Comprehensive language–image pre-training for 3D medical image understanding

Tassilo Wald[1,2,†], Ibrahim Ethem Hamamci[1,†], Yuan Gao[1,†], Sam Bond-Taylor[1], Harshita Sharma[1], Maximilian Ilse[1], Cynthia Lo[1], Olesya Melnichenko[1], Anton Schwaighofer[1], Noel C. F. Codella[1], Maria Teodora Wetscherek[1,3], Klaus H. Maier-Hein[2,4], Panagiotis Korfiatis[5], Valentina Salvatelli[1], Javier Alvarez-Valle[1], and Fernando Pérez-García[1]

[1] Microsoft
[2] German Cancer Research Center (DKFZ)
[3] Department of Radiology, University of Cambridge and Cambridge University Hospitals NHS Foundation Trust
[4] Pattern Analysis and Learning Group, Heidelberg University Hospital
[5] Department of Radiology, Mayo Clinic

Abstract. In the 3D medical image domain, vision–language pre-training is used to create vision–language encoders (VLEs) that can support radiologists by retrieving patients with similar abnormalities, predicting likelihoods of abnormality, or, with downstream adaptation, generating radiological reports. While the methodology holds promise, three challenges limit the capabilities of current 3D VLEs: data scarcity due to privacy concerns, high computational costs resulting from the volumetric nature of the images, and a domain shift between the long reports used for training and the short prompts used during inference for, *e.g.*, zero-shot classification. As a consequence, natural-image VLE recipes do not directly transfer to 3D medical imaging.

In this paper, we overcome these challenges by injecting additional supervision via a report generation objective and combining vision–language with vision-only pre-training, allowing us to leverage both image-only and paired image–text 3D datasets. Further, we propose a novel loss that addresses the domain shift between long reports and short textual prompts. Through these additional objectives, paired with best practices of the 3D medical imaging domain, we develop the **Co**mprehensive **L**anguage–**I**mage **Pr**e-trai**n**ing (COLIPRI) encoder family. Our COLIPRI encoders achieve state-of-the-art performance in report generation, semantic segmentation, classification probing, and zero-shot classification.

The model weights and inference code are freely available at `https://huggingface.co/microsoft/colipri`.

Keywords: language-image pre-training · 3D medical image analysis

† Work done during an internship at Microsoft Research.

1 Introduction

Contrastive language–image pre-training (CLIP) [28] has established itself as one of the strongest paradigms to learn general-purpose image and text representations.

Aside from being a solid starting point for adaptation to downstream tasks of interest [11,20], having language-aligned vision embeddings allows leveraging natural language for open-set classification [28] and open-set segmentation [48]. In 3D medical imaging, this training paradigm is especially relevant as every image acquired in a clinical setting is accompanied by a report, making this kind of paired data abundant in hospitals. The reports associated with the images offer (weak) supervision and enable zero-shot tasks for clinical support, such as image-to-report retrieval or open-set abnormality classification.

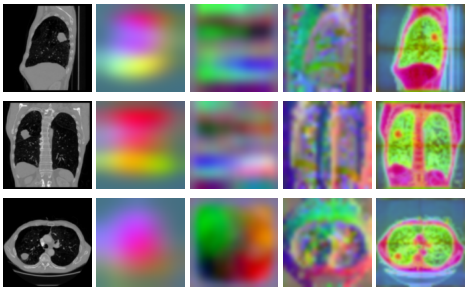
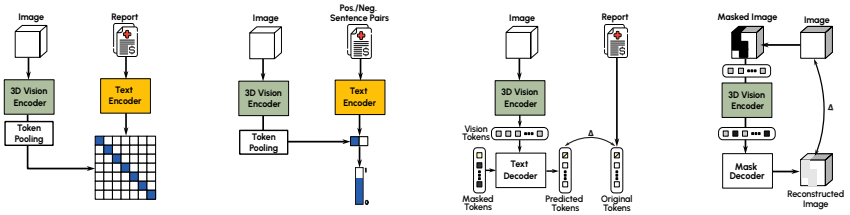


Fig. 1: Principal component analysis (PCA) maps of dense 3D features obtained from (a) a CT scan with a lung mass using different encoders. Compared to the baseline methods (b)(c)(d), our COLIPRI-CRM encoder (e) generates sharper and more coherent features. The dense embeddings have sizes $7 \times 7 \times 10$ (b), $24 \times 24 \times 8$ (c), $24 \times 24 \times 24$ (d) and $48 \times 48 \times 48$ (e), and are interpolated here using bicubic interpolation for visualisation purposes. The sagittal (top), coronal (middle) and axial (bottom) slices of the 3D scan are centred on the lung mass. The Comprehensive Language–Image Pre-training (COLIPRI) (e) map was computed from the input resampled to 1 mm isotropic, cropped and padded to $384 \times 384 \times 384$, using a sliding window of size $192 \times 192 \times 192$.

Despite these promises, the field of vision–language pre-training with medical images is not as mature as its general-domain counterpart. While CLIP [28], Perception Encoder [5], or SigLIP 2 [35] are well established in the general domain, 3D medical VLEs have only recently started to garner attention [3,12]. We believe that this can be attributed to two key issues: 1) the lack of large, publicly available, paired datasets in the medical domain and 2) domain- and modality-specific methodological and engineering hurdles.

The largest publicly available 3D datasets with image–report pairs reach a combined size of about 78k image–report pairs, which is far from the 400 million image–text pairs which CLIP [28] was trained on, and even further from the 12 billion alt-texts used to train SigLIP 2 [35]. However, the relatively large number of available unpaired images has the potential to substantially increase the amount of overall usable data, when leveraged through self-supervised learning (SSL). See Tab. S1 in the supplementary material for more information.

The number of voxels in 3D medical images is orders of magnitude larger than the number of pixels in 2D images from the general or medical domains.



(a) Contrastive image–report alignment **(b)** Opposite sentence loss **(c)** Radiology report generation **(d)** Masked autoencoder

Fig. 2: Combining a contrastive image–report pre-training, b the novel opposite sentence loss, c radiology report generation, and d masked autoencoders (MAEs) yields various versions of our Comprehensive Language–Image Pre-training (COLIPRI) vision–language encoders (VLEs).

For example, a typical chest CT volume may be composed of $512 \times 512 \times 320$ voxels, making it difficult to use entire images in native resolution during training due to excessive VRAM requirements. It is therefore common to train with image crops [27, 29] or downsampled images [3, 12]. Cropping may complicate CLIP training as clinical reports refer to the entire volume, not just the field of view (FOV) in the subvolume, and downsampling dismisses high-resolution information which might be crucial for detecting small abnormalities. Lastly, the reports of 3D images are substantially longer than the short prompts typically used during inference for, *e.g.*, zero-shot classification, which results in a large domain shift that needs to be addressed.

