Proposed	BioViL-T	✓ / ✓	**9.2 ± 0.3**	**29.6 ± 0.1**	**31.7 ± 1.0**	**17.5 ± 0.1**

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.

	Method (% of labels)	Pre-train	Consolidation	Pl. effusion	Pneumonia	Pneumothorax	Edema
Z&F	BioViL-T prompt (0%)	Temporal	53.6 ± 1.9	59.7 ± 2.1	58.0 ± 3.9	34.9 ± 1.0	64.2 ± 1.5
	BioViL-T (10%)	Temporal	59.7 ± 2.4	62.4 ± 1.4	60.1 ± 2.1	35.3 ± 2.6	62.6 ± 1.7
Supervised	CNN + Transformer	ImageNet	44.0 ± 2.0	61.3 ± 1.6	45.1 ± 3.5	31.5 ± 3.1	65.5 ± 1.1
	CheXRelNet [37]	ImageNet	47	47	47	36	49
	BioViL [9]	Static	56.1 ± 1.5	62.3 ± 1.1	59.4 ± 1.0	41.7 ± 2.8	67.5 ± 0.8
	BioViL w/reg [9]	Static	56.0 ± 1.5	63.0 ± 0.9	60.2 ± 0.7	42.5 ± 2.7	67.5 ± 0.9
	BioViL-T wout curation	Temporal	58.9 ± 1.7	65.5 ± 0.7	61.5 ± 2.2	44.4 ± 2.1	67.4 ± 0.8
	BioViL-T	Temporal	**61.1 ± 2.4**	**67.0 ± 0.8**	**61.9 ± 1.9**	42.6 ± 1.6	**68.5 ± 0.8**

approach that makes use of graph convolutions across regions of interest extracted from bounding boxes [37]. For BioViL, we perform affine image registration (with 4 DoF) for each pair of scans to cope with pose variations, and the encoded images are concatenated along the feature dimension and classified via a multilayer perceptron. For [37], we compare to the three-class setting. Lastly, we benchmark our final text model in isolation against domain specific SOTA models in a temporal sentence similarity task: CXR-BERT [9] and PubMedBert [29].

Metrics Due to class imbalance, we report macro-accuracy for temporal image classification. For phrase grounding, we use mean Intersection-Over-Union (mIoU) and Contrast-to-Noise-Ratio (CNR) [9]. The latter measures the discrepancies between cosine similarities inside and out of the bounding box region without requiring hard thresholds. To evaluate the quality of generated reports, we use both the standard lexical metrics, e.g., BLEU [50], ROUGE-L [41], and also domain-specific factuality metric: CheXbert[5] [62]. To directly probe the generation of change-related information, we introduce a new metric called temporal entity matching (TEM) to compute the match score of a fixed set of temporal entities (see Appendix D).

[5]The average of the weighted-F_1 score across 14 pathological observations labelled by CheXbert.

Table 3. Report generation results using the same train/test splits from [25], measured by lexical (BLEU-2) and factuality (CHEXBERT) metrics. Baseline results were also collected from [25]. Note the CHEXBERT score covers all 14 observations.

Method	Decoded sections	BLEU-2	CHEXBERT
R2gen [14]	Findings & Impression	21.20 ± 0.10	14.80 ± 0.30
IFCC [45]	Findings	21.70 ± 0.10	27.00 ± 0.40
CXR-RePaiR-Sel [25]	Impression	5.00 ± 0.10	27.40 ± 0.30
BioViL-T	Impression	15.86 ± 0.14	34.83 ± 0.73
BioViL-T	Findings & Impression	21.31 ± 0.19	**35.86 ± 0.35**

4.1. Temporal pre-training yields data efficiency

Downstream tasks are enabled with minimal labels.

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.

