SimCroP: Radiograph Representation Learning with Similarity-driven Cross-granularity Pre-training

Rongsheng Wang[1,2,6]⋆, Fenghe Tang[1,2]⋆, Qingsong Yao[3], Rui Yan[1,2] ✉, Xu Zhang[1,2,6], Zhen Huang[2], Haoran Lai[1,2,6], Zhiyang He[6], Xiaodong Tao[6], Zihang Jiang[1,2] ✉, and S. Kevin Zhou[1,2,4,5] ✉

[1] School of Biomedical Engineering, Division of Life Sciences and Medicine, University of Science and Technology of China (USTC), Hefei Anhui, 230026, China
[2] Center for Medical Imaging, Robotics, Analytic Computing & Learning (MIRACLE), Suzhou Institute for Advance Research, USTC, 215123, China
[3] Stanford University, Palo Alto, California, 94025, United States
[4] Jiangsu Provincial Key Laboratory of Multimodal Digital Twin Technology, Suzhou
[5] Key Laboratory of Precision and Intelligent Chemistry, USTC
[6] Anhui IFLYTEK CO., Ltd.

Abstract. Medical vision-language pre-training shows great potential in learning representative features from massive paired radiographs and reports. However, in computed tomography (CT) scans, the distribution of lesions which contain intricate structures is characterized by spatial sparsity. Besides, the complex and implicit relationships between different pathological descriptions in each sentence of the report and their corresponding sub-regions in radiographs pose additional challenges. In this paper, we propose a **Sim**ilarity-Driven **Cro**ss-Granularity **P**re-training (SimCroP) framework on chest CTs, which combines similarity-driven alignment and cross-granularity fusion to improve radiograph interpretation. We first leverage multi-modal masked modeling to optimize the encoder for understanding precise low-level semantics from radiographs. Then, similarity-driven alignment is designed to pre-train the encoder to adaptively select and align the correct patches corresponding to each sentence in reports. The cross-granularity fusion module integrates multimodal information across instance level and word-patch level, which helps the model better capture key pathology structures in sparse radiographs, resulting in improved performance for multi-scale downstream tasks. SimCroP is pre-trained on a large-scale paired CT-reports dataset and validated on image classification and segmentation tasks across five public datasets. Experimental results demonstrate that SimCroP outperforms both cutting-edge medical self-supervised learning methods and medical vision-language pre-training methods. Codes and models are available at https://github.com/ToniChopp/SimCroP.

Keywords: Medical vision language pre-training · Similarity-driven alignment · Cross-granularity fusion.

⋆ Equal contribution

1 Introduction

Deep learning (DL) has demonstrated exceptional potential in the realm of radiograph representation learning [35,21,4,14,24,15,36,27,26,28], which is trained on large-scale annotated datasets and achieves performance on par with that of clinical expert. However, the annotation of radiographs remains a resource-intensive and burdensome endeavor for clinical practitioners outside their regular duties, posing a significant bottleneck in the advancement of DL applications in medical imaging. Medical vision-language pre-training (Med-VLP) [12,30] seeks to capitalize on the detailed textual interpretations provided by radiograph-report pairs to assist radiograph representation learning, which has emerged as a prominent focus in contemporary researcher [29,17,31,32,33,13].

However, radiographs usually have complex textures and structures, especially for 3D chest CT scans. This primary challenge lies in extracting precise representations from deeper feature spaces [9]. Prior Med-VLP studies on 3D chest CTs have predominantly relied on contrastive learning (CL) [22]. For instance, CT-CLIP [8] introduces a CT-centric contrastive language-image pre-training framework, which optimizes the mutual information between global representations. Similarly, M3D [1] develops a multi-modal large language model built on CL principles. BIUD [3] bootstraps the understanding of 3D chest CT images by distilling chest-related diagnostic knowledge from an extensively pre-trained 2D X-ray expert model. Moreover, fVLM [25] adopts a fine-grained approach, aligning anatomical regions of CT images with their comparable descriptions in radiology reports and performing CL on each anatomical region individually. MG-3D [19] incorporates both intra-patient cross-modal semantic consistency and inter-patient semantic correlations into cross-modal attention mechanisms.