In this work, we overcome the aforementioned challenges, advancing the state of the art in 3D medical vision–language models. Our contributions can be summarised as follows:

1. To reduce the domain shift between long reports used for training and short textual prompts used for zero-shot classification, we introduce the novel Opposite Sentence Loss (OSL), which improves results substantially by closing the gap between domains.
2. To address the data scarcity, we increase supervision from medical reports by adding a text generation objective, and introduce a vision-only objective so that large image-only datasets can be leveraged for complementary SSL pre-training.
3. We comprehensively evaluate the resulting models through zero-shot classification, classification probes, report generation, retrieval, and semantic segmentation, demonstrating that our COLIPRI-CRM models represent the new state of the art.
4. We share model weights, code for inference, and multiple usage examples at https://huggingface.co/microsoft/colipri.

2 Related work

Contrastive vision–language models such as CLIP [28] have demonstrated that large-scale paired image–text data enables transferable multimodal representations. Later variants improved the contrastive objective [45], or proposed a language-generation objective [36, 40], with others including additional self-supervised objectives [21, 23, 35]. In parallel to this, self-supervised learning (SSL) methods [30, 49] were developed to learn general purpose features from unpaired images, achieving strong general-purpose visual features in natural imaging.

Compared to natural imaging, 3D medical imaging incurs much higher computational costs due to its higher dimensionality and faces stricter data access constraints, limiting the overall amount of available data. These factors have hindered progress of 3D image SSL, with a recent benchmark [39] showing that features generated by most 3D SSL methods [32, 42, 50] improve either dense or global downstream tasks, but generally not both. For dense downstream tasks, MAEs [22, 38] remain the strongest pre-training method, yielding features that adapt well for 3D semantic segmentation.

Medical VLEs have largely been trained on (2D) chest X-rays with paired reports [2, 4, 15, 41]. Early work on 3D medical images focused on adapting contrastive language–image pre-training to 3D settings [3, 12], incorporate anatomy awareness [29], or auxiliary supervision through chest X-rays [6]. Despite increased interest in the field of 3D SSL [25, 43] and vision–language pre-training, a research gap remains between natural imaging and 3D medical imaging. Thus far, no work has combined self-supervised and vision–language objectives, nor has any research in the 3D imaging domain introduced a text generation objective.

3 Method

We present Comprehensive Language–Image Pre-training (COLIPRI), a multimodal pre-training framework that learns joint representations between 3D medical images and associated radiology reports. COLIPRI combines contrastive vision–language alignment, increases supervision of limited data through radiology report generation (RRG), extends available data to unpaired images through masked image modelling (MIM), and closes the long-report-to-short-zero-shot-prompt distribution shift through our novel Opposite Sentence Loss (OSL) to produce 3D vision–language representations applicable to both global reasoning and dense prediction tasks (Fig. 2). We develop our models in stages, each adding a learning objective to assess the effects of each building block.

3.1 Report preprocessing

Radiological reports often contain details about all organs imaged, stating whether the findings are normal or abnormal. Should abnormalities be present, they are explicitly named; however, when absent, the abnormalities are often not listed, so as not to yield an excessive list of abnormalities a patient does not suffer from.

The exceptions are typically abnormalities that might have prompted the imaging study in the first place. This style of reporting results in reports of substantial token sequence lengths, with the average *Findings* section of CT-RATE [12] being 243 tokens long when using the CXR-BERT tokeniser [4] (Fig. S1 in the supplementary material). This is in stark contrast to Zhang *et al*. [47] or Radford *et al*. [28], which either sample a single sentence from the paired caption or whose datasets contain single-sentence captions.

To enable the objectives introduced in Secs. 3.2 and 3.3, we use a large language model (LLM) to restructure each report into eight clinical subsections, denoted as C, assigning every sentence to one subsection. We also leverage the LLM to create concise versions of all sentences s and to label each as a *positive finding* (pathology present) s^+ or *negative finding* (normal) s^-. These preprocessed reports enable 1) language regularisation, 2) the OSL, and 3) our radiological report generation formulation. Details, prompts and examples are provided in the supplementary material (Sec. A.3).

3.2 Adapting CLIP to 3D radiology

COLIPRI leverages the Primus-M transformer [37] architecture as the image encoder, and CXR-BERT [4] (pre-trained on chest X-rays) as the text encoder. The dense tokens of the vision and text encoders are aggregated through a multi-head attention-pooling mechanism to yield global representations. Due to different embedding dimensions of image and text encoders, the image embeddings are projected to the lower-dimensional text space, which can then be aligned using CLIP [28].

In contrast to natural images, radiological reports describe findings across the entire scan, requiring a sufficiently large FOV for global context. However, token sequence length increases cubically with FOVs in 3D imaging, making training computationally expensive. We use inputs of size $160 \times 160 \times 160$ with an isotropic spacing of 2 mm, balancing FOV and token length for efficiency. We train with the default CLIP objective, which outperforms a sigmoid loss [46] in our setting.

The *Findings* section in radiology reports often resembles a long list, making the text encoder prone to overfitting to sentence order or stylistic cues rather than semantics. To counter this, we introduce two complementary text augmentations: 1) the *Sentence Shuffle* augmentation randomly permutes sentences within a report, discouraging reliance on order; 2) the *Short Sentence* augmentation replaces long-form reports with LLM-shortened statements with a given probability during training. This reduces the domain shift generated by training on long reports and running inference using brief zero-shot prompts (*e.g.*, "{abnormality} present" or "No {abnormality} present") but does not fully close it. Additional ablations and detailed hyperparameters are provided in Sec. B.1 of the supplementary material.

