CLMP: Contrastive Language-MRI Pretraining at Large Scale

Haoyu Jiang[1], Chu Zhang[1], Hongyuan Zhang[1], Lumin Chen[1], Zhiying Wu[1], Hongbin Liu[1], and Dong Yi[1⋆]

Centre for Artificial Intelligence and Robotics (CAIR)
Hong Kong Institute of Science & Innovation
Chinese Academy of Sciences

Abstract. Vision-Language Models (VLMs) have demonstrated remarkable performance in medical imaging domains through contrastive learning. However, their adaptation to Magnetic Resonance Imaging (MRI) remains underdeveloped primarily due to the lack of large-scale and high-quality MRI image–text datasets. To bridge this gap, we first construct a large-scale dataset comprising 2.4 million MRI image–text pairs. The dataset spans more than 18 anatomical regions and 27 disease categories, ensuring comprehensive coverage and representative diversity of the MRI domain. Building upon this dataset, we present CLMP, a model trained under the Contrastive Language-Image Pretraining (CLIP) framework enhanced with False Negative Cancellation (FNC). FNC mitigates supervision noise arising from identical textual descriptions paired with distinct images. We evaluate CLMP on linear probing and image–text retrieval tasks. Across a wide range of evaluation datasets, it achieves superior linear probing performance and strong image-text retrieval performance in median rank and Recall@1 compared with CLIP-style medical models and MR-specific baselines. Our contributions include: (1) a large-scale MRI image–text dataset, (2) a training paradigm combining CLIP and FNC, and (3) the first diagnostic-aware and natural-language-supervised CLIP-style model for MRI. Project page: `https://goodrain553.github.io/CLMP_/`.

Keywords: Vision-Language Pretraining · MRI Image-Text Dataset · Contrastive Learning

1 Introduction

Magnetic Resonance Imaging (MRI) provides high-resolution, non-invasive visualization of soft tissue structures. Such imaging is essential for accurate clinical diagnosis. Meanwhile, medical AI can automate complex image analysis for efficient and objective decisions [23, 6]. Therefore, a powerful medical model for MRI is essential to advance MRI-related diagnosis tasks.

⋆ Corresponding author.

Among medical models, Vision-Language Models (VLMs) have shown significant effectiveness in cross-modal representation learning. CLIP-style models [8, 13, 28] use language supervision to align text with medical images. Single-modality medical image models trained with language supervision have also proven useful [22, 27, 14]. Although MRI is one of the most important medical imaging techniques, CLIP-style VLMs for MRI images are currently not extensively explored and developed. In [1], the model is trained under the contrastive learning paradigm with supervision from text generated from DICOM metadata. However, DICOM-derived text lacks critical diagnostic and clinical condition information, resulting in inevitably degraded performance on text-related tasks. In addition, the scarcity of MRI image-text pairs and lack of high-quality textual descriptions still hinder the training of a CLIP-style VLM for MRI. Thus, to the best of our knowledge, no existing CLIP-style VLM for MRI images has been trained with natural language supervision that includes explicit diagnostic information.

To this end, we construct a dataset of 2.4 million MRI image–text pairs spanning over 18 anatomies and 27 disease types. In this process, many pairs with different images but identical texts emerge, especially when sampling 2D slices from 3D MRI volumes. This redundancy leads to false negative samples during contrastive learning. Inspired by [7], we combine CLIP [18] with False Negative Cancellation (FNC), which mitigates the noise by ignoring false negative samples. We evaluate the model on linear probing and image-text retrieval tasks. In linear probing, it outperforms state-of-the-art (SOTA) generic medical models and the strongest MRI-specific baselines across five evaluation datasets on average. In image-text retrieval, it further demonstrates superior performance in median rank and Recall@1. The main contributions of this paper are summarized as follows:

1. **MRI image-text dataset.** We construct a large-scale dataset with 2.4 million MRI image-text pairs to alleviate the data scarcity issue. Importantly, this dataset was built through three approaches: data cleaning, data preprocessing, and web-source data collection.
2. **Training method combining CLIP and FNC.** We combine contrastive learning and false negative cancellation to enable low-noise training with a large-scale MRI image-text dataset.
3. **A CLIP-style VLM for MRI modality.** We trained CLMP, a dedicated VLM for MRI images. This model aligns natural language with 2D MRI images in a shared latent space, thereby supporting various cross-modality tasks. The results show that it outperforms not only generic medical image models but also MRI-specific models.