Nevertheless, prevailing methods exhibit notable limitations in effectively utilizing prior knowledge from reports for radiograph representation learning in two critical aspects. First, the spatial distribution of lesions posing intricate structures is characterized by spatial sparsity [25], presenting considerable difficulties in extracting visual features. Second, radiology reports exhibit hierarchical linguistic structures, consisted of descriptive sentences describing correlative visual features and interpretive statements synthesizing the clinical narratives [8], introduces inherent complexity that hinders effective pre-training. Furthermore, the absence of explicit spatial grounding annotations for descriptive sentences introduces substantial optimization challenges in establishing precise correspondences between sentences and massive visual feature space during pre-training.

To address these issues, we propose **Sim**ilarity-driven **Cro**ss-granularity **P**re-training (**SimCroP**), which pre-trains strong representation by aligning descriptive sentences in radiology reports and their corresponding sub-regions in radiographs. SimCroP addresses three core medical self-supervised learning (Med-SSL) objectives: (1) *Masked image modeling*, which enhances the model's ability to capture structural and textural details in sparse radiographs.(2) *Sentence-subregion alignment.* Inspired by the fact that doctors usually write sentence to describe the pathological context of some subregions in the radiograph, we first choose the sentence level as the granularity for extracting the supervision in the report. Then, we design Similiarity-driven Alignment (SA) to pull the text feature of each sentence closer to the vision features of its most similar visual patches. Without any manual annotations, SA automatically optimizes the vision encoder to select and align the correct patches which are reflect to each sentence in reports. (3) *Cross-granularity masked report modeling*, which integrates instance-level visual features via global average pooling (GAP) with word-patch level cross-modal features, to facilitate the reconstruction of the masked reports.

To comprehensively evaluate the effectiveness of SimCroP, we pre-train our Med-VLP framework on the large-scale medical dataset CT-RATE [8], which comprises paired chest CT images and reports. We conduct extensive experiments on multi-granularity downstream tasks, including linear probing classification and fine-tuning segmentation, across five public datasets. Empirical results demonstrate the superiority and generalization ability of SimCroP, significantly surpassing state-of-the-art (SOTA) Med-VLP methods with substantial performance improvements.

2 Method

In this section, we delve into the design of SimCroP for medical vision language pre-training on chest CTs. Fig. 1 illustrates our similarity-driven cross-granularity pre-training (SimCroP) framework. The fire and ice icons represent the parameters of the module that are trained and frozen, with shared weights between the text encoders. First, we briefly introduce the multi-modal masked modeling for 3D radiographs and paired reports utilized in our framework in section 2.1. Then, we illustrate how SimCroP leverages similarity-driven alignment for better fit the sparsity of radiographs like CTs in section 2.2. Finally, we clarify the detailed approaches of cross-granularity fusion in section 2.3.

2.1 Masked Modeling

Our approach is grounded on the multi-modal masked autoencoder architecture [5]. Following previous works [1,3,25], we adopt vision transformer (ViT) [6] and BERT [18] as the vision and text encoders, respectively.

Radiograph masking. Given a radiograph $I \in \mathbb{R}^{H\times W\times D}$ and its paired report T, we first split the radiograph into $\frac{H}{P_H} \times \frac{W}{P_W} \times \frac{D}{P_D}$ patches with size $P_H \times P_W \times P_D$. Following MAE [10], we mask 75% of the patches, resulting in N unmasked patches $I_u = \{I_u^s\}_{s=1}^N$. These unmasked patches are then augmented with 3D position embeddings E_{pos} as described in [34] and passed through the vision encoder to obtain the vision feature $f_v = E_v(I_u, E_{pos})$. During the vision decoding stage, f_v is fed into the vision decoder D_v along with mask tokens f_m, to reconstruct the radiograph $\hat{I} = D_v(f_v, f_m)$. We optimize the reconstructed radiograph using the mean squared error (MSE) loss:

$$\mathcal{L}_{\text{MIM}}(\hat{I}, I) = \text{MSE}(D_v(E_v(I_u, E_{pos}), f_m), I). \tag{1}$$