Closing the zero-shot domain gap Using the LLM-structured, shortened, and categorised reports (Sec. 3.1), we close the training–inference domain shift

through our novel Opposite Sentence Loss (OSL). Let x_i denote the image for case i, and let s_i^+ denote a *positive findings* sentence sampled from the report r_i. We create s_i^-, a negation of s_i^+, via the template "`No {sentence}.`". Since s_i^+ originates from the report of case i then the sentence s_i^+ correctly describes the image, yielding label $y_i = 1$. Additionally, for any clinical subsection of the current report that contains no positive findings, we sample positive findings sentences from the same subsection in other reports. These sampled positive findings do not describe the current image and therefore serve as negative examples, paired with their negated counterparts to form (s_i^+, s_i^-) pairs with supervision label $y_i = 0$. Let e_i^+ and e_i^- be the text embeddings of the positive and negated statement pair, and v_i the global image embedding. We compute a cosine similarity between image and text embeddings, followed by a temperature-scaled softmax over the pair and optimise a cross-entropy loss between probability p_i^+ and label y_i to maximise $sim(e_i^+ \mid v_i)$ if s_i^+ describes the image and minimises $sim(e_i^+ \mid v_i)$ if s_i^+ does not describe the image (described in Secs. B.4 and Alg. S1 of the supplementary material). This trains the model in a short-prompt manner and with statements of abnormality presence and absence, as used during zero-shot inference, explicitly learning to disambiguate presence vs. absence of findings. We combine our OSL objective with the CLIP objective, $\mathcal{L}_{\text{align}} = 0.5 \cdot \mathcal{L}_{\text{CLIP}} + 0.5 \cdot \mathcal{L}_{\text{OSL}}$, yielding our intermediate COLIPRI-C method (Tab. S8 in supplementary material).

3.3 Increasing supervision from limited data by generating reports

To enrich the visual representations generated by the encoder, we introduce a radiology report generation (RRG) objective. This task aims to ensure that the vision encoder captures the complete semantic content of the associated report, rather than only discriminative cues required for contrastive pairing. Following Tschannen *et al.* [36], we evaluated two decoding modes during training: *causal captioning* and *parallel captioning*. In causal captioning mode, the decoder operates autoregressively, predicting each token conditioned on all preceding tokens, $p_{\text{causal}}(y_t \mid y_{<t}, \mathbf{v})$, where $\mathbf{v}$ denotes the dense visual features and $y_{<t}$ the previously generated tokens. This setup enables sequential report generation, akin to standard language modelling. In contrast, the parallel (masked) mode predicts all tokens simultaneously from a fully masked input, $p_{\text{parallel}}(y_t \mid \mathbf{M}, \mathbf{v})$, where $\mathbf{M}$ denotes a mask applied to all text tokens. This formulation forces the text decoder to use the dense visual representations $\mathbf{v}$ rather than previously generated text $y_{<t}$. We found that *parallel captioning* yields better results, and hence discarded *causal captioning* in our final COLIPRI models (see Tab. S6 in the supplementary material).

We implement RRG using a lightweight EVA-02 [10] transformer decoder connected to the vision encoder via cross-attention. To mitigate order ambiguity inherent to free-form clinical text, we use the LLM-structured reports introduced in Sec. 3.1, where each sentence is assigned to one of eight clinical subsections. These structured reports reduce positional uncertainty by providing a consistent ordering of findings across studies, ensuring that the decoder learns to model clinical content rather than author-specific stylistic variation. During decoding,

clinical subsection headers are kept unmasked to guide generation and preserve report semantics. The RRG objective is optimised jointly with the contrastive and OSL objectives: $\mathcal{L}_{\text{VLM}} = 0.5 \cdot \mathcal{L}_{\text{CLIP}} + 0.5 \cdot \mathcal{L}_{\text{OSL}} + \lambda_{\text{RRG}}\mathcal{L}_{\text{RRG}}$ on paired image–report batches, yielding our intermediate COLIPRI-CR method (see Tab. S8 in supplementary material).

3.4 Expanding available data through vision-only self-supervision

While multimodal pre-training provides strong representations for global reasoning tasks, such methods rely on scarce image–report pairs and often underperform on dense, spatially localised tasks [24]. To increase available data to unpaired images and to improve the localised performance of the learnt representations, we include a vision-only masked autoencoder (MAE) objective [13].

Given the lack of advanced dense SSL methods such as iBOT [49] or DINOv3 [30] in 3D medical imaging, MAE pre-training provides a straightforward and effective choice. It reconstructs missing volumetric patches from context, encouraging the model to learn dense feature representations that complement the global semantic alignment achieved by the CLIP and RRG objectives. To provide a strong initialisation, we pre-train a Primus-M vision encoder first, leveraging the `nnSSL` framework [39], before joint multimodal training.

We implement MAE as a lightweight decoder attached to the vision encoder to reconstruct masked patches in voxel space using a mean-squared-error loss. To avoid interference of the vision-only masking with the vision–language objectives, we use a training scheme that alternates between vision-only and vision–language objectives, optimising either $\mathcal{L}_{\text{Align}} = 0.5 \cdot \mathcal{L}_{\text{CLIP}} + 0.5 \cdot \mathcal{L}_{\text{OSL}}$ or $\mathcal{L}_{\text{VO}} = \lambda_{\text{MAE}}\mathcal{L}_{\text{MAE}}$. This design allows benefitting from paired supervision when available, while still leveraging the often much larger image-only datasets such as NLST [33].

Finally, since the vision-only objective does not require a global FOV, we sample high-resolution (1 mm isotropic) and default-resolution (2 mm isotropic) sub-crops during MAE pre-training. This exposes the encoder to finer spatial detail, reducing the shift between pre-training and high-resolution downstream data. The resulting combination of COLIPRI-C with the MAE objective yields our COLIPRI-CM VLE. Further, combining all of the introduced objectives (CLIP, OSL, RRG, and MAE) yields our final VLE, COLIPRI-CRM (see Tab. S8 in supplementary material).

4 Experiments

We pre-train our COLIPRI models on CT-RATE [12] (for the vision-language and vision-only objectives) and NLST [33] (for the vision-only objective). Dataset and training details are presented in Secs. A and B of the supplementary material. The resulting encoders are evaluated on multiple unimodal (semantic segmentation, multilabel classification) and multimodal (zero-shot classification, report generation, retrieval) downstream tasks.

4.1 Classification and report generation

Classification performance is evaluated on a withheld test set of CT-RATE and on the publicly available subset of RAD-ChestCT [9], which comprises 3.6k chest CT volumes with 16 multi-abnormality labels that can be derived from the original CT-RATE abnormality classes. We evaluate linear classification probes, as well as zero-shot classification performance on both datasets (see Sec. C.2 in the supplementary material).

We also evaluate the quality of the frozen image encoder embeddings for report generation. To do this, we follow the LLaVA framework [19], with image tokens passed through a two-layer multilayer perceptron (MLP) to integrate them into the language space of the Qwen2.5 1B base model [34]. We fine-tune the LLM to generate the *Findings* section of each report (see details in Sec. C.4 of the supplementary material). To evaluate the clinical accuracy of generated reports, we use the text classifier trained by Hamamci *et al*. [12], based on RadBERT [44]. We also use the RadFact-CT (+/−) and RadFact-CT (+), variants of RadFact [1] with CT-specific system prompts and few-shot examples as described in Sec. C.1 of the supplementary material.

