Uni-Mlip: Unified Self-supervision for Medical Vision Language Pre-training

Ameera Bawazir
ameera.bawazir@tii.ae

Kebin Wu
Kebin.wu@tii.ae

Wenbin Li
wenbin.li@tii.ae

Technology Innovation Institute
Abu Dhabi, UAE

Abstract

Recent advancements in vision-language pre-training via contrastive learning have significantly improved performance across computer vision tasks. However, in the medical domain, obtaining multimodal data is often costly and challenging due to privacy, sensitivity, and annotation complexity. To mitigate data scarcity while boosting model performance, we introduce **Uni-Mlip**, a unified self-supervision framework specifically designed to enhance medical vision-language pre-training. Uni-Mlip seamlessly integrates cross-modality, uni-modality, and fused-modality self-supervision techniques at the data-level and the feature-level. Additionally, Uni-Mlip tailors uni-modal image self-supervision to accommodate the unique characteristics of medical images. Our experiments across datasets of varying scales demonstrate that Uni-Mlip significantly surpasses current state-of-the-art methods in three key downstream tasks: image-text retrieval, image classification, and visual question answering (VQA).

1 Introduction (Uni-Mlip)

Vision-and-Language Pre-training (VLP) techniques, such as the Contrastive Language Image Pre-training (CLIP) model [35], have set a foundational approach for leveraging self-supervision with language guidance to integrate visual and textual data effectively. By aligning image and text representations through contrastive learning, CLIP improves the model's ability to interpret complex data by linking visual concepts to linguistic descriptions. This method significantly enhances the understanding of multimodal representations by pre-training on extensive datasets and subsequently fine-tuning on targeted downstream tasks. This paradigm shift is particularly vital in the medical domain, where the acquisition of multimodal medical data presents significant challenges due to concerns over data privacy, sensitivity, and the complex, domain-specific knowledge required for annotation.

Medical Vision-and-Language Pre-training (Med-VLP) aims to address such challenges inherent to the medical imaging field. The exploration of self-supervised learning (SSL) techniques in medical VLP, as seen in models like ConVIRT [45] and GLoRIA [17], high-light the shift towards leveraging unlabelled data, indicating a growing emphasis on models

that can learn from the nuanced and complex nature of medical images and texts without extensive manual annotations. Despite advancements, previous approaches still lack the ability to systematically integrate and contextualize medical-specific knowledge, which leads to limited multimodal understanding and contextual reasoning.

In response to this, our paper introduces **Uni-Mlip**, a novel **Uni**fied **M**edical **CLIP** framework designed to improve medical vision-language pre-training with enhanced self-supervision. Uni-Mlip seamlessly integrates cross-modal, uni-modal, and fused-modal self-supervision techniques at both data-level and feature-level. This approach allows for more effective and nuanced mining of medical data, facilitating the learning of aligned and transferable features that excel in downstream tasks. Our main contributions include: **(1)** Uni-Mlip is the first unified framework that systematically explores feature-level and data-level self-supervision across both uni-modality and multimodality settings to cohesively align image and text modalities, to the best of our knowledge; **(2)** Uni-Mlip adapts image self-supervision techniques and tailors them to the unique requirements for higher precision and detail sensitivity characteristic [18] of medical images; **(3)** Comprehensive experiments demonstrate Uni-Mlip's superiority over current state-of-the-art methods, showcasing its effectiveness in tasks including image-text retrieval, image classification, and visual question answering (VQA).

2   Related work

The CLIP model [35] pioneers vision-language pre-training by aligning image-text pairs by maximizing the similarity between image and text features in a shared latent space. Subsequent models build upon CLIP by enhancing learning objectives [42, 43], modifying pre-training architecture [20, 39], and expanding downstream tasks [25, 44], which allow for better representation learning and widen the scope of vision-language pre-training.

Medical Vision-and-Language Pre-Training (Med-VLP). The models above optimized for natural images often fall short in medical settings: medical images necessitate a meticulous focus on fine-grained details, and critical diagnostic information is often embedded in the absolute pixel intensity, the scale of abnormalities, and their precise locations within the image. Hence, domain-specific Med-VLPs have been proposed to overcome these challenges. Medical VLP models such as GLoRIA [16], LoVT [32], and PRIOR [5] introduced global-local image-text alignment, enabling the learned models to capture fine-grained information lying in the medical images. PMC-CLIP [27] presents a CLIP-like model architecture that is trained on a large-scale dataset of medical image-caption pairs, highlighting the critical role of extensive datasets in effective pre-training. A significant contribution of their work is the introduction of PMC-OA, a newly curated dataset of image-caption pairs sourced from biomedical literature available on PubMedCentral's OpenAccess.

