Benchmarking Vision-Language Contrastive Methods for Medical Representation Learning

Shuvendu Roy * [1] [2] **Yasaman Parhizkar** * [1] [3] **Franklin Ogidi** [1] **Vahid Reza Khazaie** [1] **Michael Colacci** [4]
Ali Etemad [2] **Elham Dolatabadi** [1] [3] **Arash Afkanpour** [1]

Abstract

We perform a comprehensive benchmarking of contrastive frameworks for learning multimodal representations in the medical domain. Through this study, we aim to answer the following research questions: (i) How transferable are general-domain representations to the medical domain? (ii) Is multimodal contrastive training sufficient, or does it benefit from unimodal training as well? (iii) What is the impact of feature granularity on the effectiveness of multimodal medical representation learning? To answer these questions, we investigate eight contrastive learning approaches under identical training setups, and train them on 2.8 million image-text pairs from four datasets, and evaluate them on 25 downstream tasks, including classification (zero-shot and linear probing), image-to-text and text-to-image retrieval, and visual question-answering. Our findings suggest a positive answer to the first question, a negative answer to the second question, and the benefit of learning fine-grained features. Finally, our code is available at: `https://github.com/afkanpour/benchmarking-contrastive/`.

1. Introduction

The medical domain has seen rapid advancements due to the recent progress in AI methodologies, particularly in leveraging diverse data types to support complex decision-making processes and intervention planning (Acosta et al., 2022). In this domain, integrating data from multiple sources, such as imaging, pathology, and clinical notes, is essential for providing comprehensive patient care. This diversity in data modalities introduces challenges in building robust AI models, especially as it is practically impossible to curate large labeled datasets for all tasks, modalities, and outcomes necessary for training supervised models. Recently, multimodal learning has become a pivotal paradigm in AI, where models learn representations of multiple modalities by exploiting the interplay between them (Driess et al., 2023; Girdhar et al., 2023; Singh et al., 2020). Multimodal representation learning has also demonstrated its effectiveness in the medical domain (Lin et al., 2023; Eslami et al., 2021), enabling models to leverage unlabelled data to learn generalizable representations of medical data (Huang et al., 2023).

Contrastive learning is a popular, effective method for learning multi-modal representation from paired datasets. However, it requires vast unlabelled data to learn generalizable representation, often unavailable in the medical domain. Besides data size, the learning approach is vital for building a robust foundation model (Kaplan et al., 2020; Manna et al., 2024). Although prior works explored contrastive learning in the medical domain (Ikezogwo et al., 2024; Lin et al., 2023; Eslami et al., 2021), a comprehensive, systematic investigation of multimodal representation learning, crucial for understanding contrastive learning's efficacy, remains unexplored. In this work, we raise and explore three research questions (RQs).

RQ1: How effective and transferable are general-domain representations to the medical domain? By understanding the extent of knowledge transfer from general vision tasks to medical imaging tasks, we can capitalize on existing resources, potentially accelerating the development of specialized models. This understanding is critical for medical representation learning for two reasons: (a) it diminishes the necessity for large-scale medical datasets comparable to those in the general domain, and (b) it reduces the computational burden associated with training medical foundation models from scratch.

RQ2. To train effective multimodal image-text models in the medical domain, is multimodal contrastive training sufficient, or should unimodal training be added for better performance? Given that it has been shown in the general vision domain that the addition of unimodal

---

*Equal contribution [1]Vector Institute [2]Queen's University, Canada [3]York University, Canada [4]University of Toronto, Canada. Correspondence to: Arash Afkanpour <arash.afkanpour@vectorinstitute.ai>.

ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada.

learning alongside multimodal learning can improve learned unimodal representations (Wei & Hu, 2024; Zhang et al., 2023b), we question whether this approach would benefit contrastive training in the medical domain as well.

RQ3. What is the impact of feature granularity on the effectiveness of multimodal representation learning in the medical domain? Learning both global (high-level) and local (low-level) representations is necessary for medical tasks (Lu et al., 2024; Zhao et al., 2023). For instance, detecting microcalcifications in mammograms involves identifying small, clustered specks that can indicate the early stages of breast cancer (Alsheh Ali et al., 2019). Accordingly, we aim to identify the impact of different levels of feature granularity on contrastive representation learning in this context (Huang et al., 2021; Müller et al., 2022). Although some prior studies have investigated the impact of granularity on a small scale, these studies have not been conducted under a unified framework with an identical training setup.

To address these questions, we investigate eight main approaches to contrastive learning through extensive empirical analyses. First, we carefully curate an extensive set of datasets consisting of a total of 2.8 million image-text pairs. These datasets span three primary medical image modalities namely radiology (MRI, CT, ultrasound, and x-ray), histopathology, and endoscopy images. We then pretrain the contrastive methods on the paired curated samples, which is followed by evaluation on 25 downstream tasks, including classification (zero-shot and linear probing), image-to-text and text-to-image retrieval, and visual question-answering.

2. Methods

2.1. Preliminaries

To answer the questions raised above, we need to learn effective representations from paired image-text samples. Prior work has shown that contrastive learning is a suitable approach for learning representations from paired data (Radford et al., 2021; Girdhar et al., 2023). Let φ denote an image encoder and ψ denote a text encoder that maps images and text to a common representation space, respectively. Given a batch of training samples $B = \{(x_i, t_i)\}_{i=1}^{N}$, where x_i and t_i denote the i^{th} image and text instances respectively, the InfoNCE loss (Oord et al., 2018) is optimized by minimizing the distance between the representations of an image and its corresponding text, $(\varphi(x_i), \psi(t_i))$, while maximizing the distance between unrelated image-text representation pairs, $(\varphi(x_i), \psi(t_j)),\ i \neq j$:

$$\ell_{\text{con}}(x_i, t_i; B) = -\left(\log \frac{\exp(\langle \varphi(x_i), \psi(t_i)\rangle/\tau)}{\sum_{k=1}^{N} \exp(\langle \varphi(x_i), \psi(\boldsymbol{t_k})\rangle/\tau)}\right) + \log \frac{\exp(\langle \varphi(x_i), \psi(t_i)\rangle/\tau)}{\sum_{k=1}^{N} \exp(\langle \varphi(\boldsymbol{x_k}), \psi(t_i)\rangle/\tau)} \qquad (1)$$

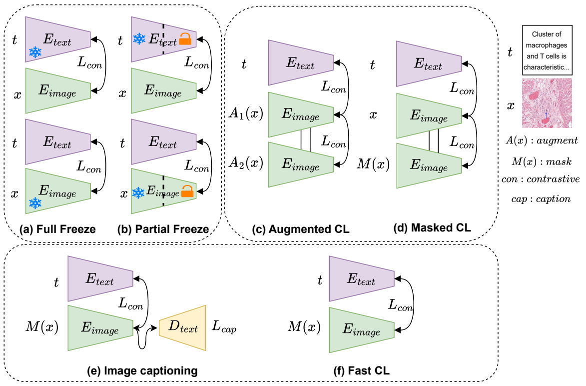


Figure 1. Illustration of eight contrastive learning approaches studied in the paper.

where $\langle\cdot,\cdot\rangle$ denotes similarity between two vectors (e.g. cosine similarity), and $\tau > 0$ is a temperature parameter. For simplicity of notation, we drop B and denote the loss for (x, t) by $\ell_{\text{con}}(x, t)$. Multimodal contrastive learning trains encoders φ and ψ by minimizing Eq. 1 over the pairs in B:

$$\ell_{\text{multimodal}} = \min_{\varphi,\psi} \ \mathbb{E}_B\left[\frac{1}{N}\sum_{i=1}^{N} \ell_{\text{con}}(x_i, t_i)\right]. \qquad (2)$$

2.2. Explored solutions

2.2.1. TRANSFERABILITY OF REPRESENTATIONS

To study the transferability and effectiveness of general-domain representations (RQ1), we consider encoders trained on a large-scale dataset of general image-text pairs. By a full or partial freeze of each of these encoders as shown in Figure 1 (a and b), we will study the effectiveness of representations of the corresponding modality for medical tasks. In particular, we consider the following cases:

1. Freeze the first α portion of layers of the image encoder and the first β portion of layers of the text encoder, but adapt the remaining layers. Here, *adapt* refers to the unsupervised fine-tuning of a model (or part of it) via contrastive learning on medical image-text pairs.

2. Fully freeze the image encoder, but adapt the text encoder ($\alpha = 1, \beta = 0$).

3. Fully freeze the text encoder, but adapt the image encoder ($\alpha = 0, \beta = 1$).

Partial fine-tuning while freezing early layers on a network allows for building high-level features on top of the early layers' features. In addition to leveraging general-domain features, this approach offers two additional benefits: (i) it reduces the number of learnable parameters, which improves robustness when training on a small amount of data, and (ii) it reduces the computational cost.

The extreme cases of a full encoder freeze (cases 2 and 3) imply a constant representation space for the corresponding modality. However, adaptation of the other modality's encoder allows for aligning its representations to that of the frozen encoder. A similar approach was taken in ImageBind (Girdhar et al., 2023), where the text and image encoders were frozen, but encoders of four other modalities were tuned to align with the frozen encoders.

2.2.2. UNIMODAL LEARNING

Vision-language contrastive learning maximizes alignment between image-text pairs. While this focuses on inter-modal alignment, a vast body of work in self-supervised learning has been devoted to unimodal representation learning. Leveraging unimodal representation learning in a multimodal framework has the potential to enhance representations by considering the relationship between examples within each modality (Wang et al., 2023). To explore this, we consider two unimodal learning approaches and combine each of them with multimodal contrastive loss.