As baselines, we compared our method against established CT models, namely CT-CLIP [12], CT-FM [26] (both trained on chest CT datasets), and Merlin [3] (trained on abdominal CTs, but we found it to be a competitive baseline for chest CT report generation). For zero-shot classification, we additionally compare against fVLM [29] and BIUD [6]. We also compare our method against two 2D VLEs which included CT slices in their pre-training data: MedImageInsight (MI2) [7], a general biomedical imaging model, and Curia [8], a radiology-specific model. We evaluate linear and zero-shot classification for the two 2D models by averaging the pixel values of a volume along the axial dimension and encoding the resulting average slice (we also tried max, mean, standard deviation and median embedding pooling [7] but found that averaging intensities across slices works best). For RRG, we instead encode all axial slices independently and perform max pooling of the embeddings.

4.2 Semantic segmentation

To evaluate the quality of the vision encoder for dense tasks, we measure 3D medical image segmentation performance after fine-tuning the encoder. In this setting, we compare against a Primus-M [37] encoder trained from scratch, as well as a state-of-the-art MAE-pre-trained [39] Primus-M encoder trained on CT-RATE and NLST (see Sec. A.1 in the supplementary material). Additionally, we compare with the default nnU-Net [16] and ResEnc-L [17] CNN encoders trained from scratch as references. All training runs are conducted using the nnU-Net framework [16], with all of the encoders being fine-tuned in a short (37.5k steps) and a long (250k steps) schedule, using the learning rate schedule described in Wald *et al*. [39], with the peak learning rate reduced to 10^{-4}. To remain partially in distribution, we chose to focus on segmentation datasets with targets in the chest region or the upper abdomen, a FOV that is often

visible during pre-training. We use five-fold cross-validation for each dataset. We evaluate segmentation performance on 1) **LiTS** [31] ($N = 131$), task 3 of the Medical Segmentation Decathlon (MSD), which contains segmentations for liver and liver tumours. 2) **Lung** [31] ($N = 64$), task 6 of the MSD, which contains cases of primary lung cancers. 3) **HVS** [31] ($N = 303$), task 8 of the MSD, which focuses on segmenting hepatic vessels and adjacent tumours. 4) **KiTS23** [14] ($N = 489$), a more recent dataset focused on segmenting tumours, cysts, and kidneys. We use the Dice similarity coefficient (DSC) averaged across all foreground classes as our summary metric for segmentation performance.

5 Results and discussion

5.1 Classification probes

Table 1: Classification probing results. We compare the embedding quality of our vision encoders against publicly available baselines. Our COLIPRI-CM and COLIPRI-CRM encoders yield the best classification results, exceeding all baselines on both datasets across all metrics. The metrics for "CT-CLIP (reported)" are from Shui *et al*. [29]. Differences in performance with the metrics we computed using the released checkpoints may be due to configuration issues. Hence, we report both sets of values for fairness and clarity. We report macro average and 95% confidence intervals (CIs) based on 100 bootstrap samples. **Bold** indicates best performance for that metric, or overlapping CIs with best. AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve.

	CT-RATE		RAD-ChestCT	
Model	AUPRC	AUROC	AUPRC	AUROC
MI2 (2D)	19.70	50.52	27.94	52.25
Curia (2D)	45.93	78.01	39.23	65.73
CT-CLIP	25.96 [24.74, 27.48]	61.21 [60.02, 62.45]	28.77 [28.23, 29.49]	54.05 [53.12, 55.01]
CT-CLIP (reported)	-	75.1	-	64.7
CT-FM	53.54 [51.93, 55.02]	82.14 [81.44, 82.82]	42.41 [41.66, 43.21]	68.49 [67.97, 69.15]
MAE	53.79 [52.55, 55.39]	82.70 [82.02, 83.31]	45.24 [44.71, 46.07]	71.20 [70.60, 71.71]
Merlin	54.81 [53.19, 56.81]	82.62 [81.84, 83.30]	45.30 [44.45, 46.20]	70.91 [70.30, 71.48]
COLIPRI-C	55.13 [53.59, 56.80]	82.63 [81.87, 83.34]	46.33 [45.61, 47.37]	71.20 [70.65, 71.78]
COLIPRI-CR	57.44 [56.00, 59.02]	83.50 [82.90, 84.04]	48.28 [47.44, 49.12]	72.78 [72.16, 73.42]
COLIPRI-CM	**61.52** [60.14, 63.03]	**86.15** [85.66, 86.78]	**51.79** [50.93, 52.76]	**76.16** [75.65, 76.61]
COLIPRI-CRM	**61.28** [59.59, 63.10]	**86.12** [85.52, 86.85]	**52.55** [51.72, 53.65]	**76.57** [76.08, 77.18]

We evaluate our COLIPRI vision encoders and other baseline vision encoders using classification probes on CT-RATE and RAD-ChestCT (Tab. 1).

Our probing setup (Sec. C.2 in the supplementary material) is largely superior to the ones used in the baselines papers, yielding AUROC values of above 80% for the majority of baseline methods (vs. approx. 75% reported in the original works) as well as our own encoders. This likely originates from our training setup, which uses multiple probes with different learning rates and token pooling methods, selecting the best probe on the validation split for testing. Moreover, not all of our token aggregation schemes are linear as some include a light-weight attention pooling block, which is more flexible than a linear layer. We believe this multi-probe scheme to be more suited for comparison due to its robustness to hyperparameter selection, and its independence of the existence or quality

of global-pooling layers attached to the vision backbones. The only exception for this performance increase is the CT-CLIP encoder, which curiously performs worse in our experiments. This might be due to a potential configuration issue, hence we additionally report the results from Sui *et al.* [29] for the metrics in common.

Our COLIPRI models exceed all baselines across all metrics and datasets, with COLIPRI-CM and COLIPRI-CRM representing the strongest encoders for classification. In particular, COLIPRI-CRM increases by 6 points AUPRC and 3.5 points AUROC over the best-performing baseline Merlin. The inclusions of our RRG and MAE objectives increases classification performance, likely due to the added objectives serving as regularisation.

5.2 Zero-shot classification

Table 2: Zero-shot classification results. We report macro AUPRC and AUROC excluding 'Medical Material' and 'Lymphadenopathy' to be consistent with [29]. An extended version of this table with all abnormalities is provided in Tab. S14 in the supplementary material. Across both metrics and datasets, our VLEs exceed the state of the art using short prompts, without relying on segmentation masks at inference as fVLM does. We report mean and 95% CIs based on 100 bootstrap samples. **Bold** indicates best performance for that metric, or overlapping CIs with best.