Self-Supervised learning (SSL). Self-Supervised Learning [1, 2, 10] addresses the issue of data annotation scarcity by training models to learn meaningful representations with unlabelled data. For natural images, adding traditional SSL to the CLIP framework for joint training can significantly improve the performance, whereas the study in [15, 41] found the opposite when directly trained on medical images. Firstly, the inherent assumptions of conventional SSL, i.e., image augmentations do not alter the semantic information of an image, may not hold when applied to medical images. This requires careful adaptation of augmentations to maintain the diagnostic integrity of the image. Secondly, conventional SSL generally applies augmentation techniques to the input space only. We argue the feature-level

SSL, through perturbation in the representation space, can further boost the performance and lead to more robust and more discriminative learning results. To this end, other approaches such as M3AE [3] and MUMC [26] respectively added a self-supervised learning paradigm to enhance the Med-VLP performance. M3AE reconstructed missing pixels and tokens in randomly masked images and texts using visual and textual features from different layers; MUMC utilized momentum contrastive self-supervised approach for uni-modal and multimodal representations learning. However, despite achieving state-of-the-art performance on the VQA task, both M3AE and MUMC have notable limitations. Firstly, MUMC is specifically designed for the VQA task, limiting its generalizability to other tasks. Secondly, M3AE and MUMC integrate several complex components, including heavy multimodal fusion modules, to create intensive interactions between image and text modalities. While this complexity enhances performance on VQA, it does not translate as effectively to other tasks, such as image-text retrieval, where M3AE, in particular, underperforms. Additionally, both models rely on large model sizes for training and inference. The integration of additional components on top of the pre-trained encoders during the inference stage, instead of adding a simple classification layer, significantly increases the model parameters. Moreover, the reliance on masked inputs for both modalities in M3AE and MUMC can potentially weaken their zero-shot capabilities when evaluated on unmasked data.

In contrast, our proposed method, Uni-Mlip, aims to provide a general-purpose Med-VLP model capable of consistently achieving strong performance across various downstream tasks. It does so by incorporating both data-level and feature-level SSL in cross-modality, uni-modality, and fused-modality settings, tailored specifically to medical data, while maintaining the simplicity of the original CLIP architecture.

3 Uni-Mlip model

In this section, we present a comprehensive overview of our proposed methodology Uni-Mlip, encompassing a detail explanation of our model architecture and pre-training objectives. As shown in Figure 1, in our pre-training model design, we follow a CLIP-like architecture that separately encodes each modality with a modality-specific encoder, and then jointly optimize both of them through image-text contrastive (**ITC**) loss, ensuring an alignment of image and text embedding. Let V be an input image, and C be the corresponding caption, given a batch of image-caption pairs $(V_1, C_1),....,(V_B, C_B)$, where $V_i \in \mathbb{R}^{H \times W \times C}$ is the i-th image and C_i is the paired caption, we encode the input image via the image encoder E_V and the input text via the text encoder E_T to get the image and text features respectively. These features are then projected to a shared embedding space and normalized to get the final image embedding $I \in \mathbb{R}^{B \times D}$ and text embeddings $T \in \mathbb{R}^{B \times D}$, where B is the batch size and D is the shared embedding dimension.

In Uni-Mlip, we integrate several complementary Self-Supervised Learning (SSL) techniques into the architecture reminiscent of CLIP. This involves incorporating cross-modality SSL at both the input and feature levels to align the embedding of images and text. Additionally, we introduce fused-modal and uni-modal SSL for text and images respectively to enhance robustness even further.
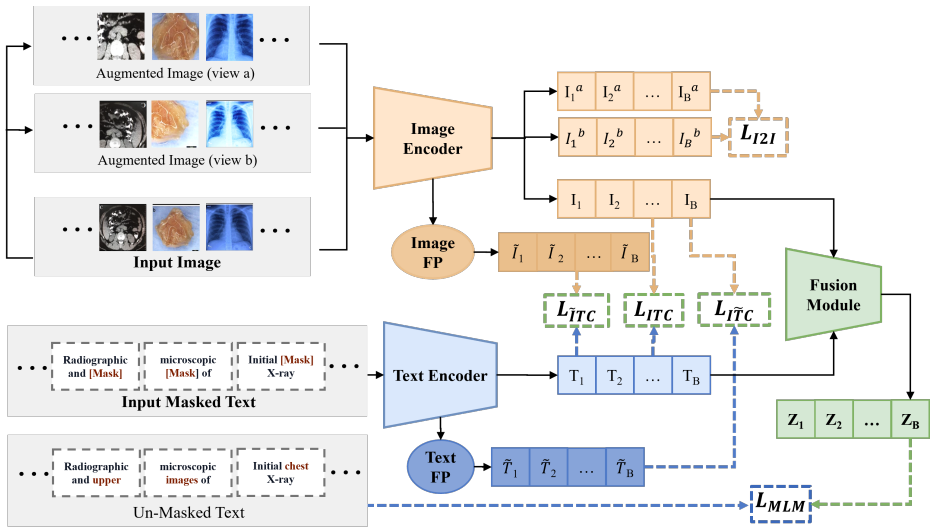
Figure 1: Illustration of our proposed model architecture Uni-Mlip. The colors indicate modality-specific components such that orange is for image, blue is for text, and green is for multimodal based components.