For this purpose, we employ SimCLR (Chen et al., 2020) which is one of the pioneering techniques for visual self-supervised learning. In this technique, two views of an image are created by applying a series of random augmentations (e.g. crop-and-resize, color jitter, Gaussian blur, etc.) to the source image. Then, an encoder is learned by minimizing the contrastive loss (i.e., minimizing the distance between two views of an image in the representation space, and maximizing the distance between two views of different images) as illustrated in Figure 1 (c). Our method slightly differs from SimCLR in the implementation of the augmentation module. Since the medical domain is sensitive to a higher degree of cropping and certain colour distortions, we follow the augmentation module of Quilt (Ikezogwo et al., 2024) to our method, which has been proven to be effective in medical representation learning. Given a batch of data, B, and random augmentations, $\mathcal{A}_1$ and $\mathcal{A}_2$, SimCLR learns encoder φ and projector ξ (usually a multi-layer perceptron) by minimizing:

$$\ell_{\text{unimodal}} = \min_{\varphi,\xi} \mathbb{E}_{\mathcal{A}_1,\mathcal{A}_2,B}\left[\ell_{\text{con}}\Big(\xi(\varphi(\mathcal{A}_1(B))), \xi(\varphi(\mathcal{A}_2(B)))\Big)\right].$$

2.2.3. FEATURE GRANULARITY

We study how changing the granularity of features affects multimodal representation learning. To learn fine-grained features, we investigate CoCa (Yu et al., 2022), which performs simultaneous image-text contrastive learning and image captioning as shown in Figure 1 (e). The addition of autoregressive caption generation via a text decoder conditioned on the image representation enforces learning more granular features (Yu et al., 2022). The loss is defined as $\lambda\ell_{\text{multimodal}} + (1-\lambda)\ell_{\text{cap}}$, where for an image-text pair (x, t), the captioning loss, ℓ_{cap}, is defined as,

$$\ell_{\text{cap}}(x, t) = -\sum_{\nu} \log \delta(t_\nu | t_{<\nu}, \varphi(x)).$$

Table 1. Vision encoder selection results. For each case, we measure validation loss on PMC-OA validation set, retrieval Recall@200 on four datasets, and VQA accuracy on two datasets.

Encoders	**Val. loss**	**ROCO**		**Quilt**		**MIMIC-CXR**		**DeepEyeNet**		**PathVQA**	**VQARAD**
		I→T	T→I	I→T	T→I	I→T	T→I	I→T	T→I	overall	overall
RN-50	1.31	0.68	0.72	0.32	0.35	0.25	0.40	0.09	0.10	46.44	62.53
ViT-B/16	**1.15**	**0.78**	**0.77**	**0.44**	**0.46**	0.27	**0.44**	0.11	**0.11**	**46.75**	61.64
ViT-B/32	1.94	0.72	0.71	0.42	0.43	**0.33**	0.43	**0.12**	0.10	46.33	**62.97**

Table 2. Retrieval (Recall@200) and VQA performance of various methods examined in RQ1 for transferability of learned representation from the general domain.

Models	**ROCO**		**Quilt**		**MIMIC-CXR**		**DeepEyeNet**		**PathVQA**	**VQARAD**
	I→T	T→I	I→T	T→I	I→T	T→I	I→T	T→I	overall	overall
Baseline	0.78	0.77	0.44	0.46	0.27	0.44	0.11	0.11	46.75	61.24
Text Full Freeze	0.64	0.65	0.28	0.28	0.24	0.25	**0.12**	**0.12**	46.69	59.87
Text Partial Freeze (3L)	0.73	0.73	0.36	0.37	0.27	0.36	0.10	0.10	**47.15**	62.08
Text Partial Freeze (6L)	0.73	0.72	0.40	0.41	0.32	0.43	0.09	0.11	46.31	**63.19**
Text Partial Freeze (9L)	0.78	0.78	0.43	0.43	0.28	0.38	0.09	0.11	45.80	**63.19**
Image Full Freeze	0.52	0.61	0.22	0.28	0.17	0.24	0.07	0.10	46.28	59.27
Image Partial Freeze (3L)	0.74	0.76	0.44	0.46	0.25	0.36	0.09	0.10	46.69	60.53
Image Partial Freeze (7L)	**0.82**	0.82	**0.48**	**0.51**	**0.43**	**0.55**	0.07	0.11	46.06	61.42
Image Partial Freeze (9L)	0.81	**0.83**	0.44	0.47	0.34	0.45	0.08	0.10	46.74	60.31

Here, λ is a trade-off parameter, δ is a text decoder, t_ν denotes the ν-th token of t, $t_{<\nu}$ denotes all tokens up to the ν-th token in t, and $\varphi(x)$ denote the embedding of x.

Going in the other direction, to study the effect of learning high-level (coarse-grained) features in multimodal contrastive learning, we follow FLIP (Li et al., 2023). This method masks a large portion of image patches and applies contrastive loss between the visible patches and text, Figure 1 (f). In addition to enforcing the vision encoder to learn higher-level features, masking 50-75% of patches also reduces the computation cost by 2-4×.

3. Experiments

Model selection. Due to the extensive nature of this study, it is challenging to perform model selection for each contrastive learning variant. Therefore, we first perform hyperparameter tuning and model selection with vanilla contrastive learning. Hyperparameter tuning was mainly performed as a grid search on batch size and learning rate. We evaluate three image encoders, ResNet-50 (He et al., 2016), ViT-B/16, and ViT-B/32 transformers (Dosovitskiy et al., 2020). For the text encoder, we used the default GPT/77 encoder used by CLIP (Radford et al., 2021). We pretrain each candidate on our pretraining data (2.8M pairs) and then evaluate its performance by measuring validation loss, retrieval Recall@200, and VQA accuracy. The validation loss is measured on the validation set of the PMC-OA dataset. The retrieval performance is measured for both image-to-text and text-to-image cases on ROCO, Quilt, MIMIC-CXR, and DeepEyeNet datasets. VQA accuracy is measured on PathVQA and VQARAD datasets. For VQA tasks, we use a context length of 12 instead of 77 for the text encoder, similar to (Do et al., 2021). Table 1 presents the results for these encoders. Findings from this experiment suggest that ViT-B/16 is the best-performing encoder compared to

Table 3. Retrieval (Recall@200) and VQA performance of various methods examined in RQ2 for integrating unimodal representation learning with multimodal learning.

Models	**ROCO**		**Quilt**		**MIMIC-CXR**		**DeepEyeNet**		**PathVQA**	**VQARAD**
	I→T	T→I	I→T	T→I	I→T	T→I	I→T	T→I	overall	overall
Baseline	**0.78**	**0.77**	**0.44**	**0.46**	**0.27**	**0.44**	0.11	**0.11**	46.75	61.64
Augmented CL	0.56	0.55	0.32	0.34	0.21	0.27	0.11	0.08	46.25	63.24
Masked CL	0.56	0.56	0.33	0.35	0.21	0.28	**0.13**	0.09	**46.86**	**64.52**

Table 4. Retrieval performance (Recall@200) of various methods examined in RQ3 for feature granularity.

models	**ROCO**		**Quilt**		**MIMIC-CXR**		**DeepEyeNet**		**PathVQA**	**VQARAD**
	I→T	T→I	I→T	T→I	I→T	T→I	I→T	T→I	overall	overall
Baseline	**0.78**	0.77	**0.44**	**0.46**	0.27	0.44	0.11	0.11	46.75	61.64
Fast CL (25%)	0.76	**0.78**	**0.44**	**0.46**	0.27	0.38	0.08	0.13	46.79	62.08
Fast CL (50%)	0.77	0.76	0.43	0.45	**0.34**	**0.50**	0.08	0.09	46.25	61.64
Fast CL (75%)	0.72	0.73	0.41	0.44	0.23	0.33	**0.12**	0.11	46.79	62.97
Fast CL (85%)	0.69	0.69	0.40	0.41	0.22	0.33	0.08	0.10	46.72	58.31
Image captioning	0.74	0.74	**0.44**	**0.46**	**0.34**	0.47	0.09	**0.12**	**47.46**	**62.53**

other encoders in our study. Specifically, it achieves the best validation loss and outperforms the other encoders on 6 out of 8 retrieval and 1 out of 2 VQA tasks. Therefore, we consider ViT-B/16 as our default vision encoder and consider the performance of this model as the baseline.

Representations learned in the general domain *can* be transferred to the medical domain. As summarized in Table 5, we investigate four contrastive learning approaches to examine the transferability of representations from the general domain to the medical domain. The retrieval and VQA results are presented in Table 2. As we find from this table, partially freezing the image encoder (7-layer unlocked) shows the best result. These findings suggest that transferability from the general domain is a viable route for representation learning in the medical domain. In particular, primitive features learned in the early layers of a vision transformer on a large, general domain dataset do not require much adaptation by further pretraining on medical data. Accordingly, a partial freeze of the image encoder trained on general domain data could lead to improvement in the medical domain. The same phenomenon, however, is not observed when partially freezing the text encoder. Only 1 in the VQA tasks benefits from partial freezing of the text encoders, with 9L showing the best results among the partial text encoder freezing experiments. We also present classification results with F1-score on linear probing in Figure 3. Here, we observe that both partial freezing of the image and text encoder show similar results. More specifically, partially freezing the image encoder outperforms the baseline on 6 datasets, and partially freezing the text encoder outperforms the baseline on 6 datasets. Finally, fully freezing either of the encoders hurts performance. This is not unexpected, as representations learned for general-domain data may not be fully transferrable to the medical domain. Yet, for full encoder freezing, 5 tasks (3 in classification and 2 in retrieval) benefit from this approach for text encoders, while 1 classification task showed improvement with full freezing of image encoders.