2 The MRI Image-Text Pairs Dataset

The constructed dataset comprises 2.4 million MRI image–text pairs from 29 publicly available sources. In this section, the dataset is introduced in detail.

The dataset construction is described in Sec. 2.1, and the data statistics are presented in Sec. 2.2.

2.1 Dataset Construction

We construct the dataset using three approaches: data cleaning, data preprocessing, and web-source data collection. The latter two approaches are prone to generating samples with distinct images but identical texts, which conventional CLIP training treats as false negatives. In contrast, the CLMP framework effectively resolves this issue.

Data Cleaning. We extract MRI image-text pairs from generic datasets. Some datasets provide explicit modality labels, such as MedTrinity-25M [24], RadImageNet [15], and LLaVA-Med [11]. MedTrinity-25M [24] integrates multiple sources with explicit textual descriptions (e.g., SAMMed2D [26]). Although RadImageNet [15] provides only labels, UniMed-CLIP [8] generates textual descriptions for it. This allows us to directly filter the MRI data and utilize it without additional processing. For datasets without modality labels (e.g., PMC-OA [13], ROCOV2 [19]), we retain only image-text pairs whose text contains the term "MRI".

Preprocessing for Usability. For datasets that cannot be directly used for contrastive language–image learning, we preprocess them to ensure usability. Some datasets provide explicit textual descriptions but in 3D volumetric format. To address this, we slice the volumes into 2D images at specific intervals along different axes and pair them with the associated text. For instance, AutoRG-Brain [10] provides doctor-written impressions for 3D BraTS-MEN [21] and WHM [9]. In addition, some datasets contain high-quality images but only provide labels without textual descriptions (e.g., ACDC [2]). Following CLIP [18], we convert these labels into text using predefined templates.

Web-source Data Collection. We collect data from Radiopaedia, an online medical imaging education and resource platform where clinicians share radiology cases. Each case contains medical images categorized by imaging modality and accompanied by expert-reviewed captions. We extract 10 images from every imaging sequence within a case.

2.2 Dataset Statistics

Overall, the constructed dataset comprises approximately 2.4 million MRI image–text pairs, covering more than 18 anatomies and 27 disease types. Compared with previous MRI-specific studies [1, 5], our dataset is fully publicly available. As shown in Fig. 1(a), it spans major organ systems including the brain, breast, heart, and knee. When excluding mixed datasets without explicit anatomical

labels, brain image–text pairs constitute the largest portion of the dataset, accounting for 23.8%. Breast data follows with 12.5%. Foot and knee each contribute approximately 5%. Other anatomies, such as shoulder, spine, and hip, account for around 1% to 3%. The dataset covers over 27 distinct diseases, such as brain tumors, breast cancer, ischemic stroke, and myocardial infarction, while further disease types are implicitly represented in the unlabeled subset.

We analyze the average caption length across all datasets. Fig. 1(c) reports an average caption length of 86.43 words, with approximately 72.3% of captions ranging between 40 and 133 words (±1 standard deviation). Such long captions provide comprehensive semantic supervision, thereby enhancing both visual and textual understanding.