	CT-RATE		RAD-ChestCT	
Model	AUPRC	AUROC	AUPRC	AUROC
MI2 (2D)	18.80	49.36	26.32	48.82
Curia (2D)	19.18	50.17	26.84	50.13
CT-CLIP	-	70.4*	-	63.2*
BIUD	-	71.3*	-	62.9*
Merlin	-	72.8*	-	64.4*
fVLM	-	77.8*	-	68.0*
COLIPRI-C	43.81 [42.9, 45.2]	76.25 [75.5, 77.2]	40.98 [40.3, 41.8]	69.12 [68.6, 69.7]
COLIPRI-CR	44.70 [43.7, 46.5]	76.41 [75.7, 77.2]	39.38 [38.7, 40.1]	67.30 [66.6, 68.1]
COLIPRI-CM	**50.32** [48.7, 51.8]	**80.52** [79.7, 81.4]	43.26 [42.5, 44.3]	72.05 [71.4, 72.7]
COLIPRI-CRM	**49.39** [48.3, 51.3]	**79.81** [79.2, 80.7]	**44.48** [43.7, 45.4]	**73.23** [72.6, 73.8]

*Values are taken from [29].

For all VLEs, we report the zero-shot classification performance on CT-RATE and RAD-ChestCT in Tab. 2, using a short-form prompting scheme for our models (see Sec. A.4 in the supplementary material). Values for some baselines are taken from the fVLM paper [29]. However, in their work, results are reported for only 16 of the 18 abnormalities in CT-RATE, excluding the non-localisable 'Lymphadenopathy' and 'Medical Material' as

Table 3: The Opposite Sentence Loss (OSL) enables short-form zero-shot classification. Effect of the OSL on the validation set results. Without the OSL, zero-shot classification with short prompts performs substantially worse than using a long-form, more native prompting style.

Prompt Style	Native		Short	
	AUPRC	AUROC	AUPRC	AUROC
COLIPRI-CRM	51.17	81.91	50.70	81.66
without OSL	47.48	80.48	39.08	74.09

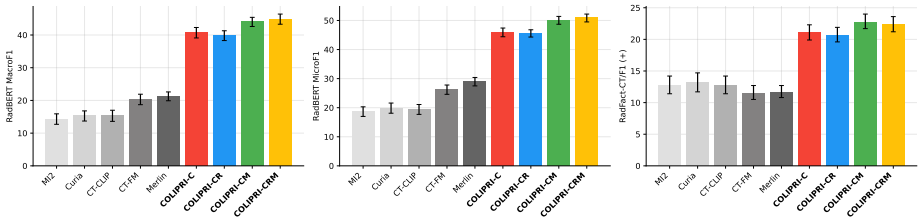


Fig. 3: Report generation results (summary). Our COLIPRI vision encoders enable generating reports substantially better than the state of the art. In particular, the LLM trained on top of our models generates more accurate statements about abnormalities being present, as measured by RadFact-CT $(+)/F_1$ and the RadBERT F_1. The exact values are presented in Tab. S15 in the supplementary material. RadFact-CT $(+)/F_1$: summary of logical precision and recall of positive findings (i.e., describing a present abnormality as judged by GPT-4o) [1]. We report median and 95% CIs (error bars) based on 500 bootstrap samples.

they are not associated with an organ[§], which is a limitation of fVLM. We report results including all abnormalities in Tab. S14 of the supplementary material.

Our COLIPRI-CM and COLIPRI-CRM encoders outperform all baselines on CT-RATE, while the COLIPRI-C and COLIPRI-CR encoders perform worse. When evaluated on RAD-ChestCT, COLIPRI-CRM generalises better than our other encoders or fVLM, with fVLM decreasing 10 points in AUROC on RAD-ChestCT compared to CT-RATE.

Including the OSL during training closes the gap between a 'native', longer-form prompting scheme where we average the embeddings of 50 randomly sampled reports with the abnormality and 50 without, versus the short-form prompts where we query with "{abnormality} present" and "no {abnormality} present" (Tab. 3). The OSL improves the utility of the encoder as short queries are substantially easier to create than the 'native', longer-form queries. We hypothesise that the OSL can generalise beyond 3D medical imaging to other domains with highly detailed reports or captions.

5.3 Report generation

We evaluate the impact of our pre-training strategy on downstream RRG (details in Sec. C.4 of the supplementary material) using both lexical and clinical metrics (Fig. 3, and Tab. S15 in the supplementary material). Across all lexical metrics (ROUGE-L, BLEU1, BLEU4, and METEOR), our COLIPRI-CRM encoder performs slightly better than the baselines, meaning that they all generate lexically accurate reports. However, high lexical overlap does not necessarily imply clinically accurate reports [18].

[§] `https://github.com/alibaba-damo-academy/fvlm/issues/12#issuecomment-3283463870`

Table 4: Report-to-image retrieval results. Retrieval results on our CT-RATE test set, after deduplicating identical reports, yielding a total sample size of $N = 1493$. CT-CLIP retrieval values are taken from Hamamci *et al.* [12].

	CT-CLIP	COLIPRI-C	COLIPRI-CR	COLIPRI-CM	COLIPRI-CRM
Recall @1	-	4.49	7.23	14.53	**15.27**
Recall @5	2.90	13.46	19.36	34.16	**35.10**
Recall @10	5.00	19.69	26.99	45.34	**46.01**

When looking at clinical metrics, our COLIPRI vision encoders outperform all baselines by a very large margin. We use F_1 scores from the RadBERT classifier and RadFact-CT (+), which measure the correctness of medical entities and factual statements. Specifically, COLIPRI-CRM improves by 23 points in RadBERT MacroF_1 and by 9 points in RadFact-CT (+)/F_1 over the strongest baseline, reflecting that the produced reports contain fewer omissions and more specific diagnostic statements. This demonstrates that our pre-training paradigm yields representations that encode more clinically relevant semantics. In absolute terms, however, the overall accuracy of generated reports for 3D medical is still low, with Macro F_1-Scores of around 45% for the abnormalities measured by RadBERT and F_1-scores of around 22% for all abnormalities as measured by RadFact-CT (+). COLIPRI-C and COLIPRI-CR exhibit slightly lower performance compared to the encoders including MAE pre-training. Interestingly, the inclusion of the RRG objective appears to have a minimal impact on overall RRG performance, despite explicitly optimising embeddings to facilitate report generation. The inclusion of the MAE objective, on the other hand, appears to substantially improve performance. As such, we hypothesise that the MAE objective, which encourages the vision encoder to learn low-level semantic representations, better supports the LLaVA framework [19]. The RRG supervision might lead to more language-aligned features, which may result in the loss of more fine-grained features. Comparing the values of RadFact-CT (+), which considers only statements mentioning the presence of abnormalities, and RadFact-CT (+/−), which considers statements about healthy organs and statements about abnormalities, reveals very different behaviour (see Tab. S15 in the supplementary material). While our encoders reach substantially better metrics in the (+) setting, some of the baselines achieve slightly higher scores than our encoders as measured in the (+/−) setting.