Unimodal representation learning *may not* enhance multimodal learning. The results of our experimentation on two contrastive learning approaches enabling joint unimodal and multimodal representation learning are presented in Table 3 (retrieval and VQA) and Figure 4 (classification). While neither of the variants in this study performs well across most tasks, masked CL does improve performance in the 2 VQA tasks, 2 classification tasks, and one retrieval task (I→T). A few explanations are possible. For instance, a different trade-off in the loss combination could result in a better performance. It is also possible that optimizing the unimodal and multimodal loss functions requires very different feature sets, making the resulting representation space ineffective for our downstream tasks. Further investigation is required to better understand these results.

Fine-grained representation learning *can enhance* multimodal medical representations. In this study, Fast CL and Image Captioning are the two contrastive learning methods aimed at encouraging learning of the coarse-grained and fine-grained representations, respectively. As shown in Table 4, fine-grained learning improves (or is on par with) performance in 4 retrieval and 2 VQA tasks. Additionally, Figure 5 demonstrates that learning fine-grained features enhanced performance in 4 classification tasks. On the other hand, Fast CL is on par or outperforms the baseline in 3 retrieval tasks, but no VQA task, and 5 classification tasks with 25% masking setting. These results suggest that learning fine-grained features could be useful in the medical domain, compared to the coarse-grained representation. This observation is supported by the fact that local details are important in medical diagnosis tasks (Lu et al., 2024).

Other observations. Based on a holistic view of the results obtained from the experiments, we observe that in the medical domain, for the task of retrieval, Image Partial Freezing yields the overall highest results based on Tables 2, 3 and 4. For both zero-shot learning and linear probing, the standard CLIP baseline shows the overall best results according to Figures 3 through 5. Finally, for VQA, according to Tables 3 and 4, Masked CL and Image Captioning can yield the best performances.

4. Conclusion

We performed an extensive study of eight contrastive learning approaches for multimodal representation learning in the medical domain. This includes evaluating these variants on 19 classification tasks, 4 retrieval tasks, and 2 VQA tasks. Our results demonstrate that partial freezing of early layers of vision transformers and fine-tuning the remaining layers via contrastive learning could improve performance in medical downstream tasks. Additionally, incorporating techniques in contrastive learning that lead to fine-grained visual features could potentially improve performance.

References

Acosta, J. N., Falcone, G. J., Rajpurkar, P., and Topol, E. J. Multimodal biomedical ai. *Nature Medicine*, 28(9):1773–1784, 2022.

Alsheh Ali, M., Czene, K., Hall, P., and Humphreys, K. Association of microcalcification clusters with short-term invasive breast cancer risk and breast cancer risk factors. *Scientific reports*, 9(1):14604, 2019.

Baliah, S., Maani, F. A., Sanjeev, S., and Khan, M. H. Exploring the transfer learning capabilities of clip in domain generalization for diabetic retinopathy. In *International Workshop on Machine Learning in Medical Imaging*, pp. 444–453, 2023.

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 *ECCV*, pp. 1–21, 2022.

Chaitanya, K., Erdil, E., Karani, N., and Konukoglu, E. Contrastive learning of global and local features for medical image segmentation with limited annotations. *NeurIPS*, 33:12546–12558, 2020.

Chen, S., Gu, J., Han, Z., Ma, Y., Torr, P., and Tresp, V. Benchmarking robustness of adaptation methods on pre-trained vision-language models. *NeurIPS*, 36, 2024.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *ICML*, pp. 1597–1607, 2020.

Cui, Y., Zhao, L., Liang, F., Li, Y., and Shao, J. Democratizing contrastive language-image pre-training: A clip benchmark of data, model, and supervision. *arXiv preprint arXiv:2203.05796*, 2022.

Dattakumar, R. and Jagadeesh, R. A review of literature on benchmarking. *Benchmarking: An International Journal*, 10(3):176–209, 2003.

Do, T., Nguyen, B. X., Tjiputra, E., Tran, M., Tran, Q. D., and Nguyen, A. Multiple meta-model quantifying for medical visual question answering. In *MICCAI*, pp. 64–74, 2021.

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.

Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al. Palm-e: An embodied multimodal language model. In *ICML*, pp. 8469–8488, 2023.

Eslami, S., de Melo, G., and Meinel, C. Does clip benefit visual question answering in the medical domain as much as it does in the general domain? *arXiv preprint arXiv:2112.13906*, 2021.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., and Misra, I. Imagebind: One embedding space to bind them all. In *CVPR*, pp. 15180–15190, 2023.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

He, X., Zhang, Y., Mou, L., Xing, E., and Xie, P. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Huang, S.-C., Shen, L., Lungren, M. P., and Yeung, S. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *ICCV*, pp. 3942–3951, 2021.

Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J., and Zou, J. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023.

Ikezogwo, W., Seyfioglu, S., Ghezloo, F., Geva, D., Sheikh Mohammed, F., Anand, P. K., Krishna, R., and Shapiro, L. Quilt-1m: One million image-text pairs for histopathology. *NeurIPS*, 36, 2024.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pp. 4904–4916. PMLR, 2021.

Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., and Horng, S. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kim, I., Kim, J., Choi, J., and Kim, H. J. Concept bottleneck with visual concept filtering for explainable medical image classification. In *MICCAI*, pp. 225–233, 2023.

Kim, W., Son, B., and Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, pp. 5583–5594. PMLR, 2021.

Lau, J. J., Gayen, S., Ben Abacha, A., and Demner-Fushman, D. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S. C. H. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS*, 34:9694–9705, 2021a.

Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., Wu, H., and Wang, H. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020.

Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., and Yan, J. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021b.

Li, Y., Fan, H., Hu, R., Feichtenhofer, C., and He, K. Scaling language-image pre-training via masking. In *CVPR*, pp. 23390–23400, 2023.

Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., and Xie, W. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *MICCAI*, pp. 525–536, 2023.

Liu, C., Cheng, S., Shi, M., Shah, A., Bai, W., and Arcucci, R. Imitate: Clinical prior guided hierarchical vision-language pre-training. *arXiv preprint arXiv:2310.07355*, 2023.

Liu, C., Wan, Z., Cheng, S., Zhang, M., and Arcucci, R. Etp: Learning transferable ecg representations via ecg-text pre-training. In *ICASSP*, pp. 8230–8234, 2024.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Lu, M. Y., Chen, B., Williamson, D. F., Chen, R. J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L. P., Gerber, G., et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024.

Manna, S., Bhattacharya, S., and Pal, U. Self-supervised visual representation learning for medical image analysis: A comprehensive survey. *TMLR*, 2024.

Müller, P., Kaissis, G., Zou, C., and Rueckert, D. Joint learning of localized representations from medical images and reports. In *ECCV*, pp. 685–701. Springer, 2022.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Pang, T., Li, P., and Zhao, L. A survey on automatic generation of medical imaging reports based on deep learning. *BioMedical Engineering OnLine*, 22(1):48, 2023.

Pelka, O., Koitka, S., Rückert, J., Nensa, F., and Friedrich, C. M. Radiology objects in context (roco): a multimodal image dataset. In *MICCAI*, pp. 180–189, 2018.

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.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *IJCV*, 115:211–252, 2015.

Seibold, C., Reiß, S., Sarfraz, M. S., Stiefelhagen, R., and Kleesiek, J. Breaking with fixed set pathology recognition through report-guided contrastive training. In *MICCAI*, pp. 690–700, 2022.

Singh, A., Goswami, V., Natarajan, V., Jiang, Y., Chen, X., Shah, M., Rohrbach, M., Batra, D., and Parikh, D. Mmf: A multimodal framework for vision and language research. `https://github.com/facebookresearch/mmf`, 2020.

Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., and Kiela, D. Flava: A foundational language and vision alignment model. In *CVPR*, pp. 15638–15650, 2022.

Tiu, E., Talius, E., Patel, P., Langlotz, C. P., Ng, A. Y., and 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.

Tu, W., Deng, W., and Gedeon, T. A closer look at the robustness of contrastive language-image pre-training (clip). *NeurIPS*, 36, 2024.

Wang, P., Wang, S., Lin, J., Bai, S., Zhou, X., Zhou, J., Wang, X., and Zhou, C. One-peace: Exploring one general representation model toward unlimited modalities. *arXiv preprint arXiv:2305.11172*, 2023.

Wei, Y. and Hu, D. Mmpareto: Boosting multimodal learning with innocent unimodal assistance. In *ICML*, 2024.

Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., and Xu, C. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., and Beyer, L. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, pp. 18123–18133, 2022.

Zhan, X., Wu, Y., Dong, X., Wei, Y., Lu, M., Zhang, Y., Xu, H., and Liang, X. Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. In *ICCV*, pp. 11782–11791, 2021.

Zhang, J., Huang, J., Jin, S., and Lu, S. Vision-language models for vision tasks: A survey. *IEEE PAMI*, 2024.

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 preprint arXiv:2303.00915*, 2023a.

Zhang, X., Yoon, J., Bansal, M., and Yao, H. Multimodal representation learning by alternating unimodal adaptation. *arXiv preprint arXiv:2311.10707*, 2023b.

Zhang, Y., Gao, J., Zhou, M., Wang, X., Qiao, Y., Zhang, S., and Wang, D. Text-guided foundation model adaptation for pathological image classification. In *MICCAI*, pp. 272–282, 2023c.

Zhao, Z., Liu, Y., Wu, H., Li, Y., Wang, S., Teng, L., Liu, D., Li, X., Cui, Z., Wang, Q., et al. Clip in medical imaging: A comprehensive survey. *arXiv preprint arXiv:2312.07353*, 2023.