Method	% of Labels	Supervision	Acc.	F1	AUROC
GLoRIA [32]	✗	Zero-shot	0.70	0.58	-
BioViL [9]	✗	Zero-shot	0.732	0.665	0.831
BioViL-T	✗	Zero-shot	**0.805**	**0.706**	**0.871**
BioViL [9]	1%	Few-shot	0.805	0.723	0.881
BioViL-T	1%	Few-shot	**0.814**	**0.730**	**0.890**

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, curation of temporal training data is observed to improve the classification results by .68 pp aggregated across the findings, 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 temporal 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 F.6 for further details.

Table 5. Results on *MS-CXR* benchmark [10] (5-runs with different seeds), "Multi-image" column indicates the input images used at test time.

Method	Multi-Image	Avg. CNR	Avg. mIoU
BioViL [9]	✗	1.07 ± 0.04	0.229 ± 0.005
+ Local loss [9, 32]	✗	1.21 ± 0.05	0.202 ± 0.010
BioViL-T	✗	**1.33 ± 0.04**	**0.243 ± 0.005**
BioViL-T	✓	**1.32 ± 0.04**	**0.240 ± 0.005**

Table 6. Results on *MS-CXR-T* sentence similarity benchmark.

	MS-CXR-T (361 pairs)		**RadNLI** (145 pairs)	
Text Model	**Accuracy**	**ROC-AUC**	**Accuracy**	**ROC-AUC**
PubMedBERT [29]	60.39	.542	81.38	.727
CXR-BERT-G [9]	62.60	.601	87.59	.902
CXR-BERT-S [9]	78.12	.837	89.66	.932
BioViL-T	**87.77 ± 0.5**	**.933 ± .003**	90.52 ± 1.0	**.947 ± .003**

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 temporal 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 already 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,

Table 7. Ablation study on image encoder, pre-training settings, and report generation (one component at a time, and repeated for 4 random seeds). Note that for temporal classification, linear probing is applied to frozen image embeddings. In report generation, the baseline method is fine-tuned with both prior image and report.

	Ablation	Avg. CNR (mIoU)	Pl. Effusion Acc.
Encoder	Baseline	1.33 ± 0.02 (.248)	64.8 ± 0.6
	– Temporal pos. encoding	1.32 ± 0.02 (.242)	62.9 ± 1.0
	– Feature decomposition	1.11 ± 0.08 (.203)	64.0 ± 0.6
Pre-training	Baseline	1.33 ± 0.02 (.248)	64.8 ± 0.6
	– Use of findings section	1.32 ± 0.01 (.246)	63.8 ± 0.8
	– MLM loss	1.28 ± 0.02 (.238)	63.2 ± 0.7
	– Local contrastive loss	1.18 ± 0.02 (.236)	60.2 ± 0.6
	Ablation	**ROUGE**	**TEM**
Report gen.	Baseline	29.64 ± 0.08	17.54 ± 0.11
	– Prior image	29.35 ± 0.25	16.30 ± 0.40
	– Prior report	28.67 ± 0.12	16.00 ± 0.30
	– (Prior image and report)	27.78 ± 0.09	13.65 ± 0.48
	– Separation token	26.00 ± 0.40	15.50 ± 1.06

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 Δ *image loss* of [8], we define $\Delta_{\text{img}}^{\text{prior}}$ as the change in loss by conditioning the estimation with a prior image for a given token w as follows:

$$\Delta_{\text{img}}^{\text{prior}}(w) = l(w, \mathbf{x}_{\text{img}}^{\text{curr}}, \varnothing) - l(w, \mathbf{x}_{\text{img}}^{\text{curr}}, \mathbf{x}_{\text{img}}^{\text{prior}}) \quad (2)$$

where $l(w, \mathbf{x}_{\text{img}}^{\text{curr}}, \mathbf{x}_{\text{img}}^{\text{prior}})$ 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. $\Delta_{\text{img}}^{\text{prior}}$ 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 $\Delta_{\text{img}}^{\text{prior}}$ 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.