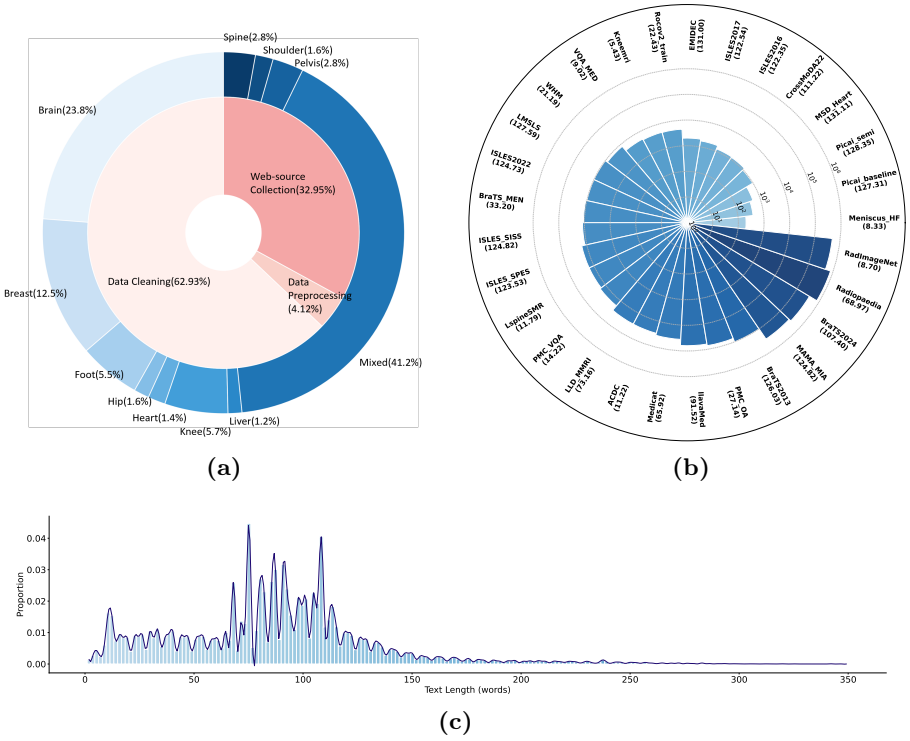

Fig. 1. Overview of data statistics. (a) Pie chart of anatomies and dataset source types, (b) Dataset magnitude comparison and average caption length at dataset scale, (c) Caption length distribution.

3 Method

We train CLMP on the constructed dataset. CLMP follows the paradigm of CLIP [18] and incorporates the FNC strategy. This method aligns image features and text embeddings in the latent space and addresses the issue of false negatives generated during dataset construction. The architecture is described in Sec. 3.1. The training objective is introduced in Sec. 3.2.

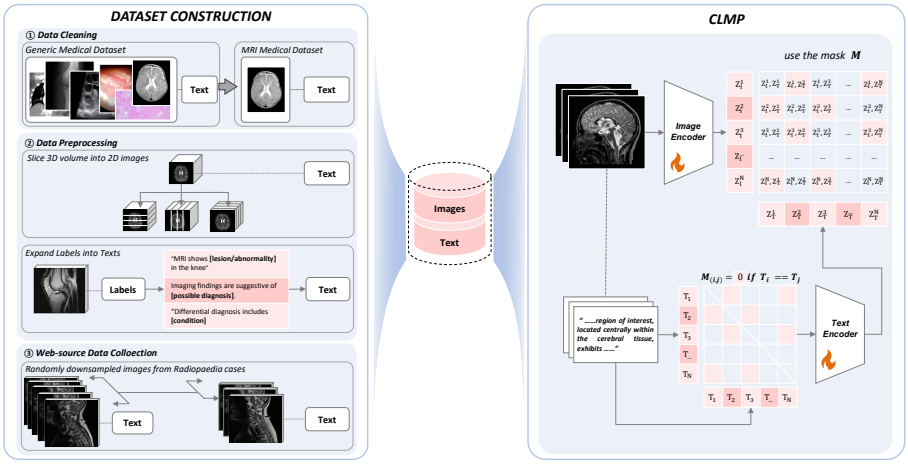

Fig. 2. Overview of dataset construction pipeline and CLMP architecture.