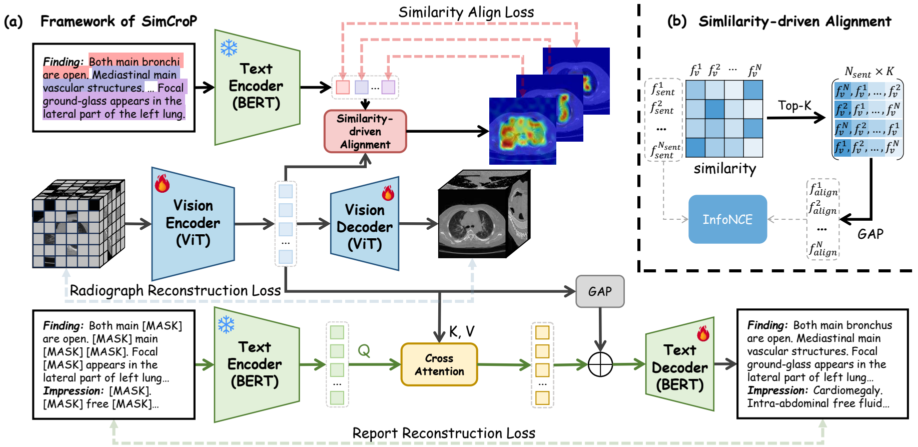
(a) Framework of SimCroP
Finding: Both main bronchi are open. Mediastinal main vascular structures. ... Focal ground-glass appears in the lateral part of the left lung.
Text Encoder (BERT)
Similarity Align Loss
Similarity-driven Alignment
Vision Encoder (ViT)
Vision Decoder (ViT)
Radiograph Reconstruction Loss
GAP
Finding: Both main [MASK] are open. [MASK] main [MASK] [MASK]. Focal [MASK] appears in the lateral part of left lung...
Impression: [MASK]. [MASK] free [MASK]...
Text Encoder (BERT)
Q
K, V
Cross Attention
Text Decoder (BERT)
Finding: Both main bronchus are open. Mediastinal main vascular structures. Focal ground-glass appears in the lateral part of left lung...
Impression: Cardiomegaly. Intra-abdominal free fluid...
Report Reconstruction Loss
(b) Simlilarity-driven Alignment
f_v^1 f_v^2 ... f_v^N
f_{sent}^1
f_{sent}^2
...
$f_{sent}^{N_{sent}}$
similarity
Top-K
$N_{sent} \times K$
$f_v^N, f_v^1, \ldots, f_v^2$
$f_v^2, f_v^1, \ldots, f_v^N$
$f_v^N, f_v^2, \ldots, f_v^1$
$f_v^1, f_v^2, \ldots, f_v^N$
GAP
f_{align}^1
f_{align}^2
...
f_{align}^N
InfoNCE

Fig. 1: (a) Overall framework of proposed SimCroP for medical vision-language pre-training. We combine multi-modal masked modeling, similarity-driven alignment, and cross-granularity fusion to achieve effective radiograph representation learning. (b) Details of similarity-driven alignment. For each descriptive sentence, we calculate the similarity with each patch in the radiograph and select the top-K most similar patches to ensure better alignment.

Report masking. For the original report consisting of N_t words, we first tokenize the words to tokens. Then, we randomly mask a proportion (γ) of tokens from the report T, resulting in a masked set of tokens $T_m = \{T_m^l\}_{l=1}^{N_m}$ and the unmasked tokens $T_u = \{T_u^l\}_{l=1}^{N_t - N_m}$. Both masked and unmasked tokens are passed together through the text encoder to obtain the masked report feature $f_t = E_t(T_m, T_u)$. The decoding procedure will be introduced in secion 2.3.