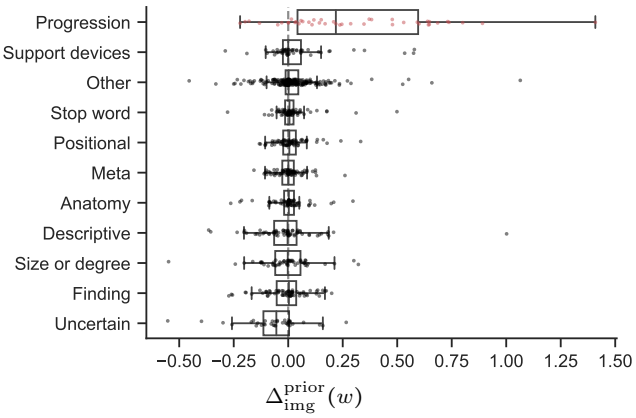

Figure 4. Mean token-level increase in image-guided MLM loss when prior image is discarded, grouped by token category. The prior image is excluded during inference to measure its impact on masked token predictions. *Progression* tokens are significantly better predicted when prior images are incorporated into image embeddings. The top five *Progression* tokens are 'persist', 'improving', 'remains', 'unchanged', and 'residual'.

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[6] and code[7] 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.

[6]Models can be found at: https://aka.ms/biovil-t-model
[7]Code can be found at: https://aka.ms/biovil-t-code

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online, July 2020. Association for Computational Linguistics. 3, 5, 13, 14

[2] Monica N Agrawal, Hunter Lang, Michael Offin, Lior Gazit, and David Sontag. Leveraging time irreversibility with order-contrastive pre-training. In *International Conference on Artificial Intelligence and Statistics*, pages 2330–2353. PMLR, 2022. 3

[3] Uwa O. Aideyan, Kevin Berbaum, and Wilbur L. Smith. Influence of prior radiologic information on the interpretation of radiographic examinations. *Academic Radiology*, 2(3):205–208, 1995. 1

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 2, 3

[5] B.B. Avants, C.L. Epstein, M. Grossman, and J.C. Gee. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41, 2008. Special Issue on The Third International Workshop on Biomedical Image Registration – WBIR 2006. 2

[6] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3

[7] Nadine Behrmann, Mohsen Fayyaz, Juergen Gall, and Mehdi Noroozi. Long short view feature decomposition via contrastive video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9244–9253, 2021. 3

[8] Yonatan Bitton, Gabriel Stanovsky, Michael Elhadad, and Roy Schwartz. Data efficient masked language modeling for vision and language. In *2021 Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021*, pages 3013–3028. Association for Computational Linguistics (ACL), 2021. 8

[9] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the most of text semantics to improve biomedical vision–language processing. In *European Conference on Computer Vision (ECCV)*, pages 1–21. Springer, 2022. 1, 2, 3, 4, 5, 6, 7, 17, 18, 19, 20

[10] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. MS-CXR: Making the most of text semantics to improve biomedical vision–language processing (version 0.1). PhysioNet, 2022. 2, 3, 5, 7, 15

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 5, 19

[12] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3

[13] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: UNiversal Image-TExt Representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 2, 4

[14] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Nov. 2020. 5, 6

[15] Ching-Yao Chuang, R Devon Hjelm, Xin Wang, Vibhav Vineet, Neel Joshi, Antonio Torralba, Stefanie Jegelka, and Yale Song. Robust contrastive learning against noisy views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16670–16681, 2022. 1

[16] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020. 1

[17] R. C. Daudt, B. L. Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 4063–4067, 2018. 2

[18] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. 15

[19] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11162–11173, 2021. 1, 2

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 3, 13

[22] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022. 2, 4

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

prehensive framework for the spatiotemporal statistical analysis of longitudinal shape data. *International Journal of Computer Vision*, 103:22–59, 2013. 2

[24] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pages 2286–2296. PMLR, 2021. 3, 4

[25] Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y Ng, and Pranav Rajpurkar. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Machine Learning for Health*, pages 209–219. PMLR, 2021. 5, 6, 18