3.1 Architecture

Given a dataset $S = \{S_1, S_2, \ldots, S_n\}$, where $S_i = \{T_i, I_i\}$, each image I_i and text T_i is mapped into a shared latent space, producing representations z_I^i and z_T^i through the image encoder f_I and text encoder f_T, respectively.

$$z_I^i = f_I(I_i), \quad z_T^i = f_T(T_i) \tag{1}$$

3.2 Training Objective

To address the issue of false negatives in contrastive learning, we adopt FNC. The dataset construction pipeline often generates samples with different images but identical text, which may cause semantically related pairs to be mistakenly treated as negatives. FNC mitigates this by using a matrix M to record duplicated textual descriptions within each batch, where $M(i, j) = 0$ if $T_i = T_j$ and $i \neq j$, and $M(i, i) = 1$. For every anchor, non-matching samples with the same text are ignored, ensuring that false negatives do not distort the optimization.

Based on this, we implement the InfoNCE Loss [17] following CLIP [18], defined as the cross-entropy loss over similarity scores in both directions. The InfoNCE loss with FNC is formulated as:

$$\mathcal{L}_{I\to T}^{FNC} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\big(\mathrm{sim}(z_I^i,\, z_T^i)/\tau\big)}{\sum_{j=1}^{N} M(i,j)\exp\big(\mathrm{sim}(z_I^i,\, z_T^j)/\tau\big)} \tag{2}$$

$$\mathcal{L}_{T\to I}^{FNC} = -\frac{1}{N}\sum_{j=1}^{N}\log\frac{\exp\big(\mathrm{sim}(z_T^j,\, z_I^j)/\tau\big)}{\sum_{i=1}^{N} M(j,i)\exp\big(\mathrm{sim}(z_T^j,\, z_I^i)/\tau\big)} \tag{3}$$

$$L^{FNC} = \frac{1}{2}\left(\mathcal{L}_{I\to T}^{FNC} + \mathcal{L}_{T\to I}^{FNC}\right). \tag{4}$$

4 Experimental Settings

We evaluate the performance of our model using linear probing and image-text retrieval. Our model is compared with SOTA CLIP-style generic medical models, including UniMed-CLIP [8], BioMed-CLIP [28], as well as the MRI-specific models, MRI-CORE [5] and MR-CLIP [1]. Details of the downstream tasks are provided in Sec. 4.1. The implementation details are provided in Sec. 4.2.

4.1 Downstream Tasks

Linear Probing. We evaluate linear separability via linear probing on five medical tasks: brain tumor (BT), Anterior Cruciate Ligament (ACL), meniscus (MEN), stroke (STR), and multiple sclerosis (MS). Our study relies on publicly available datasets, namely Kaggle Brain Tumor Classification [16], MRNet [3], Atlas Stroke [12], and the MICCAI 2008 Multiple Sclerosis Lesion Segmentation Challenge [20]. The pretrained encoder is frozen while a linear classifier is fine-tuned. Datasets are split into 80% training, 10% validation, and 10% testing, except for the brain tumor classification task. The brain tumor classification task adopts the official dataset split. Few-shot probing uses 1% and 5% of training data to evaluate the impact of data availability. Accuracy (ACC) is reported as the metric.

Image-Text Retrieval. Image–text retrieval is evaluated in both directions, namely image-to-text (I2T) and text-to-image (T2I). Experiments are conducted on the ROCOv2 [19] test set, where 745 MRI image–text pairs are used for evaluation. Cosine similarity is applied to measure the alignment between image and text features. The performance is reported with Recall@1 (R@1), Recall@5 (R@5) and Median Rank (MR).