A. Related Work

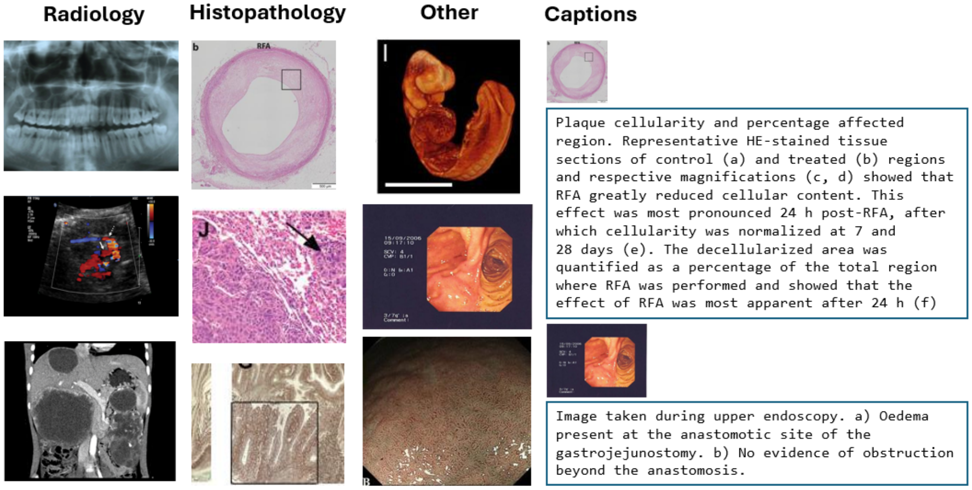
Vision language contrastive learning. Learning multimodal visual and text representations without human supervision using contrastive cross-entropy loss is a straightforward yet powerful pretraining paradigm that has gained significant interest in self-supervised tasks. The pioneering work of CLIP (Radford et al., 2021) builds a task-agnostic model by leveraging this method, which performs competitively with task-specific supervised models (Zhang et al., 2024). ALIGN (Jia et al., 2021) further scales up CLIP by leveraging a noisy dataset containing over one billion image-alt-text pairs. Given the data and compute-intensive nature of these Vision-Language Models (VLMs), several studies have explored more data and compute-efficient VLM pretraining methods using fewer image-text pairs. DeCLIP (Li et al., 2021b) introduces nearest-neighbor supervision to utilize information from similar pairs, enabling effective pretraining on limited data. LiT (Zhai et al., 2022) proposed contrastive tuning with a frozen image encoder, demonstrating another approach to enhance pretraining efficiency without extensive datasets. FLIP proposes a simple method of randomly masking out a large portion of image patches during training, which improves performance by encouraging the model to learn higher-level representations (Li et al., 2023). Another line of research has proposed combining dual-encoder contrastive pretraining with unimodal pretraining (Li et al., 2021a; Yu et al., 2022; Singh et al., 2022). This approach allows models to inherit the strengths of both methods, resulting in strong performance on vision-language benchmarks. For instance, ALBEF (Li et al., 2021a) integrates contrastive loss with masked language modelling using a dual-encoder design. CoCa (Yu et al., 2022) presents a simpler and more efficient approach by jointly training with contrastive and captioning losses. Another line of follow-up studies has focused on capturing finer-level information, such as the relationship between visual objects and textual words, by performing image-text contrastive learning across various semantic levels (Zhan et al., 2021; Li et al., 2020; Kim et al., 2021). For example, FILIP (Yao et al., 2021) introduces fine-grained semantic alignment (image patches and text tokens) through a novel cross-modal late interaction mechanism in contrastive learning, enabling the model to learn detailed vision-language correspondence.

Contrastive learning in the medical domain. The development and extension of contrastive vision language representations, particularly through models like CLIP (Zhao et al., 2023), has garnered significant interest in the medical domain. These models have demonstrated promising results and explored transferability of original CLIP in their application to chest X-ray (Boecking et al., 2022; Tiu et al., 2022; Seibold et al., 2022), ECG data (Liu et al., 2024), histological analysis of stomach tissue (Zhang et al., 2023c), lung CT scans (Liu et al., 2023), dermatology (Kim et al., 2023), and eye fundus images (Baliah et al., 2023). To satisfy the needs of medical CLIP-style (or foundational models) and resolve data scarcity, some studies have focused on generating novel image-text pairs across multiple domains (Lin et al., 2023; Eslami et al., 2021; Zhang et al., 2023a) and histopathology (Ikezogwo et al., 2024; Huang et al., 2023). Due to notable differences between general web images and medical domain images, where small abnormalities can significantly influence diagnostic results, incorporating fine-grained contrastive learning has been shown to improve performance (Chaitanya et al., 2020). Similarly, in diagnostic reports consisting of multiple sentences that describe image findings in specific regions, the inclusion of local vision features has been demonstrated to be important for accurate interpretation (Pang et al., 2023). Studies such as GLoRIA (Huang et al., 2021) and LoVT (Müller et al., 2022) have explored local-level cross-modal contrastive learning between both text-to-image and image-to-text local features. Additionally, CONCH (Lu et al., 2024) incorporated (generative) loss for fine-grained region-level features, thereby benefiting various tasks such as visual recognition, crossmodal alignment, image captioning, and multimodal understanding.

Benchmarking in vision-language contrastive learning. Benchmarking is recognized as an essential tool for continuous improvement of quality (Dattakumar & Jagadeesh, 2003). In the deep learning era, characterized by rapid advancements and the proliferation of novel methods, the role of benchmarking has become increasingly pronounced (Russakovsky et al., 2015). It has evolved to provide a fair and consistent basis for evaluating different models, facilitating comparison, and driving innovation (Chen et al., 2024). In the realm of vision-language contrastive learning, surprisingly, there have been only a few comprehensive benchmark studies. In a recent study, (Cui et al., 2022) benchmarked contrastive learning across the three dimensions of data, supervision, and model architecture. They discovered that the quality of the data and the application of appropriate supervision significantly enhance the performance of contrastive learning. (Tu et al., 2024) explored the safety objectives of 83 CLIP models in terms of resilience to variations in visual factors, calibrated uncertainty estimations, and the ability to detect anomalous inputs. The differences between these models were not in their architecture but in their training methods, usage scenarios (such as few-shot learning and fine-tuning), and the types of encoders employed. In the medical domain, the importance of benchmarking is even more pronounced due to the critical nature of healthcare applications. To the best of our knowledge, no study has yet explored various contrastive learning frameworks in this domain.

Table 5. Different contrastive learning (CL) methods.

RQ	Method name	Description
RQ1	**Image Partial Freeze**	Freeze the first 50% of the image encoder
	Text Partial Freeze	Freeze the first 50% of the text encoder
	Image Full Freeze	Fully freeze the image encoder
	Text Full Freeze	Fully freeze the text encoder
RQ2	**Augmented CL**	Combined unimodal (SimCLR (Chen et al., 2020)) and multimodal CL
	Masked CL	Combined unimodal masked and multimodal CL
RQ3	**Image Captioning**	Combined CL and captioning (Yu et al., 2022)
	Fast CL	CL between text and the masked image (Li et al., 2023)



Plaque cellularity and percentage affected region. Representative HE-stained tissue sections of control (a) and treated (b) regions and respective magnifications (c, d) showed that RFA greatly reduced cellular content. This effect was most pronounced 24 h post-RFA, after which cellularity was normalized at 7 and 28 days (e). The decellularized area was quantified as a percentage of the total region where RFA was performed and showed that the effect of RFA was most apparent after 24 h (f)

Image taken during upper endoscopy. a) Oedema present at the anastomotic site of the gastrojejunostomy. b) No evidence of obstruction beyond the anastomosis.

Figure 2. Samples from the datasets used in this study.

Table 6. Evaluation datasets used in this study.

Task	Setup	Dataset	Modality	Nb. Samples
Retrieval	I→T & T→I	Quilt-1M	Histopathology	13,559
		MIMIC-IV-CXR	Chest X-ray	3,269
		ROCO	Chest X-ray	8,176
		DeepEyeNet	Retina	3,140
Zero-shot classification & Linear probing	6 classes	PAD-UFES-20	Dermatology	460
	7 classes	SkinCancer	Dermatology	2,003
	2 classes	PatchCamelyon (PCam)	Histopathology	32,768
	8 classes	NCT-CRC-HE-100K	Histopathology	6,333
	3 classes	LC25000Lung	Histopathology	3,000
	2 classes	LC25000Colon	Histopathology	2,000
	4 classes	BACH	Histopathology	100
	4 classes	SICAPv2	Histopathology	2,122
	14 classes*	ChestMNIST+	Chest X-ray	112,120
	9 classes	PathMNIST+	Colon Pathology	107,180
	7 classes	DermaMNIST+	Dermatoscope	10,015
	4 classes	OctMNIST+	Retinal OCT	109,309
	2 classes	PneumoniaMNIST+	Chest X-Ray	5,856
	5 classes	RetinaMNIST+	Fundus Camera	1,600
	2 classes	BreastMNIST+	Breast Ultrasound	780
	8 classes	BloodMNIST+	Blood Cell Microscope	17,092
	8 classes	TissueMNIST+	Kidney Cortex Microscope	236,386
	11 classes	OrganAMNIST+	Abdominal CT	58,830
	11 classes	OrganCMNIST+	Abdominal CT	23,583
	11 classes	OrganSMNIST+	Abdominal CT	25,211
VQA	-	VQA-RAD	Radiology	3,515
		PathVQA	Histopathology	32,799

*Denotes multi-label classification.