2.2 Simlilarity-driven Alignment

Given an input report $T = [T_\mathrm{F}, T_\mathrm{I}]$, T_F and T_I indicates the *finding* and *impression* section, respectively. We strategically disaggregate the "*Finding*" section T_F due to its clinical relevance in encapsulating comprehensive visual observations. For a "*Finding*" component comprising N_{sent} linguistically independent clauses, we directly feed the tokenized sentence ensemble $T_{sent} = \{T_{sent}^l\}_{l=1}^{N_{sent}}$ into the text encoder to derive sentence-level embeddings $f_{sent}^l = E_t(T_{sent}^l)$. As illustrated in Fig. 1(b), similarity-driven alignment between the l-th sentence T_{sent}^l and s-th unmasked patch I_u^s can be calculated as:

$$Sim_{l,s}(T_{sent}^l, I_u^s) = [E_t(T_{sent}^l)]^T [E_v(I_u^s, E_{pos})]. \tag{2}$$

The top-K most relevant spatial correspondences per sentence are then identified through:

$$Sim_{l,K}(T_{sent}^l, I_u^s) = \underset{0 \le s < N}{\mathrm{TopK}}\ Sim_{l,s}(T_{sent}^l, I_u^s). \tag{3}$$

Table 1: Evaluating our method against other SOTA Med-SSL and Med-VLP approaches on the linear probing classification task with **3D ViT-B backbone**. The best and second-best results are in **bold** and underlined, respectively.

Method	CT-RATE (AUC)			CC-CCII (ACC)			Rad-ChestCT (AUC)		LUNA16 (AUC)	
	1%	10%	100%	1%	10%	100%	10%	100%	10%	100%
Random init	58.2	60.9	63.8	57.8	74.7	77.2	56.4	60.0	57.6	60.4
3D Med-SSL										
MAE [10]	75.6	78.1	79.9	67.4	81.6	88.5	68.6	71.8	66.3	70.5
3D Med-VLP										
M3AE [5]	77.2	79.5	81.0	69.8	81.7	89.9	69.3	72.2	**67.1**	70.8
CT-CLIP [8]	74.1	78.6	80.4	65.0	76.7	83.6	67.5	69.1	64.0	65.0
MRM [33]	77.7	81.7	82.1	70.2	81.3	90.3	72.1	72.6	66.6	72.2
M3D [1]	74.3	79.1	80.7	65.4	77.0	83.8	68.0	69.8	65.2	68.4
fVLM [25]	79.4	81.8	82.2	72.3	82.9	90.7	71.1	74.2	65.9	71.3
SimCroP (Ours)	**81.0**	**82.4**	**82.9**	**73.1**	**83.2**	**91.3**	**73.4**	**75.8**	67.0	**73.3**

These discriminative visual features f_v^K undergo spatial aggregation via GAP to produce similarity-driven aligned feature for the l-th sentence $f_{align}^l = \mathrm{GAP}(f_v^K)$. To enforce semantic coherence between subregion visual features and sentence features, we employ similarity-driven fine-grained contrastive learning with the following symmetric InfoNCE loss [20]:

$$\mathcal{L}_{align} = -\frac{1}{N_{sent}} \sum_{i=1}^{N_{sent}} [\log \frac{\exp(s_{i,i}^{vt}/\tau)}{\sum_{j=1}^{N_{sent}} \exp(s_{i,j}^{vt}/\tau)} + \log \frac{\exp(s_{i,i}^{tv}/\tau)}{\sum_{j=1}^{N_{sent}} \exp(s_{i,j}^{tv}/\tau)}], \quad (4)$$

where $s_{i,j}^{vt} = (f_{align}^i)^T f_{sent}^j$, $s_{i,j}^{tv} = (f_{sent}^i)^T f_{align}^j$, τ denotes the temperature, which is set to 0.07 following common practice.

2.3 Cross-granularity Fusion

Diverging from M3AE [5], which singularly employs cross attention mechanisms to aggregat vision feature f_v to the report feature f_t, we propose a hierarchical fusion architecture comprising two complementary components:

- Instance-level vision features f^I are derived through global average pooling: $f^I = GAP(f_v)$;
- Word-patch level cross-modal features f^W are computed via scaled dot-product cross attention:

$$f^W = \mathrm{CrossAttention}(f_t, f_v) = \mathrm{Softmax}(\frac{f_t W_q (f_v W_k)^T}{\sqrt{d_k}}) f_v W_v, \quad (5)$$

where W_q, W_k, and W_v are linear projection matrix, d_k is the feature dimension. Finally, the fused features are passed through the text decoder to reconstruct

Table 2: Segmentation results of Med-SSL and Med-VLP approaches on the fine-tuning segmentation task with **3D ViT-B backbone**. The best and second-best results are in **bold** and <u>underlined</u>, respectively.