4.2 Implementation Details

Following previous works [8, 28, 27, 13], we adopt ViT-B-16-quickgelu [25] as the vision encoder and BioMed-BERT [4] as the text encoder. We train CLMP for 40 epochs on a node with 6 A100 GPUs. The learning rate is set to 4×10^{-5} with a warm-up of 5,000 iterations, and the effective batch size is configured as 2,688. The total training time for our model is approximately 10 hours.

5 Results

In most downstream tasks, CLMP surpasses CLIP-style generic models as well as MRI-specific models. In Sec. 5.1, the results of the downstream tasks are provided. In Sec. 5.2, we perform an ablation study to demonstrate the effectiveness of the FNC module.

5.1 Downstream Tasks

Linear Probing. As shown in Table 1, CLMP ranks first in 4 of 5 tasks. Compared with CLIP-style VLMs, it achieves an average gain of 5.5 percentage points over the best-performing UniMed-CLIP. Against MRI-specific models,

Table 1. Results of linear probing. ACC is adopted as the evaluation metric. Best results are in **bold** and the second best results are underlined.

Models	BT	ACL	MEN	STR	MS
UniMed-CLIP	0.8833	0.7843	0.7864	0.8100	0.7978
BioMed-CLIP	0.8207	0.7255	0.7157	0.8207	0.8925
MRI-CORE	0.7323	0.7184	0.6961	0.7323	0.9120
MR-CLIP	0.6743	0.7476	0.7353	0.8250	**0.9454**
CLMP	**0.8879**	**0.8461**	**0.8365**	**0.8316**	0.9356

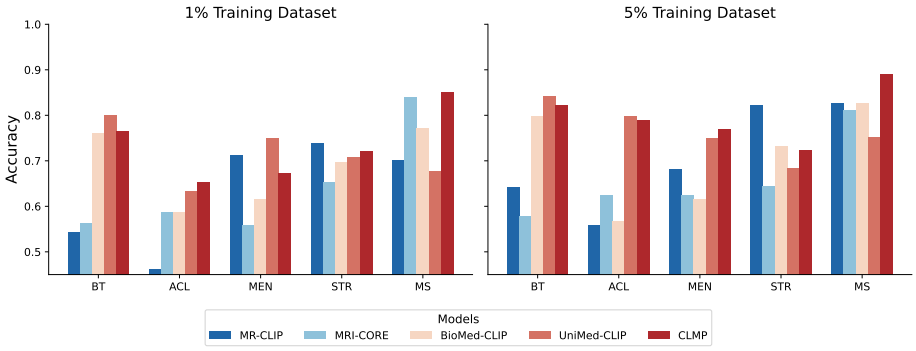

Fig. 3. Few-shot linear probing with 1% training data on the left and 5% on the right. CLMP achieves the best average rank under both settings.

Table 2. Results of image-text retrieval. Best results are in **bold** and the second best results are underlined.

Models	I2T			T2I		
	MR	R@1	R@5	MR	R@1	R@5
UniMed-CLIP	2.00	0.49	**0.81**	2.00	0.48	**0.80**
BioMed-CLIP	2.00	0.44	0.72	2.00	0.43	0.76
CLMP	**1.00**	**0.57**	0.76	**1.00**	**0.56**	0.75

Table 3. Ablative analysis of linear probing and image-text retrieval.

Models	BT	ACL	MEN	STR	MS
CLMP w/o FNC	0.8291	0.8173	0.8118	0.8032	0.7843
CLMP	0.8879	0.8461	0.8365	0.8316	0.9356

Models	I2T			T2I		
	MR	R@1	R@5	MR	R@1	R@5
CLMP w/o FNC	1.00	0.61	0.77	1.00	0.61	0.78
CLMP	1.00	0.57	0.76	1.00	0.56	0.75

CLMP surpasses the strongest baseline, MR-CLIP, with an 8.2 percentage-point margin.In few-shot linear probing, CLMP shows consistently strong results. As illustrated in Fig. 3, the histogram reports accuracies with 1% and 5% of training data. Under both conditions, CLMP attains an average rank of 1.8, representing the highest performance among all compared models.