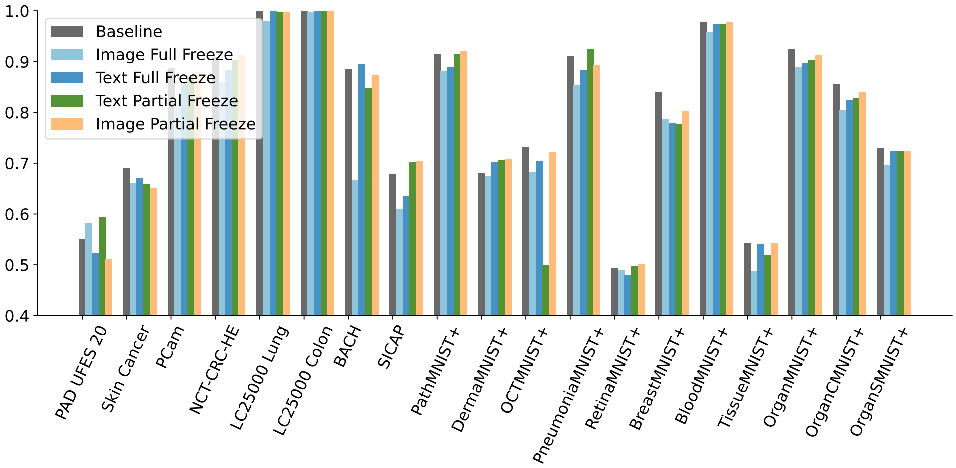


Figure 3. F1 score for linear probing on the study of transferability. Here, Image Partial Freeze (7L) outperforms the baseline on 6 datasets, while Text Partial Freeze (9L) also performs better than the baseline in 6 datasets. Text Full Freeze and Image Full Freeze perform better than the baseline in 3, and 1 datasets, respectively. Here, We only show the best-performing partial freezing experiments.

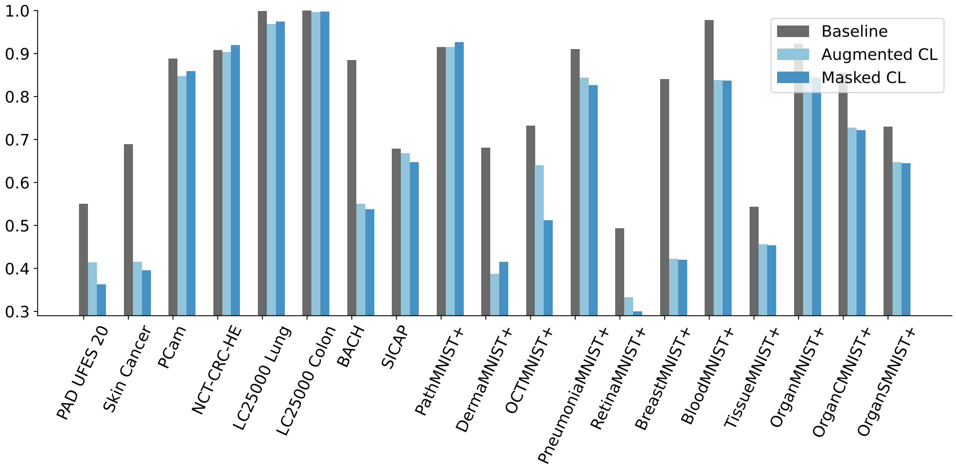


Figure 4. F1 score for linear probing on the study of unimodal learning. Here, Masked CL performs better than the baseline on two datasets.

B. Experimental Setups

Pretraining. We train *eight* contrastive vision-language models, categorized along three key dimensions: transferability (RQ1), unimodal learning (RQ2), and fine-grained representation learning (RQ3). Table 5 presents the benchmarked methods and their descriptions. Here, **Fast CL** follows the prescription of contrastive learning between a masked image and the corresponding text, proposed by FLIP (Yao et al., 2021). The training data consists of 2.8 million image-text pairs encompassing radiology (1.4M), histopathology (1.2M), and general medical images (e.g., endoscopy). Please see Figure 2 for a few samples. These pairs are sourced from four medical datasets including PMC-OA (Lin et al., 2023), Quilt-1M (Ikezogwo et al., 2024), MIMIC-CXR (Johnson et al., 2019), and ROCO (Pelka et al., 2018). These datasets contain a broad spectrum of textual content such as scholarly articles, clinical reports, and social media posts. We use the training splits of Quilt-1M, MIMIC-CXR, and ROCO for pretraining, while the test splits are later used for retrieval in downstream evaluation. The PMC-OA dataset is entirely used for pretraining. Please refer to Appendix Section C for more details on the pretraining data.

Evaluation. We perform an extensive evaluation of all models on a total of 25 retrieval, classification, and visual question-

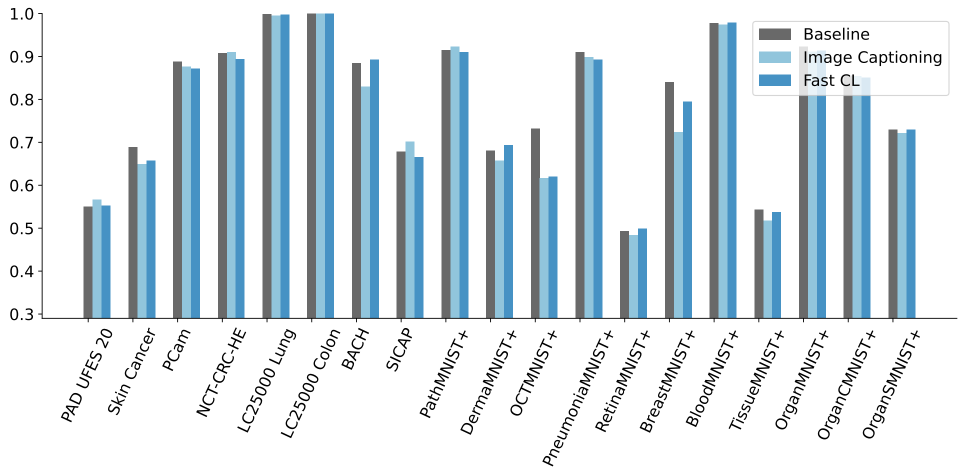


Figure 5. F1 score for linear probing on the study of feature granularity. Here, Image Captioning and Fast CL outperform the baseline on 4 datasets each.

Table 7. Training Datasets

Name	Image Sub-modality	Training Split Size
PMC-OA	Radiology and Histopathology	1.3M
Quilt-1M	Histopathology	1.0M
MIMIC-IV-CXR	Radiology	369.0K
ROCO	Radiology	65.6K
SUM		2.7M

answering (VQA) tasks. Within our evaluation, we also explore models' generalization performances on two modalities including dermatology and ophthalmology that are not present in pretraining data. Table 6 presents the datasets used for all evaluations. The image and text retrieval tasks encompass both image-to-text (I→T) and text-to-image (T→I) retrieval. To evaluate the (I→T) performance, we utilize test sets from the ROCO, Quilt, and MIMIC-CXR datasets. For each case, we report Recall@200 results. The image classification tasks include both zero-shot classification and linear probing across 22 tasks, including histopathology (9 tasks), radiology (8 tasks), dermatology (3 tasks), and ophthalmology (2 tasks). Visual question-answering is performed using the Mixture of Enhanced Visual Features (MEVF) method (Do et al., 2021) without using a decoder. We use our trained vision and text encoders to encode the image and question, respectively. This task contains open- and close-ended questions and is mapped as a classification problem and performed on PathVQA (He et al., 2020), VQARAD (Lau et al., 2018).

Inspired by (Wang et al., 2023), we also perform contrastive learning between representations of an original image and its masked version, Figure 1 (d). Following (Dosovitskiy et al., 2020), we divide the image into non-overlapping patches and then use a masking operator, $\mathcal{M}$, to randomly mask a certain fraction of patches. The unimodal contrastive loss in this case is defined as,

$$\ell_{\text{unimodal}} = \min_{\varphi,\xi} \; \mathbb{E}_{\mathcal{M},B}\left[\ell_{\text{con}}\Big(\xi(\varphi(B)), \xi(\varphi(\mathcal{M}(B)))\Big)\right].$$

Finally, in each case, the text and image encoders are trained by optimizing,

$$\lambda\ell_{\text{multimodal}} + (1-\lambda)\ell_{\text{unimodal}},$$

where, $\ell_{\text{multimodal}}$ is the image-text contrastive loss (Eq. 2), and $0 \leq \lambda \leq 1$ is a tradeoff parameter.

C. Datasets

Table 7 shows the datasets we used for training all models and each dataset's size and image sub-modality. The training data is balanced between radiology and histopathology sub-modalities.

MIMIC-CXR: MIMIC-CXR contains 377,110 images from 65,379 patients, with de-identified free-text reports describing the images. This dataset is the largest public chest X-ray dataset, acquired in the emergency department of Beth Israel Deaconess Medical Center in the US. For each patient, there are multiple views and a corresponding report labelled for 13 common radiological conditions using the CheXpert labeller (Irvin et al., 2019) or with "no finding" if no condition is present. Available labels include atelectasis, cardiomegaly, consolidation, edema, enlarged cardiomediastinum, fracture, lung lesion, lung opacity, pleural effusion, pleural other, pneumonia, pneumothorax, support devices, and no finding.

Quilt-1M: Quilt-1M contains more than one million histopathology image-text pairs. This dataset comprises four subsets. The main subset, Quilt, contains $802,144$ image-text pairs sourced from 1,087 hours of education histopathology videos on YouTube. Images and textual captions were automatically extracted from the videos using a mixture of models, including large language models, handcrafted algorithms, human knowledge databases, and automatic speech recognition. Three other subsets are sourced from PubMed Open Access Articles, The Large-scale Artificial Intelligence Open Network (LAION-5B), and Twitter data from OpenPath. Image-text pairs which are related to histopathology were extracted from these three datasets and combined with Quilt to constitute more than one million image-text pairs.