Method	LUNA16 (Dice)		BTCV (Dice)
	10%	100%	100%
Random init	91.9	92.3	78.9
3D Med-SSL			
MAE [10]	92.7	93.3	80.3
3D Med-VLP			
M3AE [5]	93.1	**93.7**	<u>80.5</u>
CT-CLIP [8]	92.9	93.4	79.4
MRM [33]	<u>93.2</u>	<u>93.5</u>	80.3
M3D [1]	92.9	93.2	79.6
fVLM [25]	93.1	93.4	80.0
SimCroP (Ours)	**93.5**	**93.7**	**80.7**

Table 3: Ablation study on each design component in our framework on linear-probe classification and fine-tuning segmentation. "SA" refers to Similarity-driven Alignment, "IL" denotes Instance-level vision features, and "WL" stands for Word-patch level cross-modal features. ✓ and × denote whether the component is included.

SA	IL	WL	RadChestCT (AUC) 10%	LUNA16 (Dice) 10%
×	✓	×	71.6	92.0
×	×	✓	69.8	91.7
×	✓	✓	72.4	92.4
✓	✓	×	73.3	92.8
✓	×	✓	72.7	92.6
✓	✓	✓	**73.4**	**93.5**

the masked tokens $\hat{T}_m = D_t(f^I + f^W)$. The masked text modeling loss for report reconstructing can be formulated as

$$\mathcal{L}_{\text{MLM}} = -\frac{1}{N_m}\sum_{l=1}^{N_m} \log P(\hat{T}_m^l = T_m^l \mid \hat{T}_m). \tag{6}$$

The overall loss function of SimCroP is as follows:

$$\mathcal{L} = \mathcal{L}_{\text{MIM}} + \lambda_1 \mathcal{L}_{align} + \lambda_2 \mathcal{L}_{\text{MLM}}. \tag{7}$$

Empirically, we configure $\lambda_1 = \lambda_2 = 1$ to maintain equilibrium between alignment and reconstruction objectives.

3 Experiments and Results

Pre-training datasets. We employ CT-RATE [8], a public large-scale dataset comprising 50,188 CT volumes with paired reports. Per official split, we leverage 47,149 volume-report pairs from the official training subset during pre-training. **Fine-tuning datasets.** CT-RATE [8] and RadChestCT [7] establish multi-label classification benchmarks with official data partition. The CC-CCII [11] dataset addresses multi-class pneumonia classification task with a 7:3 random partition. LUNA16 [23] provides dual objectives of nodule classification and pulmonary segmentation, with analogous 7:3 randomized data stratification. To

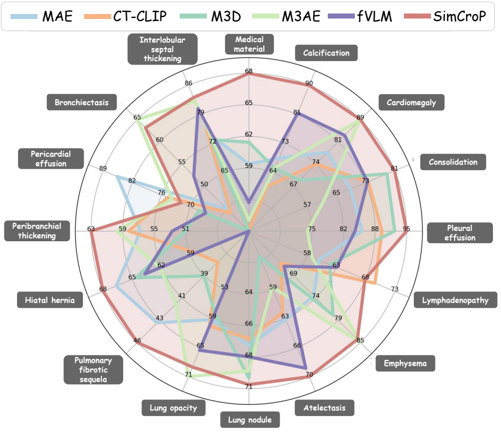

Fig. 2: Test results for each label on external dataset RadChestCT.