[26] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pretrained model clip. In *Proceedings of the AAAI conference on artificial intelligence*, 2022. 5

[27] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34:7068–7081, 2021. 5

[28] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. 5

[29] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021. 6, 7, 20

[30] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Advances in Neural Information Processing Systems*, 33:5679–5690, 2020. 3

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3

[32] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951, 2021. 1, 2, 4, 5, 7

[33] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI conference on artificial intelligence*, 33(01):590–597, 2019. 2

[34] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, Curtis Langlotz, et al. Radgraph: Extracting clinical entities and relations from radiology reports. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 17

[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

[36] A. Johnson, T. Pollard, S.J. Berkowitz, R. Mark, and S. Horng. MIMIC-CXR database (version 2.0.0). PhysioNet, 2019. 2, 5

[37] Gaurang Karwande, Amarachi B Mbakwe, Joy T Wu, Leo A Celi, Mehdi Moradi, and Ismini Lourentzou. Chexrelnet: An anatomy-aware model for tracking longitudinal relationships between chest x-rays. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022*, pages 581–591, 2022. 2, 3, 5, 6, 7

[38] Jiann-Shu Lee, Jing-Wein Wang, Hsing-Hsien Wu, and Ming-Zheng Yuan. A nonparametric-based rib suppression method for chest radiographs. *Computers & Mathematics with Applications*, 64(5):1390–1399, 2012. 20

[39] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR, 17–23 Jul 2022. 2, 4

[40] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 1, 2

[41] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. 6

[42] Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*, 2018. 4

[43] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 17, 19

[44] Bradley C Lowekamp, David T Chen, Luis Ibáñez, and Daniel Blezek. The design of simpleitk. *Frontiers in neuroinformatics*, 7:45, 2013. 20

[45] Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304, Online, June 2021. Association for Computational Linguistics. 5, 6, 7, 18

[46] Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics*, 2022. 2, 4

[47] American College of Radiology (ACR). ACR practice guideline for communication of diagnostic imaging findings. *Practice guidelines & technical standards*, 2020. 1

[48] Dong Yul Oh, Jihang Kim, and Kyong Joon Lee. Longitudinal change detection on chest X-rays using geometric correlation maps. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 748–756. Springer, 2019. 2

[49] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1, 4

[50] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. 6

[51] Namuk Park and Songkuk Kim. How do vision transformers work? In *International Conference on Learning Representations*, 2022. 3, 4

[52] Daifeng Peng, Yongjun Zhang, and Haiyan Guan. End-to-end change detection for high resolution satellite images using improved UNet++. *Remote Sensing*, 11(11):1382, 2019. 2

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 4

[54] Vignav Ramesh, Nathan Andrew Chi, and Pranav Rajpurkar. Improving radiology report generation systems by removing hallucinated references to non-existent priors. *arXiv preprint arXiv:2210.06340*, 2022. 2

[55] Adria Recasens, Pauline Luc, Jean-Baptiste Alayrac, Luyu Wang, Florian Strub, Corentin Tallec, Mateusz Malinowski, Viorica Pătrăucean, Florent Altché, Michal Valko, et al. Broaden your views for self-supervised video learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1255–1265, 2021. 3

[56] Nils Reimers, Iryna Gurevych, Nils Reimers, Iryna Gurevych, Nandan Thakur, Nils Reimers, Johannes Daxenberger, Iryna Gurevych, Nils Reimers, Iryna Gurevych, et al. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 671–688. Association for Computational Linguistics, 2019. 7, 20

[57] Liqa A Rousan, Eyhab Elobeid, Musaab Karrar, and Yousef Khader. Chest x-ray findings and temporal lung changes in patients with covid-19 pneumonia. *BMC Pulmonary Medicine*, 20(1):1–9, 2020. 1

[58] Ruggiero Santeramo, Samuel Joseph Withey, and G. Montana. Longitudinal detection of radiological abnormalities with time-modulated LSTM. In *MICCAI 2018 Workshop on Deep Learning in Medical Imaging Analysis*, 2018. 2