ROCO: Radiology Objects in COntext (ROCO) is a set of over 81K radiology image-text pairs with several medical imaging modalities including Computer Tomography, Ultrasound, X-Ray, Fluoroscopy, Positron Emission Tomography, Mammography, Magnetic Resonance Imaging, Angiography. ROCO is sourced from open-access biomedical articles on PubMedCentral.

PMC-OA: PMC-OA contains 1.65M image-text pairs encompassing histopathology, radiology, digital camera output, and other modalities. PMC-OA comprises many diagnostic procedures including X-ray, MRI, CT, Fluorescence, Ultrasound, fMRI, ENG, Radioisotope, Endoscope, Mitotic, DOT and PET. The pairs are automatically collected from PubMedCentral's open-access articles via a pipeline proposed in the original paper (Lin et al., 2023).

D. Hyperparameter and vision encoder configuration

All hyperparameters were established using vanilla contrastive learning, achieving the lowest validation loss across four retrieval tasks. We tuned the learning rate and batch size by training a selected model (ViT-B/32 and GPT/77) with various values of each hyperparameter and choosing the pair of hyperparameters which yielded the least validation loss. The chosen hyperparameters are $lr = 5e^{-5}$ and batch size of 32. Validation was done on the validation split of the PMC-OA dataset. After selecting the learning rate and batch size, we fixed the text encoder, GPT/77, and tested three different vision encoders to identify the optimal combination for all experiments. The least validation loss and retrieval scores for each encoder are shown in Table 1. As observed in Table 1, the combination of the vision encoder ViT-B/16 and the GPT/77 text encoder outperformed the other two vision encoders. This combination and the above hyperparameter will be used for all subsequent experiments.

E. Pretraining

For pretraining, we use a cosine decay learning rate scheduler with no warmup steps. We use the AdamW optimizer (Loshchilov & Hutter, 2017) with a weight decay of 0.1, $\beta_1 = 0.9$ and $\beta_2 = 0.999$, along with gradient accumulation with a frequency of 4. Gradient accumulation performs gradient updates after processing 4 batches. We use a cosine decay learning rate scheduler with no warmup steps. Our experiments are distributed across four A40 or A100 GPUs.

F. Downstream evaluation

Tables 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 show AUC and F1 score of linear probing and zero-shot classification for the three RQs. Table 20 shows VQA results for encoder selection. Tables 21, 22, 23 show VQA results for the three RQs.

Table 8. Linear Probing AUC across various medical datasets for models in RQ1.

Dataset	**Baseline**	**Image Full Freeze**	**Text Full Freeze**	**Text P. F. 6**	**Image P. F. 7**	**Dataset**	**Baseline**	**Image Full Freeze**	**Text Full Freeze**	**Text P. F. 6**	**Image P. F. 7**
sicap	0.91	0.88	0.88	0.91	0.98	dermamnist+	0.96	0.96	0.96	0.96	0.96
pad_ufes_20	0.88	0.90	0.88	0.88	0.89	octmnist+	0.97	0.96	0.96	0.96	0.96
skin_cancer	0.95	0.96	0.96	0.96	0.96	pneumoniamnist+	0.99	0.97	0.98	0.99	0.99
pcam	0.95	0.91	0.94	0.95	0.95	retinamnist+	0.86	0.83	0.85	0.86	0.87
nct_crc_he_100k	0.98	0.99	0.99	0.99	0.99	breastmnist+	0.88	0.88	0.89	0.89	0.88
lc25000_lung	0.99	1.00	1.00	1.00	1.00	bloodmnist+	0.99	0.79	0.91	0.91	0.91
lc25000_colon	1.00	1.00	1.00	1.00	1.00	tissuemnist+	0.91	0.89	0.91	0.91	0.91
bach	0.98	0.90	0.99	0.97	0.98	organamnist+	0.99	0.99	0.99	1.00	0.99
pathmnist+	0.99	0.99	0.99	0.99	0.99	organcmnist+	0.99	0.99	0.99	0.99	0.99
chestmnist+	0.78	0.75	0.77	0.79	0.80	organsmnist+	0.97	0.97	0.97	0.98	0.97

Dataset	**Baseline**	**Image P. F. 3**	**Text P. F. 3**	**Image P. F. 9**	**Text P. F. 9**	**Dataset**	**Baseline**	**Image P. F. 3**	**Text P. F. 3**	**Image P. F. 9**	**Text P. F. 9**
sicap	0.91	0.90	0.90	0.92	0.92	dermamnist+	0.96	0.96	0.96	0.96	0.96
pad_ufes_20	0.88	0.90	0.88	0.90	0.87	octmnist+	0.97	0.97	0.97	0.99	0.97
skin_cancer	0.95	0.96	0.96	0.96	0.96	pneumoniamnist+	0.99	0.99	0.99	0.99	0.99
pcam	0.95	0.93	0.96	0.95	0.95	retinamnist+	0.86	0.86	0.85	0.86	0.86
nct_crc_he_100k	0.98	0.99	0.99	1.00	0.99	breastmnist+	0.88	0.85	0.89	0.90	0.88
lc25000_lung	0.99	1.00	1.00	1.00	1.00	bloodmnist+	0.99	1.00	1.00	1.00	1.00
lc25000_colon	1.00	1.00	1.00	1.00	1.00	tissuemnist+	0.91	0.90	0.91	0.91	0.91
bach	0.98	0.96	0.99	0.98	0.98	organamnist+	0.99	0.99	1.00	0.99	1.00
pathmnist+	0.99	0.99	0.99	1.00	0.99	organcmnist+	0.99	0.99	0.99	0.99	0.99
chestmnist+	0.78	0.96	0.79	0.80	0.79	organsmnist+	0.97	0.97	0.98	0.98	0.99

Table 9. Linear Probing F1-score across various medical datasets for models in RQ1.

Dataset	Baseline	Image Full Freeze	Text Full Freeze	Text Partial Freeze	Image Partial Freeze	Dataset	Baseline	Image Full Freeze	Text Full Freeze	Text Partial Freeze	Image Partial Freeze
sicap	0.67	0.61	0.88	0.91	0.98	dermamnist+	0.68	0.68	0.70	0.70	0.71
pad_ufes_20	0.54	0.58	0.52	0.58	0.59	octmnist+	0.73	0.68	0.70	0.71	0.71
skin_cancer	0.68	0.66	0.67	0.67	0.66	pneumoniamnist+	0.91	0.85	0.88	0.90	0.93
pcam	0.88	0.82	0.85	0.88	0.87	retinamnist+	0.49	0.49	0.48	0.49	0.50
nct_crc_he_100k	0.90	0.86	0.88	0.91	0.90	breastmnist+	0.84	0.79	0.78	0.77	0.78
lc25000_lung	0.99	0.98	1.00	1.00	1.00	bloodmnist+	0.97	0.89	0.91	0.91	0.91
lc25000_colon	1.00	1.00	1.00	1.00	1.00	tissuemnist+	0.54	0.49	0.54	0.54	0.52
bach	0.88	0.67	0.99	0.97	0.98	organamnist+	0.92	0.89	0.99	1.00	0.99
pathmnist+	0.91	0.88	0.99	0.92	0.99	organcmnist+	0.85	0.81	0.99	0.99	0.99
chestmnist+	0.08	0.02	0.06	0.07	0.09	organsmnist+	0.73	0.70	0.99	0.98	0.97

Dataset	Baseline	Image P. F. 3	Text P. F. 3	Image P. F. 9	Text P. F. 9	Dataset	Baseline	Image P. F. 3	Text P. F. 3	Image P. F. 9	Text P. F. 9
sicap	0.67	0.63	0.66	0.70	0.70	dermamnist+	0.68	0.62	0.69	0.69	0.71
pad_ufes_20	0.54	0.58	0.54	0.55	0.51	octmnist+	0.73	0.73	0.63	0.73	0.72
skin_cancer	0.68	0.67	0.66	0.68	0.65	pneumoniamnist+	0.91	0.88	0.92	0.91	0.89
pcam	0.88	0.85	0.88	0.89	0.88	retinamnist+	0.49	0.48	0.47	0.54	0.50
nct_crc_he_100k	0.90	0.88	0.92	0.93	0.91	breastmnist+	0.84	0.77	0.83	0.81	0.80
lc25000_lung	0.99	0.99	1.00	1.00	1.00	bloodmnist+	0.97	0.96	0.98	0.98	0.98
lc25000_colon	1.00	1.00	1.00	1.00	1.00	tissuemnist+	0.54	0.50	0.54	0.54	0.54
bach	0.88	0.82	0.87	0.85	0.87	organamnist+	0.92	0.90	0.92	0.91	0.91
pathmnist+	0.91	0.90	0.93	0.94	0.92	organcmnist+	0.85	0.81	0.85	0.84	0.84
chestmnist+	0.08	0.05	0.06	0.08	0.07	organsmnist+	0.73	0.69	0.73	0.73	0.72

Table 10. Zero-shot classification AUC across various medical datasets for models in RQ1.