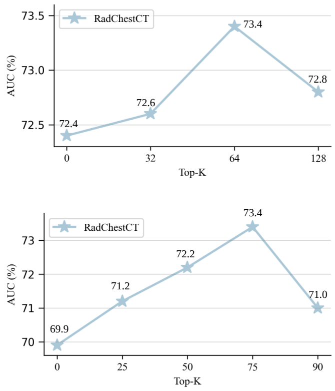

Fig. 3: Ablation on Top-K and report masking ratio selection.

assess cross-domain transferability of learned chest CT representations to abdominal imaging, we employ the BTCV [16] dataset under its prescribed organ segmentation protocol with official data splits. Unless official validation set is defined, testing sets serve as validation sets for all experimental configurations.

Implementation. We resample the volume to $1.5 \times 1.5 \times 3.0$ spacing using trilinear interpolation and map the Hounsfield unit range from $(-1000, 1000)$ to $(-1, 1)$, with clipping. The image size is set to be $224 \times 224 \times 112$. We leverage an 8-layer 3D ViT-B [6] as vision encoder initialized with MAE ImageNet-1K pre-trained weights [10] and a 4-layer 3D ViT-B for vision decoder, with patch size of $16 \times 16 \times 8$. Additionally, we employ pre-trained CXR-BERT [2] as our text encoder and a 6-layer BERT as our text decoder. The model is trained for 140 epochs on A800 GPUs with a batch size of 48, using AdamW as the optimizer with a learning rate of 1.5e-4 and weight decay of 0.05.

Baselines. We compare SimCroP with one Med-SSL method, MAE [10], and five Med-VLP methods: M3AE [5], CT-CLIP [8], MRM [33], M3D [1], and fVLM [25]. For fair comparison, we directly use the official pre-trained weights of M3D and re-implement the other methods under the same data settings as our approach.

Classification results. We perform linear-probe classification with our SimCroP on four datasets, as shown in Table 1, our SimCroP significantly outperforms 3D Med-SSL and 3D Med-VLP approaches across different training data ratios on all four datasets. Notably, on CT-RATE, which includes the largest number of disease labels and test data, SimCroP surpasses the current state-of-the-art (SOTA) method by 1.6% AUC while using 1% labeled data. Interestingly, we find the masked modeling-based methods M3AE [5], MRM [33] and SimCroP outperform contrastive-based methods like M3D [1] and fVLM [25] especially on fine-grained nodule classification task introduced by LUNA16. Additionally, Fig.2 shows the test results for each label on the external dataset Rad-ChestCT,

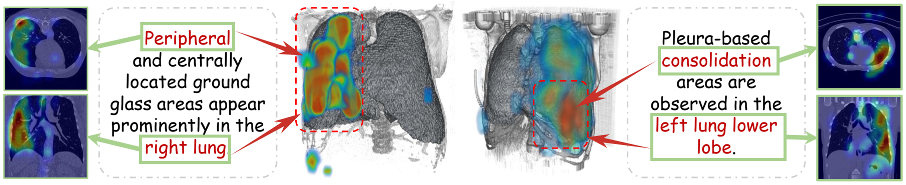

Fig. 4: Visualization of top-K patches correlated to descriptive sentences.

while training on 100% of the training data from CT-RATE. The label transfer setting follows CT-CLIP[8], and the transfer results further demonstrate the robustness and superiority of the representation learned by SimCroP.

Segmentation results. To evaluate the effectiveness of fine-grained radiograph representations learned by SimCroP, fine-tuning segmentation tasks are conducted in Table 2 using UNETR [9] framework. SimCroP consistently outperforms all aforementioned methods across lung segmentation and abdominal organ segmentation. This demonstrates that SimCroP effectively leverages its strong ability to utilize cross-granularity information between radiographs and reports. Therefore, SimCroP generalizes well across different medical vision tasks, establishing itself as a robust and highly effective Med-VLP method.