[59] Wenzhong Shi, Min Zhang, Rui Zhang, Shanxiong Chen, and Zhao Zhan. Change detection based on artificial intelligence: State-of-the-art and challenges. *Remote Sensing*, 12(10):1688, 2020. 2

[60] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019. 2, 5, 7

[61] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15617–15629, 2022. 2

[62] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, Online, Nov. 2020. Association for Computational Linguistics. 6

[63] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*, 2020. 2, 16

[64] Anuroop Sriram, Matthew Muckley, Koustuv Sinha, F. Shamout, Joelle Pineau, K. Geras, L. Azour, Y. Aphinyanaphongs, N. Yakubova, and William H. Moore. COVID-19 prognosis via self-supervised representation learning and multi-image prediction. *arXiv preprint arXiv:2101.04909*, 2021. 2, 3

[65] Philippe Thévenaz and Michael Unser. Optimization of mutual information for multiresolution image registration. *IEEE transactions on image processing*, 9(12):2083–2099, 2000. 20

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 4

[67] Yen Nhi Truong Vu, Richard Wang, Niranjan Balachandar, Can Liu, Andrew Y Ng, and Pranav Rajpurkar. Medaug: Contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation. In *Machine Learning for Healthcare Conference*, pages 755–769. PMLR, 2021. 2

[68] Chuang Wang, Andreas Rimner, Yu chi Hu, Neelam Tyagi, Jue Jiang, Ellen Yorke, Sadegh Riyahi, Gig S. Mageras, Joseph O. Deasy, and Pengpeng Zhang. Towards predicting the evolution of lung tumors during radiotherapy observed

on a longitudinal MR imaging study via a deep learning algorithm. *Medical Physics*, 2019. 2

[69] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, 2021. 2

[70] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. ChestX-Ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2097–2106. IEEE Computer Society, 2017. 2, 5

[71] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2021. 5

[72] Joy Wu, Nkechinyere Agu, Ismini Lourentzou, Arjun Sharma, Joseph Paguio, Jasper Seth Yao, Edward Christopher Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo Anthony Celi, Tanveer Syeda-Mahmood, and Mehdi Moradi. Chest imagenome dataset (version 1.0.0). PhysioNet, 2021. 5, 15, 18

[73] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10033–10041, 2021. 19

[74] Yiwen Xu, Ahmed Hosny, Roman Zeleznik, Chintan Parmar, Thibaud P. Coroller, Idalid Ivy Franco, Raymond H. Mak, and Hugo J.W.L. Aerts. Deep learning predicts lung cancer treatment response from serial medical imaging. *Clinical Cancer Research*, 25:3266 – 3275, 2019. 2

[75] Bin Yan and Mingtao Pei. Clinical-BERT: Vision-language pre-training for radiograph diagnosis and reports generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):2982–2990, 2022. 2

[76] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 2, 4

[77] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 1, 2

[78] Sukmin Yun, Jaehyung Kim, Dongyoon Han, Hwanjun Song, Jung-Woo Ha, and Jinwoo Shin. Time is matter: Temporal self-supervision for video transformers. In *International Conference on Machine Learning*, pages 25804–25816. PMLR, 2022. 3

[79] Dewen Zeng, John N Kheir, Peng Zeng, and Yiyu Shi. Contrastive learning with temporal correlated medical images: A case study using lung segmentation in chest x-rays. In *2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, pages 1–7. IEEE, 2021. 2

[80] Zhaoyang Zeng, Daniel McDuff, Yale Song, et al. Contrastive learning of global and local video representations. *Advances in Neural Information Processing Systems*, 34:7025–7040, 2021. 3

[81] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020. 1, 2, 4, 5

[82] Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D Manning, and Curtis P Langlotz. Biomedical and clinical english model packages for the stanza python nlp library. *Journal of the American Medical Informatics Association*, 28(9):1892–1899, 2021. 16, 17, 18