Dataset	Baseline	Image Full Freeze	Text Full Freeze	Text Partial Freeze	Image Partial Freeze	Dataset	Baseline	Image Full Freeze	Text Full Freeze	Text Partial Freeze	Image Partial Freeze
sicap	0.71	0.50	0.72	0.75	0.79	dermamnist+	0.60	0.50	0.64	0.63	0.64
pad_ufes_20	0.59	0.54	0.61	0.55	0.66	octmnist+	0.68	0.55	0.73	0.72	0.65
skin_cancer	0.57	0.48	0.66	0.72	0.63	pneumoniamnist+	0.94	0.52	0.93	0.84	0.85
pcam	0.71	0.62	0.84	0.57	0.78	retinamnist+	0.59	0.48	0.41	0.47	0.56
nct_crc_he_100k	0.95	0.78	0.86	0.91	0.92	breastmnist+	0.53	0.56	0.44	0.62	0.46
lc25000_lung	0.97	0.59	0.98	0.97	0.94	bloodmnist+	0.65	0.56	0.67	0.56	0.59
lc25000_colon	0.99	0.91	0.88	0.99	1.00	tissuemnist+	0.45	0.41	0.50	0.50	0.48
bach	0.82	0.62	0.66	0.73	0.79	organamnist+	0.85	0.73	0.82	0.84	0.85
pathmnist+	0.88	0.79	0.81	0.83	0.86	organcmnist+	0.78	0.66	0.77	0.77	0.78
chestmnist+	0.50	0.52	0.47	0.50	0.50	organsmnist+	0.81	0.63	0.78	0.78	0.80

Dataset	Baseline	Image P. F. 3	Text P. F. 3	Image P. F. 9	Text P. F. 9	Dataset	Baseline	Image P. F. 3	Text P. F. 3	Image P. F. 9	Text P. F. 9
sicap	0.71	0.59	0.67	0.80	0.68	dermamnist+	0.60	0.53	0.64	0.60	0.65
pad_ufes_20	0.59	0.59	0.65	0.60	0.62	octmnist+	0.68	0.69	0.56	0.72	0.69
skin_cancer	0.57	0.52	0.61	0.64	0.66	pneumoniamnist+	0.94	0.83	0.56	0.83	0.78
pcam	0.71	0.79	0.80	0.71	0.59	retinamnist+	0.59	0.58	0.51	0.45	0.47
nct_crc_he_100k	0.95	0.90	0.88	0.95	0.90	breastmnist+	0.53	0.36	0.53	0.55	0.58
lc25000_lung	0.97	0.94	0.98	0.98	0.96	bloodmnist+	0.65	0.61	0.56	0.62	0.65
lc25000_colon	0.99	0.98	0.97	1.00	0.95	tissuemnist+	0.45	0.43	0.50	0.50	0.35
bach	0.82	0.63	0.74	0.82	0.72	organamnist+	0.85	0.84	0.78	0.82	0.85
pathmnist+	0.88	0.84	0.88	0.85	0.91	organcmnist+	0.78	0.78	0.71	0.76	0.78
chestmnist+	0.50	0.50	0.50	0.50	0.50	organsmnist+	0.81	0.78	0.77	0.79	0.80

Table 11. Zero-shot classification F1-score across various medical datasets for models in RQ1.

Dataset	**Baseline**	**Image Full Freeze**	**Text Full Freeze**	**Text Partial Freeze**	**Image Partial Freeze**	**Dataset**	**Baseline**	**Image Full Freeze**	**Text Full Freeze**	**Text Partial Freeze**	**Image Partial Freeze**
sicap	0.40	0.12	0.30	0.35	0.32	dermamnist+	0.07	0.05	0.09	0.07	0.06
pad_ufes_20	0.12	0.08	0.21	0.07	0.15	octmnist+	0.23	0.10	0.26	0.25	0.24
skin_cancer	0.07	0.04	0.11	0.08	0.03	pneumoniamnist+	0.52	0.38	0.39	0.71	0.77
pcam	0.66	0.57	0.69	0.54	0.65	retinamnist+	0.05	0.14	0.09	0.08	0.12
nct_crc_he_100k	0.54	0.33	0.41	0.56	0.53	breastmnist+	0.52	0.55	0.22	0.47	0.47
lc25000_lung	0.73	0.26	0.90	0.85	0.77	bloodmnist+	0.03	0.56	0.10	0.47	0.59
lc25000_colon	0.95	0.82	0.41	0.93	0.94	tissuemnist+	0.03	0.04	0.09	0.04	0.03
bach	0.54	0.23	0.32	0.34	0.42	organamnist+	0.28	0.13	0.20	0.21	0.23
pathmnist+	0.44	0.37	0.27	0.38	0.43	organcmnist+	0.19	0.11	0.20	0.19	0.16
chestmnist+	0.09	0.10	0.10	0.10	0.10	organsmnist+	0.21	0.09	0.19	0.19	0.18

Dataset	**Baseline**	**Image P. F. 3**	**Text P. F. 3**	**Image P. F. 9**	**Text P. F. 9**	**Dataset**	**Baseline**	**Image P. F. 3**	**Text P. F. 3**	**Image P. F. 9**	**Text P. F. 9**
sicap	0.4	0.15	0.20	0.56	0.28	dermamnist+	0.07	0.04	0.07	0.09	0.07
pad_ufes_20	0.12	0.07	0.16	0.03	0.10	octmnist+	0.23	0.30	0.10	0.11	0.10
skin_cancer	0.07	0.07	0.06	0.10	0.18	pneumoniamnist+	0.05	0.55	0.53	0.60	0.67
pcam	0.66	0.67	0.58	0.66	0.33	retinamnist+	0.52	0.12	0.02	0.08	0.12
nct_crc_he_100k	0.54	0.42	0.55	0.63	0.59	breastmnist+	0.52	0.42	0.46	0.25	0.43
lc25000_lung	0.73	0.77	0.90	0.90	0.73	bloodmnist+	0.03	0.15	0.04	0.05	0.14
lc25000_colon	0.95	0.91	0.90	0.98	0.89	tissuemnist+	0.03	0.04	0.02	0.03	0.03
bach	0.54	0.33	0.37	0.52	0.36	organamnist+	0.28	0.24	0.22	0.18	0.23
pathmnist+	0.44	0.46	0.44	0.43	0.57	organcmnist+	0.19	0.21	0.15	0.17	0.21
chestmnist+	0.09	0.10	0.10	0.10	0.10	organsmnist+	0.21	0.20	0.20	0.20	0.20

Table 12. Linear probing AUC across various medical datasets for models in RQ2.

Dataset	**Augmented CL**	**Masked CL**	**Dataset**	**Augmented CL**	**Masked CL**
sicap	0.91	0.91	dermamnist+	0.90	0.90
pad_ufes_20	0.80	0.82	octmnist+	0.95	0.93
skin_cancer	0.90	0.90	pneumoniamnist+	0.97	0.96
pcam	0.93	0.94	retinamnist+	0.84	0.78
nct_crc_he_100k	0.99	0.99	breastmnist+	0.76	0.82
lc25000_lung	0.97	0.97	bloodmnist+	0.98	0.98
lc25000_colon	1.00	1.00	tissuemnist+	0.89	0.89
bach	0.97	0.98	organamnist+	0.99	0.99
pathmnist+	0.99	0.99	organcmnist+	0.99	0.99
chestmnist+	0.79	0.79	organsmnist+	0.98	0.98

Table 13. Linear probing F1-score across various medical datasets for models in RQ2.

Dataset	**Augmented CL**	**Masked CL**	**Dataset**	**Augmented CL**	**Masked CL**
sicap	0.67	0.60	dermamnist+	0.39	0.42
pad_ufes_20	0.41	0.36	octmnist+	0.64	0.51
skin_cancer	0.42	0.40	pneumoniamnist+	0.84	0.84
pcam	0.85	0.86	retinamnist+	0.33	0.30
nct_crc_he_100k	0.50	0.42	breastmnist+	0.42	0.42
lc25000_lung	0.97	0.97	bloodmnist+	0.84	0.84
lc25000_colon	1.00	1.00	tissuemnist+	0.45	0.45
bach	0.55	0.54	organamnist+	0.85	0.84
pathmnist+	0.91	0.92	organcmnist+	0.65	0.64
chestmnist+	0.01	0.01	organsmnist+	0.64	0.64

Table 14. Zero-shot classification AUC across various medical datasets for models in RQ2.

Dataset	**Augmented CL**	**Masked CL**	**Dataset**	**Augmented CL**	**Masked CL**
sicap	0.59	0.64	dermamnist+	0.44	0.49
pad_ufes_20	0.51	0.48	octmnist+	0.50	0.62
skin_cancer	0.47	0.53	pneumoniamnist+	0.20	0.65
pcam	0.82	0.76	retinamnist+	0.47	0.48
nct_crc_he_100k	0.89	0.87	breastmnist+	0.57	0.58
lc25000_lung	0.98	0.96	bloodmnist+	0.48	0.44
lc25000_colon	0.92	0.96	tissuemnist+	0.49	0.50
bach	0.60	0.65	organamnist+	0.79	0.67
pathmnist+	0.78	0.84	organcmnist+	0.72	0.69
chestmnist+	0.41	0.44	organsmnist+	0.75	0.72

Table 15. Zero-shot classification F1-score across various medical datasets for models in RQ2.

Dataset	**Augmented CL**	**Masked CL**	**Dataset**	**Augmented CL**	**Masked CL**
sicap	0.23	0.23	dermamnist+	0.12	0.03
pad_ufes_20	0.09	0.07	octmnist+	0.10	0.18
skin_cancer	0.13	0.12	pneumoniamnist+	0.20	0.18
pcam	0.74	0.69	retinamnist+	0.05	0.05
nct_crc_he_100k	0.44	0.37	breastmnist+	0.25	0.50
lc25000_lung	0.89	0.86	bloodmnist+	0.09	0.12
lc25000_colon	0.67	0.88	tissuemnist+	0.05	0.12
bach	0.39	0.29	organamnist+	0.13	0.12
pathmnist+	0.30	0.32	organcmnist+	0.12	0.12
chestmnist+	0.10	0.10	organsmnist+	0.12	0.13

Table 16. Linear probing AUC across various medical datasets for models in RQ3.