Ablation study. As demonstrated in Table 3, our similarity-driven alignment module achieves substantial performance gains in radiograph representation learning, with improvement differentials exceeding 1% and 1.1% on multi-label classification and pulmonary segmentation task respectively. A critical observation reveals that the ablation configuration devoid instance-level visual features underperforms the variant excluding word-patch cross-modal features for lung segmentation. This phenomenon can be attributed to the domain-specific characteristic of chest CTs where macro-anatomical structures occupy significantly larger voxel distributions compared to localized pathologies, thereby rendering holistic instance-level features more discriminative for lung segmentation. In Fig. 3, ablation studies investigating the selection of top-K and report masking ratios demonstrate that subregions comprising of 64 patches (appropriately 10% of the radiograph area) optimally align with the semantic granularity of descriptive sentences. This empirical evidence validates SimCroP's capability to address the inherent spatial sparsity of radiographs. Furthermore, a report masking ratio of 75% determines to maximize radiograph representation learning efficacy.

Representation visualization. Fig. 4 exhibits the subregions consisted of top-K similar patches correlated to the given descriptive sentences. Lesions and anatomy locations with higher similarity to the sentences are highlighted in red, showcasing SimCroP's superiority of similarity-driven alignment design.

4 Conclusion

We propose a novel medical vision-language pre-training method specifically designed for radiographs with inherent sparsity. To address the challenge of spatial sparsity and complex fine-grained connections between sentences and subregions, we pre-train the model to select and align each descriptive sentences with corresponding subregions. Furthermore, to tackle the issue of the massive visual feature space in 3D radiographs, we introduce a cross-granularity fusion module that simultaneously aggregates instance-level and word-patch level features. The effectiveness of our proposed modules is validated through multi-scale downstream tasks, including linear-probe classification and fine-tuning segmentation across multiple datasets, demonstrating the robustness of our method. However, the absence of instance-level cross-modal alignment hinders the zero-shot performance of our approach. In future work, we aim to address this limitation, further enhancing SimCroP as a more powerful foundation model.

Acknowledgments. This work was supported by National Natural Science Foundation of China (62271465, 62402473), Suzhou Basic Research Program under Grant SYG202338, and Jiangsu Province Science Foundation for Youths (NO. BK20240464).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bai, F., Du, Y., Huang, T., Meng, M.Q.H., Zhao, B.: M3d: Advancing 3d medical image analysis with multi-modal large language models. arXiv preprint arXiv:2404.00578 (2024)
2. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: ECCV. pp. 1–21 (2022)
3. Cao, W., Zhang, J., Xia, Y., Mok, T.C., et al.: Bootstrapping chest ct image understanding by distilling knowledge from x-ray expert models. In: CVPR. pp. 11238–11247 (2024)
4. Chen, X., Wang, X., Zhang, K., Fung, K.M., Thai, T.C., Moore, K., Mannel, R.S., Liu, H., Zheng, B., Qiu, Y.: Recent advances and clinical applications of deep learning in medical image analysis. Medical image analysis **79**, 102444 (2022)
5. Chen, Z., Du, Y., Hu, J., Liu, Y., Li, G., Wan, X., Chang, T.H.: Multi-modal masked autoencoders for medical vision-and-language pre-training. In: MICCAI. pp. 679–689. Springer (2022)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2020)
7. Draelos, R.L., Dov, D., Mazurowski, M.A., Lo, J.Y., Henao, R., Rubin, G.D., Carin, L.: Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. Medical image analysis **67**, 101857 (2021)
8. Hamamci, I.E., Er, S., Almas, F., Simsek, A.G., Esirgun, S.N., Dogan, I., Dasdelen, et al.: A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. CoRR (2024)
9. Hatamizadeh, A., Tang, Y., et al.: Unetr: Transformers for 3d medical image segmentation. In: WACV. pp. 574–584 (2022)