This is attributable to the vast imbalance of statements about presence vs. absence of abnormalities, with the latter dominating the metric. Thus, the inclusion of normal (negative) findings is an important but double-edged aspect of medical report generation. Since statements about the absence of abnormalities improve apparent completeness and boost (+/−) metrics, they can mask a low diagnostic sensitivity, as highlighted in our (+) results.

Table 5: Segmentation fine-tuning results. Mean DSC results of five-fold cross-validation trained for 37.5k steps or 250k steps. Across all datasets, our COLIPRI-CM and COLIPRI-CRM vision encoders perform on par or better than the state-of-the-art (SOTA) pre-training method MAE, when adapted for downstream segmentation. We highlight the best short and long fine-tuning results of the same architecture in bold, and second best underlined.

	References (250k steps)		**Primus-M** (37.5k steps)						**Primus-M** (250k steps)		
Dataset	**nnU-Net**	**ResEnc-L**	**Scratch**	**MAE**	**C**	**CR**	**CM**	**CRM**	**Scratch**	**MAE**	**CRM**
LiTS	80.09	81.60	74.39	<u>80.27</u>	76.77	77.09	79.97	**80.46**	79.75	<u>80.32</u>	**81.11**
Lung	70.32	70.34	62.74	67.12	65.89	65.32	<u>68.74</u>	**68.98**	**69.11**	67.46	<u>67.56</u>
HVS	68.38	67.73	64.67	**67.10**	64.64	65.08	66.58	<u>67.05</u>	65.34	<u>65.73</u>	**66.42**
KiTS23	86.04	88.17	78.90	85.34	81.02	80.71	**86.03**	<u>85.79</u>	86.25	<u>87.46</u>	**87.68**

5.4 Retrieval

Report-to-image retrieval results on our CT-RATE test set are presented in Tab. 4. All our COLIPRI VLEs are substantially better in retrieving the associated image given a report.

In particular, our COLIPRI-CRM VLE yields a R@5 of 35.1% compared against a R@5 of 2.9% of CT-CLIP, highlighting the strong multimodal alignment of our encoders. The inclusion of the RRG and MAE objectives positively benefit retrieval performance, with the MAE yielding larger performance improvements.

5.5 Segmentation

We evaluate semantic segmentation performance on the LiTS, Lung, HVS and KiTS23 datasets, against a baseline trained from scratch and an MAE pre-trained on NLST and CT-RATE at 1-mm and 2-mm isotropic resolutions, using the `nnSSL` framework [39] (Tab. 5).

In the short training regime of 37.5k steps, our encoders including MAE pre-training consistently exceed training from scratch, exceed or match the strong purely MAE pre-trained baseline. The encoders without the MAE objective also exceed training from scratch, but show lower segmentation performance, highlighting the importance of including the MAE objective for dense downstream task adaptation. This low performance of the COLIPRI-C and COLIPRI-CR models is not surprising, as it was shown that contrastive baselines struggle to surpass a baseline trained from scratch for segmentation tasks in Wald *et al.* [39].

When fine-tuning for 250k steps, our final COLIPRI-CRM model exceeds the pure MAE pre-training and improves for LiTS and KiTS, but decreases for Lung and HVS relative to the short fine-tuning schedule. For the Lung dataset, training from scratch yields better results than using pre-training. This is not uncommon as fine-tuning for longer may lead to overfitting and decreasing performance [39]. Comparing the nnU-Net default and ResEnc-L references with the results of the Primus-M encoder trained from scratch, it is clear that, on LiTS and KiTS, where Primus-M is close in performance to the CNN references, pre-training allows to exceed the default nnU-Net, while the gap to the ResEnc-L is not

closed. This suggests that transformer encoders are generally still lagging behind their convolutional neural network (CNN) counterparts, indicating the need for transformer architectures with stronger segmentation capabilities.

5.6 Qualitative analysis

Aside from quantitative results, we provide a PCA visualisation of the 3D embeddings of Merlin, CT-FM, CT-CLIP, and our COLIPRI-CRM encoder, on a lung cancer case from the MSD Lung dataset (Fig. 1). The embedding resolution is very low for Merlin and CT-FM, providing hardly any semantic localisation. CT-CLIP yields embeddings of higher resolution, allowing features to be visually mapped from the input CT to the PCA map. However, the PCA map is inconsistent and noisy, and exhibits a strong bias towards absolute position within the scan, as visible through the anteroposterior green/red shift. On the other hand, our COLIPRI encoders yield higher-resolution embeddings, which are sharper and more consistent, allowing for clear recognition of the boundaries of the patient, lungs, and the abdominal organs, as well as the lung mass present in the right lung (on the left-hand side of the coronal and axial slice views).

5.7 Summary of the results

Our encoders learn clinically relevant features, as measured by the linear separability of features to abnormality classes (Sec. 5.1) and by our higher clinical accuracy in report generation, yielding SOTA results in both aspects.

They also show great text-to-image alignment, with substantial improvements in report-to-image retrieval, and are sensitive to fine-grained similarity measurements, as evidenced by our SOTA zero-shot classification results(Sec. 5.2). While the former is an effect of our additional MAE and RRG supervision signals, the latter is enabled by our novel Opposite Sentence Loss (OSL), which teaches our text encoder to embed short statements in a semantically meaningful way. However, a substantial gap remains between classification probes and zero-shot classification, indicating potential for further improvement in alignment. Regarding dense tasks, our encoders including the MAE objective yield results that are better or on par with the current SOTA SSL pre-training method MAE (Sec. 5.5). Compared to SOTA CNNs, however, our transformer architecture still limits the performance of our encoders, yielding worse performance than ResEnc-L.

Across all our experiments, we find that our COLIPRI-CRM encoder provides embeddings that are potent in global tasks and is adaptable to dense downstream tasks, demonstrating the strength and versatility of our method.

6 Limitations

We discuss below various opportunities for improvement, from a technical and clinical point of view.

On the technical side, the inclusion of the radiology report generation (RRG) objective yields only slight improvements, indicating insufficiencies in the objective formulation. This positions the radiology report generation (RRG) objective more as an auxiliary objective improving retrieval. Should one not be interested in retrieval, the radiology report generation (RRG) objective can be considered optional, with the COLIPRI-CM configuration performing similarly to the full COLIPRI-CRM. On a more general note, there remains a large gap between the trained classification probe and the zero-shot classification performance, highlighting the need for better alignment.