Dataset	**Image Captioning**	**Fast CL 25%**	**Fast CL 50%**	**Fast CL 75%**	**Fast CL 85%**	**Dataset**	**Image Captioning**	**Fast CL 25%**	**Fast CL 50%**	**Fast CL 75%**	**Fast CL 85%**
sicap	0.92	0.90	0.92	0.87	0.87	dermamnist+	0.95	0.96	0.96	0.97	0.96
pad_ufes_20	0.90	0.88	0.87	0.89	0.86	octmnist+	0.96	0.96	0.96	0.96	0.95
skin_cancer	0.95	0.95	0.96	0.96	0.96	pneumoniamnist+	0.98	0.99	0.98	0.99	0.98
pcam	0.96	0.94	0.91	0.91	0.92	retinamnist+	0.85	0.87	0.86	0.86	0.85
nct_crc_he_100k	0.99	0.99	0.99	0.99	0.98	breastmnist+	0.87	0.86	0.90	0.90	0.88
lc25000_lung	1.00	1.00	1.00	1.00	1.00	bloodmnist+	1.00	1.00	1.00	1.00	1.00
lc25000_colon	1.00	1.00	1.00	1.00	1.00	tissuemnist+	0.91	0.91	0.91	0.91	0.91
bach	0.97	0.98	0.96	0.97	0.97	organamnist+	0.99	1.00	0.99	1.00	0.99
pathmnist+	0.99	0.99	0.99	0.99	0.99	organcmnist+	0.99	0.99	0.99	0.99	0.99
chestmnist+	0.96	0.79	0.96	0.78	0.78	organsmnist+	0.98	0.98	0.98	0.97	0.98

Table 17. Linear probing F1-score across various medical datasets for models in RQ3.

Dataset	**Image Captioning**	**Fast CL 25%**	**Fast CL 50%**	**Fast CL 75%**	**Fast CL 85%**	**Dataset**	**Image Captioning**	**Fast CL 25%**	**Fast CL 50%**	**Fast CL 75%**	**Fast CL 85%**
sicap	0.65	0.67	0.62	0.59	0.57	dermamnist+	0.39	0.69	0.39	0.70	0.68
pad_ufes_20	0.57	0.55	0.53	0.51	0.53	octmnist+	0.42	0.62	0.39	0.70	0.48
skin_cancer	0.65	0.66	0.66	0.67	0.66	pneumoniamnist+	0.84	0.89	0.84	0.87	0.88
pcam	0.88	0.87	0.88	0.84	0.83	retinamnist+	0.50	0.50	0.50	0.50	0.50
nct_crc_he_100k	0.61	0.89	0.59	0.90	0.85	breastmnist+	0.79	0.79	0.77	0.77	0.77
lc25000_lung	1.00	1.00	0.53	1.00	1.00	bloodmnist+	0.49	0.98	0.53	0.98	0.98
lc25000_colon	0.95	1.00	0.53	1.00	1.00	tissuemnist+	0.54	0.98	0.54	0.53	0.52
bach	0.83	0.89	0.80	0.80	0.84	organamnist+	0.90	0.91	0.88	0.92	0.91
pathmnist+	0.92	0.91	0.92	0.91	0.87	organcmnist+	0.73	0.85	0.72	0.85	0.85
chestmnist+	0.06	0.07	0.07	0.06	0.06	organsmnist+	0.77	0.73	0.77	0.73	0.73

Table 18. Zero-shot classification AUC across various medical datasets for models in RQ3.

Dataset	**Image Captioning**	**Fast CL 25%**	**Fast CL 50%**	**Fast CL 75%**	**Fast CL 85%**	**Dataset**	**Image Captioning**	**Fast CL 25%**	**Fast CL 50%**	**Fast CL 75%**	**Fast CL 85%**
sicap	0.70	0.75	0.67	0.67	0.64	dermamnist+	0.60	0.55	0.60	0.60	0.52
pad_ufes_20	0.66	0.60	0.57	0.54	0.52	octmnist+	0.57	0.65	0.56	0.56	0.42
skin_cancer	0.67	0.67	0.59	0.59	0.57	pneumoniamnist+	0.90	0.90	0.80	0.79	0.52
pcam	0.61	0.75	0.61	0.61	0.78	retinamnist+	0.50	0.55	0.50	0.50	0.60
nct_crc_he_100k	0.91	0.90	0.91	0.91	0.80	breastmnist+	0.71	0.66	0.53	0.53	0.56
lc25000_lung	0.95	0.98	0.91	0.96	0.95	bloodmnist+	0.60	0.56	0.66	0.66	0.69
lc25000_colon	0.99	0.99	0.99	0.99	0.96	tissuemnist+	0.53	0.47	0.50	0.50	0.57
bach	0.72	0.77	0.72	0.72	0.77	organamnist+	0.78	0.82	0.81	0.82	0.81
pathmnist+	0.84	0.84	0.85	0.87	0.76	organcmnist+	0.76	0.76	0.74	0.79	0.77
chestmnist+	0.50	0.50	0.50	0.50	0.50	organsmnist+	0.78	0.79	0.75	0.79	0.76

Table 19. Zero-shot classification F1-score across various medical datasets for models in RQ3.

Dataset	**Image Captioning**	**Fast CL 25%**	**Fast CL 50%**	**Fast CL 75%**	**Fast CL 85%**	**Dataset**	**Image Captioning**	**Fast CL 25%**	**Fast CL 50%**	**Fast CL 75%**	**Fast CL 85%**
sicap	0.29	0.33	0.29	0.22	0.24	dermamnist+	0.05	0.08	0.04	0.14	0.12
pad_ufes_20	0.18	0.11	0.12	0.10	0.07	octmnist+	0.12	0.21	0.10	0.10	0.18
skin_cancer	0.16	0.16	0.13	0.12	0.13	pneumoniamnist+	0.78	0.40	0.71	0.71	0.41
pcam	0.47	0.62	0.45	0.45	0.34	retinamnist+	0.14	0.07	0.14	0.14	0.12
nct_crc_he_100k	0.53	0.51	0.53	0.59	0.39	breastmnist+	0.42	0.52	0.21	0.21	0.43
lc25000_lung	0.62	0.84	0.53	0.79	0.79	bloodmnist+	0.49	0.05	0.53	0.02	0.02
lc25000_colon	0.95	0.96	0.53	0.91	0.84	tissuemnist+	0.08	0.02	0.12	0.03	0.06
bach	0.32	0.33	0.29	0.28	0.49	organamnist+	0.13	0.12	0.12	0.20	0.27
pathmnist+	0.51	0.33	0.51	0.51	0.28	organcmnist+	0.13	0.14	0.13	0.18	0.22
chestmnist+	0.10	0.10	0.10	0.10	0.10	organsmnist+	0.13	0.14	0.13	0.17	0.20

Table 20. Visual question answering Accuracy score across two medical datasets for models in Table 1.

Vision encoders	**VQARAD**			**PathVQA**			
	Open-ended	Close-ended	Overall	Yes/No	Number	Other	Overall
RN50	**45.81**	73.53	62.53	82.31	**33.33**	**10.23**	46.44
ViT-B/16	42.46	74.26	61.64	**83.19**	27.78	9.99	**46.75**
ViT-B/32	44.13	**75.37**	**62.97**	82.63	27.78	9.7	46.33

Table 21. Visual question answering Accuracy score across two medical datasets for models in RQ1.

Models	**VQARAD**			**PathVQA**			
	Open-ended	Close-ended	Overall	Yes/No	Number	Other	Overall
Baseline	42.46	74.26	61.64	83.19	27.78	9.99	46.75
Text Full Freeze	40.78	72.47	59.87	82.96	27.78	10.11	46.69
Text Partial Freeze (3 Layers Unlocked)	41.34	75.74	62.08	**83.81**	**33.33**	**10.14**	**47.15**
Text Partial Freeze (6 Layers Unlocked)	43.58	**76.10**	**63.19**	82.6	**33.33**	9.67	46.31
Text Partial Freeze (9 Layers Unlocked)	**46.93**	73.90	**63.19**	82.37	27.78	8.92	45.81
Image Full Freeze	38.55	73.90	59.87	82.1	**33.33**	10.11	46.28
Image Partial Freeze (3 Layers Unlocked)	36.87	**76.10**	60.53	83.52	27.78	9.55	46.69
Image Partial Freeze (7 Layers Unlocked)	41.9	74.26	61.42	82.54	27.78	9.25	46.06
Image Partial Freeze (9 Layers Unlocked)	40.22	73.53	60.31	83.28	**33.33**	9.84	46.74

Table 22. Visual question answering Accuracy score across two medical datasets for models in RQ2.

Models	**VQARAD**			**PathVQA**			
	Open-ended	Close-ended	Overall	Yes/No	Number	Other	Overall
Baseline	42.46	74.26	61.64	**83.19**	27.78	9.99	46.75
Augmented CL	46.37	**75**	63.64	82.54	**33.33**	9.61	46.25
Masked CL	**49.16**	74.63	**64.52**	83.13	**33.33**	**10.23**	**46.86**

Table 23. Visual question answering Accuracy score across two medical datasets for models in RQ3.

Models	**VQARAD**			**PathVQA**			
	Open-ended	Close-ended	Overall	Yes/No	Number	Other	Overall
Baseline	42.46	74.26	61.64	83.19	27.78	9.99	46.75
FastCL 25%	41.90	75.37	62.08	83.19	27.78	10.08	46.80
FastCL 50%	40.22	75.74	61.64	82.31	27.78	9.88	46.25
FastCL 75%	**45.25**	74.63	**62.97**	83.10	27.78	10.17	46.80
FastCL 85%	40.78	69.85	58.31	**83.43**	27.78	9.70	46.72
Image captioning	41.34	**76.47**	62.53	82.6	**61.11**	**11.84**	**47.46**