Image-Text Retrieval. For I2T retrieval, our model achieves a median rank of 1 with Recall@1 of 0.57. For T2I retrieval, the median rank is also 1 and Recall@1 reaches 0.56, demonstrating consistent improvements in median rank and Recall@1 over CLIP-style medical models. Although CLMP has slightly lower Recall@5 than UniMed-CLIP, it provides stronger top-rank retrieval performance. In our evaluation, MR-CLIP [1] yields a median rank of 373 in T2I retrieval and 380 in I2T retrieval, likely because its language supervision is derived solely from DICOM metadata, which lacks radiologist-written diagnostic information.

5.2 Ablation Study

As shown in Table 3, CLMP consistently outperforms its variant without FNC in linear probing tasks by 5.8 percentage points on average. In the retrieval task, CLMP shows a 1–5 percentage-point drop in Recall@1 and Recall@5. This drop may be attributed to text-based false negative cancellation: false negatives are identified through duplicated texts rather than direct image similarity. Consequently, certain visual differences are not sufficiently compared, which reduces the granularity of semantic supervision and weakens the strength of contrastive learning for fine-grained retrieval.

6 Conclusion

In this paper, we construct a large-scale MRI image–text dataset from publicly accessible sources. Based on this dataset, we train our model, termed CLMP, within a paradigm that combines CLIP with FNC. CLMP consistently outperforms both CLIP-style generic models and MRI-specific models across most evaluation metrics in linear probing and image-text retrieval. We believe that the constructed dataset helps alleviate the scarcity of MRI image–text pairs for natural language-supervised learning. Furthermore, the proposed method can enhance the efficiency of data utilization.

References

1. Avci, M.Y., Borges, P., Fernandez, V., Wright, P., Yigitsoy, M., Ourselin, S., Cardoso, J.: Metadata-aligned 3d mri representations for contrast understanding and quality control (2025), https://arxiv.org/abs/2511.00681
2. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE transactions on medical imaging **37**(11), 2514–2525 (2018)
3. Bien, N., Rajpurkar, P., Ball, R.L., Irvin, J., Park, A., Jones, E., Bereket, M., Patel, B.N., Yeom, K.W., Shpanskaya, K., et al.: Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of mrnet. PLoS medicine **15**(11), e1002699 (2018)
4. Chakraborty, S., Bisong, E., Bhatt, S., Wagner, T., Elliott, R., Mosconi, F.: Biomedbert: A pre-trained biomedical language model for qa and ir. In: Proceedings of the 28th international conference on computational linguistics. pp. 669–679 (2020)
5. Dong, H., Chen, Y., Gu, H., Konz, N., Chen, Y., Li, Q., Mazurowski, M.A.: Mri-core: A foundation model for magnetic resonance imaging. arXiv preprint arXiv:2506.12186 (2025)
6. Eslami, S., Meinel, C., De Melo, G.: Pubmedclip: How much does clip benefit visual question answering in the medical domain? In: Findings of the Association for Computational Linguistics: EACL 2023. pp. 1181–1193 (2023)
7. Huynh, T., Kornblith, S., Walter, M.R., Maire, M., Khademi, M.: Boosting contrastive self-supervised learning with false negative cancellation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2785–2795 (2022)
8. Khattak, M.U., Kunhimon, S., Naseer, M., Khan, S., Khan, F.S.: Unimed-clip: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities (2024), https://arxiv.org/abs/2412.10372
9. Kuijf, H., Biesbroek, M., de Bresser, J., Heinen, R., Chen, C., van der Flier, W., Barkhof, Viergever, M., Biessels, G.J.: Data of the White Matter Hyperintensity (WMH) Segmentation Challenge (2022). https://doi.org/10.34894/AECRSD, https://doi.org/10.34894/AECRSD
10. Lei, J., Zhang, X., Wu, C., Dai, L., Zhang, Y., Zhang, Y., Wang, Y., Xie, W., Li, Y.: Autorg-brain: Grounded report generation for brain mri (2024), https://arxiv.org/abs/2407.16684

11. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
12. Liew, S.L., Anglin, J.M., Banks, N.W., Sondag, M., Ito, K.L., Kim, H., Chan, J., Ito, J., Jung, C., Khoshab, N., et al.: A large, open source dataset of stroke anatomical brain images and manual lesion segmentations. Scientific data **5**(1), 1–11 (2018)
13. Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., Xie, W.: Pmc-clip: Contrastive language-image pre-training using biomedical documents (2023), https://arxiv.org/abs/2303.07240
14. Maani, F., Saeed, N., Saleem, T., Farooq, Z., Alasmawi, H., Diehl, W., Mohammad, A., Waring, G., Valappi, S., Bricker, L., Yaqub, M.: Fetalclip: A visual-language foundation model for fetal ultrasound image analysis (2025), https://arxiv.org/abs/2502.14807
15. Mei, X., Liu, Z., Robson, P.M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K.E., Yang, T., et al.: Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. Radiology: Artificial Intelligence **4**(5), e210315 (2022)
16. Nickparvar, M.: Brain tumor mri dataset (2021). https://doi.org/10.34740/KAGGLE/DSV/2645886, https://www.kaggle.com/dsv/2645886
17. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021), https://arxiv.org/abs/2103.00020
19. Rückert, J., Bloch, L., Brüngel, R., Idrissi-Yaghir, A., Schäfer, H., Schmidt, C.S., Koitka, S., Pelka, O., Abacha, A.B., G. Seco de Herrera, A., et al.: Rocov2: Radiology objects in context version 2, an updated multimodal image dataset. Scientific Data **11**(1), 688 (2024)
20. Styner, M., Lee, J., Chin, B., Chin, M., Commowick, O., Tran, H., Markovic-Plese, S., Jewells, V., Warfield, S.: 3d segmentation in the clinic: A grand challenge ii: Ms lesion segmentation. MIDAS journal **2008**, 1–6 (2008)
21. de Verdier, M.C., Saluja, R., Gagnon, L., LaBella, D., Baid, U., Tahon, N.H., Foltyn-Dumitru, M., Zhang, J., Alafif, M., Baig, S., et al.: The 2024 brain tumor segmentation (brats) challenge: Glioma segmentation on post-treatment mri. arXiv preprint arXiv:2405.18368 (2024)
22. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text (2022), https://arxiv.org/abs/2210.10163
23. Wu, C., Zhang, X., Zhang, Y., Hui, H., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. Nature Communications **16**(1), 7866 (2025)
24. Xie, Y., Zhou, C., Gao, L., Wu, J., Li, X., Zhou, H.Y., Liu, S., Xing, L., Zou, J., Xie, C., Zhou, Y.: Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine (2025), https://arxiv.org/abs/2408.02900
25. Xu, H., Xie, S., Tan, X.E., Huang, P.Y., Howes, R., Sharma, V., Li, S.W., Ghosh, G., Zettlemoyer, L., Feichtenhofer, C.: Demystifying clip data. arXiv preprint arXiv:2309.16671 (2023)

26. Ye, J., Cheng, J., Chen, J., Deng, Z., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., Sun, H., Zhu, M., Zhang, S., He, J., Qiao, Y.: Sa-med2d-20m dataset: Segment anything in 2d medical imaging with 20 million masks (2023), https://arxiv.org/abs/2311.11969
27. You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E.K., Baek, W., Roh, B.: Cxr-clip: Toward large scale chest x-ray language-image pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 101–111 (2023)
28. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arxiv 2023. arXiv preprint arXiv:2303.00915 **70** (2023)