10. He, K., Chen, X., Xie, S., Li, Y., Doll'ar, P., Girshick, R.B.: Masked autoencoders are scalable vision learners. In: CVPR. pp. 15979–15988 (2021)
11. He, X., Wang, S., Shi, S., Chu, X., Tang, J., Liu, X., Yan, C., Zhang, J., Ding, G.: Benchmarking deep learning models and automated model design for covid-19 detection with chest ct scans. MedRxiv pp. 2020–06 (2020)
12. 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: ICCV. pp. 3942–3951 (2021)
13. Huang, W., Li, C., Zhou, H.Y., Yang, H., Liu, J., et al.: Enhancing representation in radiography-reports foundation model: A granular alignment algorithm using masked contrastive learning. Nature Communications **15**(1), 7620 (2024)
14. Huang, Z., Li, H., Shao, S., Zhu, H., Hu, H., Cheng, Z., Wang, J., Kevin Zhou, S.: Pele scores: pelvic x-ray landmark detection with pelvis extraction and enhancement. IJCARS **19**(5), 939–950 (2024)
15. Huang, Z., Zhou, X., He, X., Wei, Y., Yang, W., Wang, S., Sun, X., Li, H.: Casemark: A hybrid model for robust anatomical landmark detection in multi-structure x-rays. JKS University Computer and Information Sciences **37**(3), 1–18 (2025)
16. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop. vol. 5, p. 12 (2015)
17. Li, Z., Yang, L.T., Ren, B., et al.: Mlip: Enhancing medical visual representation with divergence encoder and knowledge-guided contrastive learning. In: CVPR. pp. 11704–11714 (2024)
18. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in NeurIPS **32** (2019)
19. Ni, X., Wu, L., Zhuang, J., Wang, Q., et al.: Mg-3d: Multi-grained knowledge-enhanced 3d medical vision-language pre-training. arXiv preprint arXiv:2412.05876 (2024)
20. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. ArXiv **abs/1807.03748** (2018)
21. Pathak, Y., Shukla, P.K., Tiwari, A., Stalin, S., Singh, S.: Deep transfer learning based classification model for covid-19 disease. Irbm **43**(2), 87–92 (2022)
22. 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: ICML. pp. 8748–8763. PMLR (2021)
23. Setio, A.A.A., Traverso, A., De Bel, T., et al.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. Medical image analysis **42**, 1–13 (2017)
24. Shao, S., Yuan, X., Huang, Z., et al.: Diffuseexpand: Expanding dataset for 2d medical image segmentation using diffusion models. arXiv preprint arXiv:2304.13416 (2023)
25. Shui, Z., Zhang, J., Cao, W., et al.: Large-scale and fine-grained vision-language pre-training for enhanced ct image understanding. In: ICLR (2025)
26. Tang, F., Nian, B., Li, Y., et al.: Mambamim: Pre-training mamba with state space token interpolation and its application to medical image segmentation. Medical Image Analysis p. 103606 (2025)
27. Tang, F., Xu, R., Yao, Q., et al.: Hyspark: Hybrid sparse masking for large scale medical image pre-training. In: MICCAI. pp. 330–340. Springer (2024)
28. Tang, F., Yao, Q., et al.: Hi-end-mae: Hierarchical encoder-driven masked autoencoders are stronger vision learners for medical image segmentation. arXiv preprint arXiv:2502.08347 (2025)

29. Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P.: Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. Nature Biomedical Engineering **6**(12), 1399–1406 (2022)
30. Wang, F., Zhou, Y., et al.: Multi-granularity cross-modal alignment for generalized medical visual representation learning. In: NeurIPS. vol. 35, pp. 33536–33549 (2022)
31. Wang, R., Yao, Q., Lai, H., He, Z., Tao, X., Jiang, Z., Zhou, S.: Ecamp: Entity-centered context-aware medical vision language pre-training. arXiv preprint arXiv:2312.13316 (2023)
32. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: MedCLIP: Contrastive learning from unpaired medical images and text. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 3876–3887 (Dec 2022)
33. Zhou, H.Y., Lian, C., Wang, L., Yu, Y.: Advancing radiograph representation learning with masked record modeling. In: ICLR (2023)
34. Zhou, L., Liu, H., Bae, J., He, J., Samaras, D., Prasanna, P.: Self pre-training with masked autoencoders for medical image classification and segmentation. In: 2023 ISBI. pp. 1–6. IEEE (2023)
35. Zhou, S.K., Greenspan, H., Davatzikos, C., Duncan, J.S., et al.: A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE (2021)
36. Zhou, X., Huang, Z., Zhu, H., Yao, Q., Zhou, S.K.: Hybrid attention network: An efficient approach for anatomy-free landmark detection. arXiv preprint arXiv:2412.06499 (2024)