On the clinical side, despite the improvements achieved by COLIPRI-CRM, the clinical performance metrics remain below the thresholds one would expect in clinical practice. We anticipate that training the model on a substantially larger and more diverse dataset would further enhance its performance. Additionally, a more comprehensive evaluation, stratified by clinical findings and patient characteristics, would be necessary to fully assess and demonstrate the model's potential clinical utility.

7 Conclusion

COLIPRI achieves SOTA performance across all standard tasks, including classification, retrieval, segmentation, and report generation. Extensive evaluation on established benchmarks demonstrates consistent improvements over prior work, including stronger generalisation across datasets. These results confirm that COLIPRI advances the empirical frontier of the field and establishes a new performance baseline for the commonly studied tasks.

Authors contributions

Conceptualisation: TW, IEH, YG, FPG. Data curation: TW, IEH, OM, MTW, FPG. Formal analysis: TW, YG, SBT, HS, MI, CL, AS, FPG. Funding acquisition: KHMH, JAV. Investigation: TW, FPG. Methodology: TW, IEH, YG, MTW, FPG. Software: TW, IEH, YG, SBT, HS, MI, CL, AS, FPG. Supervision: VS, JAV, FPG. Validation: TW, YG, SBT, HS, MI, CL, FPG. Visualization: TW, FPG. Writing (original draft): TW, YG, FPG. Writing (review & editing): all authors.

Acknowledgements

The German AI Service Center WestAI provided some of the computational resources used in this work to Tassilo Wald after he left Microsoft.

We thank Matthew Lungren for his insightful comments regarding report translation and structuring.

References

1. Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., Ilse, M., Pérez-García, F., Salvatelli, V., Sharma, H., et al.: Maira-2: Grounded radiology report generation. arXiv preprint arXiv:2406.04449 (2024)
2. Bannur, S., Hyland, S., Liu, Q., Pérez-García, F., Ilse, M., Castro, D.C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., Schwaighofer, A., Wetscherek, M., Lungren, M.P., Nori, A., Alvarez-Valle, J., Oktay, O.: Learning to exploit temporal structure for biomedical vision-language processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15016–15027 (June 2023)
3. Blankemeier, L., Kumar, A., Cohen, J.P., Liu, J., Liu, L., Van Veen, D., Gardezi, S.J.S., Yu, H., Paschali, M., Chen, Z., Delbrouck, J.B., Reis, E., Holland, R., Truyts, C., Bluethgen, C., Wu, Y., Lian, L., Jensen, M.E.K., Ostmeier, S., Varma, M., Valanarasu, J.M.J., Fang, Z., Huo, Z., Nabulsi, Z., Ardila, D., Weng, W.H., Junior, E.A., Ahuja, N., Fries, J., Shah, N.H., Zaharchuk, G., Willis, M., Yala, A., Johnston, A., Boutin, R.D., Wentland, A., Langlotz, C.P., Hom, J., Gatidis, S., Chaudhari, A.S.: Merlin: a computed tomography vision–language foundation model and dataset. Nature (Mar 2026)
4. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: European conference on computer vision. pp. 1–21. Springer (2022)
5. Bolya, D., Huang, P.Y., Sun, P., Cho, J.H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Rasheed, H., Wang, J., Monteiro, M., Xu, H., Dong, S., Ravi, N., Li, D., Dollár, P., Feichtenhofer, C.: Perception Encoder: The best visual embeddings are not at the output of the network (Apr 2025). https://doi.org/10.48550/arXiv.2504.13181, http://arxiv.org/abs/2504.13181, arXiv:2504.13181 [cs]
6. Cao, W., Zhang, J., Xia, Y., Mok, T.C.W., Li, Z., Ye, X., Lu, L., Zheng, J., Tang, Y., Zhang, L.: Bootstrapping Chest CT Image Understanding by Distilling Knowledge from X-ray Expert Models. pp. 11238–11247 (2024), https://openaccess.thecvf.com/content/CVPR2024/html/Cao_Bootstrapping_Chest_CT_Image_Understanding_by_Distilling_Knowledge_from_X-ray_CVPR_2024_paper.html
7. Codella, N.C.F., Jin, Y., Jain, S., Gu, Y., Lee, H.H., Abacha, A.B., Santamaria-Pang, A., Guyman, W., Sangani, N., Zhang, S., Poon, H., Hyland, S., Bannur, S., Alvarez-Valle, J., Li, X., Garrett, J., McMillan, A., Rajguru, G., Maddi, M., Vijayrania, N., Bhimai, R., Mecklenburg, N., Jain, R., Holstein, D., Gaur, N., Aski, V., Hwang, J.N., Lin, T., Tarapov, I., Lungren, M., Wei, M.: MedImageInsight: An Open-Source Embedding Model for General Domain Medical Imaging (Oct 2024). https://doi.org/10.48550/arXiv.2410.06542, http://arxiv.org/abs/2410.06542, arXiv:2410.06542 [eess]
8. Dancette, C., Khlaut, J., Saporta, A., Philippe, H., Ferreres, E., Callard, B., Danielou, T., Alberge, L., Machado, L., Tordjman, D., Dupuis, J., Floch, K.L., Terrail, J.D., Moshiri, M., Dercle, L., Boeken, T., Gregory, J., Ronot, M., Legou, F., Roux, P., Sapoval, M., Manceron, P., Hérent, P.: Curia: A multi-modal foundation model for radiology (2025), https://arxiv.org/abs/2509.06830
9. Draelos, R.L., Dov, D., Mazurowski, M.A., Lo, J.Y., Henao, R., Rubin, G.D., Carin, L.: RAD-ChestCT Dataset (Oct 2020). https://doi.org/10.5281/zenodo.6406114, https://zenodo.org/records/6406114

10. Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva-02: A visual representation for neon genesis. Image and Vision Computing **149**, 105171 (2024). https://doi.org/https://doi.org/10.1016/j.imavis.2024.105171, https://www.sciencedirect.com/science/article/pii/S0262885624002762
11. Goyal, S., Kumar, A., Garg, S., Kolter, Z., Raghunathan, A.: Finetune like you pretrain: Improved finetuning of zero-shot vision models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19338–19347 (2023)
12. Hamamci, I.E., Er, S., Wang, C., Almas, F., Simsek, A.G., Esirgun, S.N., Dogan, I., Durugol, O.F., Hou, B., Shit, S., Dai, W., Xu, M., Reynaud, H., Dasdelen, M.F., Wittmann, B., Amiranashvili, T., Simsar, E., Simsar, M., Erdemir, E.B., Alanbay, A., Sekuboyina, A., Lafci, B., Kaplan, A., Lu, Z., Polacin, M., Kainz, B., Bluethgen, C., Batmanghelich, K., Ozdemir, M.K., Menze, B.: Generalist foundation models from a multimodal dataset for 3D computed tomography. Nat. Biomed. Eng. (Feb 2026)
13. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
14. Heller, N., Isensee, F., Trofimova, D., Tejpaul, R., Zhao, Z., Chen, H., Wang, L., Golts, A., Khapun, D., Shats, D., et al.: The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct. arXiv preprint arXiv:2307.01984 (2023)
15. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: 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. pp. 3942–3951 (2021)
16. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
17. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F.: nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 488–498. Springer (2024)
18. Li, R., Li, J., Jian, B., Yuan, K., Zhu, Y.: Reevalmed: Rethinking medical report evaluation by aligning metrics with real-world clinical judgment. arXiv preprint arXiv:2510.00280 (2025)
19. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
20. Lüddecke, T., Ecker, A.: Image segmentation using text and image prompts. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7086–7096 (2022)
21. Maninis, K.K., Chen, K., Ghosh, S., Karpur, A., Chen, K., Xia, Y., Cao, B., Salz, D., Han, G., Dlabal, J., Gnanapragasam, D., Seyedhosseini, M., Zhou, H., Araujo, A.: TIPS: Text-Image Pretraining with Spatial awareness (Mar 2025). https://doi.org/10.48550/arXiv.2410.16512, http://arxiv.org/abs/2410.16512, arXiv:2410.16512 [cs]
22. Munk, A., Ambsdorf, J., Llambias, S., Nielsen, M.: Amaes: Augmented masked autoencoder pretraining on public brain mri data for 3d-native segmentation. arXiv preprint arXiv:2408.00640 (2024)

23. Naeem, M.F., Xian, Y., Zhai, X., Hoyer, L., Van Gool, L., Tombari, F.: Silc: Improving vision language pretraining with self-distillation. In: European Conference on Computer Vision. pp. 38–55. Springer (2024)
24. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=a68SUt6zFt, featured Certification
25. Pai, S., Hadzic, I., Bontempi, D., Bressem, K., Kann, B.H., Fedorov, A., Mak, R.H., Aerts, H.J.W.L.: Vision Foundation Models for Computed Tomography (Jan 2025), https://arxiv.org/abs/2501.09001v1
26. Pai, S., Hadzic, I., Bontempi, D., Bressem, K., Kann, B.H., Fedorov, A., Mak, R.H., Aerts, H.J.: Vision foundation models for computed tomography. arXiv preprint arXiv:2501.09001 (2025)
27. Pérez-García, F., Sparks, R., Ourselin, S.: TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Computer Methods and Programs in Biomedicine p. 106236 (2021). https://doi.org/https://doi.org/10.1016/j.cmpb.2021.106236, https://www.sciencedirect.com/science/article/pii/S0169260721003102
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
29. Shui, Z., Zhang, J., Cao, W., Wang, S., Guo, R., Lu, L., Yang, L., Ye, X., Liang, T., Zhang, Q., Zhang, L.: Large-scale and fine-grained vision-language pre-training for enhanced CT image understanding. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=nYpPAT4L3D
30. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
31. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019)
32. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20730–20740 (June 2022)
33. Team, N.L.S.T.R.: The national lung screening trial: overview and study design. Radiology **258**(1), 243–253 (2011)
34. Team, Q.: Qwen2.5 technical report. arXiv preprint 2412.15115 (2024)
35. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., Hénaff, O., Harmsen, J., Steiner, A., Zhai, X.: SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features (Feb 2025). https://doi.org/10.48550/arXiv.2502.14786, http://arxiv.org/abs/2502.14786, arXiv:2502.14786 [cs]

36. Tschannen, M., Kumar, M., Steiner, A., Zhai, X., Houlsby, N., Beyer, L.: Image captioners are scalable vision learners too. Advances in Neural Information Processing Systems **36**, 46830–46855 (2023)
37. Wald, T., Roy, S., Isensee, F., Ulrich, C., Ziegler, S., Trofimova, D., Stock, R., Baumgartner, M., Köhler, G., Maier-Hein, K.: Primus: Enforcing attention usage for 3d medical image segmentation. arXiv preprint arXiv:2503.01835 (2025)
38. Wald, T., Ulrich, C., Lukyanenko, S., Goncharov, A., Paderno, A., Miller, M., Maerkisch, L., Jaeger, P., Maier-Hein, K.: Revisiting mae pre-training for 3d medical image segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5186–5196 (2025)
39. Wald, T., Ulrich, C., Suprijadi, J., Ziegler, S., Nohel, M., Peretzke, R., Kohler, G., Maier-Hein, K.: An openmind for 3d medical vision self-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23839–23879 (2025)
40. Wan, B., Tschannen, M., Xian, Y., Pavetic, F., Alabdulmohsin, I.M., Wang, X., Susano Pinto, A., Steiner, A., Beyer, L., Zhai, X.: Locca: Visual pretraining with location-aware captioners. Advances in Neural Information Processing Systems **37**, 116355–116387 (2024)
41. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing. vol. 2022, p. 3876 (2022)
42. Wu, L., Zhuang, J., Chen, H.: Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22873–22882 (June 2024)
43. Xu, T., Hosseini, S., Anderson, C., Rinaldi, A., Krishnan, R.G., Martel, A.L., Goubran, M.: A generalizable 3D framework and model for self-supervised learning in medical imaging. NPJ Digit. Med. **8**(1), 639 (Nov 2025)
44. Yan, A., McAuley, J., Lu, X., Du, J., Chang, E.Y., Gentili, A., Hsu, C.N.: Radbert: adapting transformer-based language models to radiology. Radiology: Artificial Intelligence **4**(4), e210258 (2022)
45. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid Loss for Language Image Pre-Training. pp. 11975–11986 (2023), https://openaccess.thecvf.com/content/ICCV2023/html/Zhai_Sigmoid_Loss_for_Language_Image_Pre-Training_ICCV_2023_paper.html
46. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11975–11986 (October 2023)
47. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine learning for healthcare conference. pp. 2–25. PMLR (2022)
48. Zhou, C., Loy, C.C., Dai, B.: Extract Free Dense Labels from CLIP. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 696–712. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-19815-1_40
49. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: Image BERT pre-training with online tokenizer. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=ydopy-e6Dg
50. Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J.: Models genesis. Medical image analysis **67**, 101840 (2021)