3.1 Cross-modal input level self-supervision

We perform an input level SSL by contrasting the image and text embeddings using the Image-Text Contrastive (ITC) loss from CLIP [65]. The ITC loss is defined as a sum of image-to-text L_{I2T} and text-to-image L_{T2I} NT-Xent losses [37], which is formulated as:

$$\mathcal{L}_{ITC} = -\frac{1}{2B}\sum_{i=1}^{B} log \frac{e^{sim(I_i,T_i)/\tau}}{\sum_{j=1}^{B} e^{sim(I_i,T_j)/\tau}} - \frac{1}{2B}\sum_{i=1}^{B} log \frac{e^{sim(I_i,T_i)/\tau}}{\sum_{j=1}^{B} e^{sim(I_j,T_i)/\tau}} \qquad (1)$$

The goal of the ITC loss is to maximize the similarity between the embedding of the B image-text pairs i.e., (I_i, T_i) while minimizing the similarity with the rest of the $B^2 - B$ non-pair samples i.e., (I_j, T_i) in the batch, where $i \neq j$. The temperature τ is a learnable parameter for scaling the measured similarity. The image embedding I and text embedding T are obtained via the image and text encoder, and *sim* is the cosine similarity metric.

3.2 Cross-modal feature level self-supervision

To further boost the cross-modality alignment, we also introduce a feature-level self-supervision by performing a feature perturbation. The motivation is that image/text features after some minor perturbations are expected to remain semantically aligned with the corresponding pair of text/image embeddings, ensuring robust and better representation learning. The feature perturbation can be implemented by a simple Dropout [14] or its variant such as DropBlock [9]. Perturbation applied to the image features is done after the last residual layer and before the attention pooling layer in the vision encoder, where a DropBlock method is utilized. On the other hand, to generate the perturbed text features, a dropout is applied after getting the

text features from the text encoder. In Figure 1, we refer to the DropBlock operation as Image FP, and to the dropout operation as Text FP.

Regarding the training objectives, we align the image/text features with the perturbed features of text/image using contrastive learning. Hence, the ITC loss with perturbed image features is denoted as $\mathcal{L}_{\tilde{I}TC}$, where the image embedding I from $\mathcal{L}_{ITC}$ is replaced with the perturbed image embedding $\tilde{I}$. Similarly, we define the ITC loss with the perturbed text embeddings as $\mathcal{L}_{I\tilde{T}C}$, such that the text embedding T in $\mathcal{L}_{ITC}$ equation is replaced with perturbed text embedding $\tilde{T}$.

3.3 Uni-modal self supervision - image

Based on the observation [31] that combining image-only self-supervised learning with language supervision can effectively improve the representation quality of the images, we move one step further and add uni-modal, image-only or text-only, self-supervision on top of cross-modal supervision. While reconstructing masked images is a popular approach for natural images [13], this method may be less suitable for medical imaging. This is because medical images often depict localized anomalies associated with diseases, and masking these parts could hinder the model's ability to accurately represent these localized anomalies. Thus, we adopt the contrastive self-supervised learning objective from SimCLR [2], which enables better capturing of subtle patterns and anomalies in medical images.

Given a batch size B of images, we create two strongly augmented views for each image. Then the uni-modal contrastive loss $\mathcal{L}_{I2I}$ is defined in Eq. 2 to maximize the similarity between the embedding of the positive pairs (views of the same image) while minimizing the similarity between the negative pairs (views of different images). Such loss term promotes the model to learn semantically robust representations for images with different augmentations.

$$\mathcal{L}_{I2I} = -\frac{1}{2B}\sum_{i=1}^{B} log \frac{e^{sim(I_i^a,I_i^b)/\tau}}{\sum_{j=1}^{B} e^{sim(I_i^a,I_j^b)/\tau}} - \frac{1}{2B}\sum_{i=1}^{B} log \frac{e^{sim(I_i^a,I_i^b)/\tau}}{\sum_{j=1}^{B} e^{sim(I_j^a,I_i^b)/\tau}} \tag{2}$$

where I_i^a is the feature embedding of the augmented image V^a, and I_i^b is the feature embedding of the augmented image V^b.

While SLIP [31] empirically demonstrates SimCLR's enhancement of vision-language pre-training with natural images, simply adding Eq. 2 worsens the CLIP performance in the medical domain [15]. One reason may be the fundamental differences in characteristics between natural and medical images. Unlike natural images, for which relative intensity information is key in recognition, medical images rely heavily on precise absolute intensity values. When integrating SimCLR into CLIP-like architecture, the same vision encoder will process three different views for each input image: one weakly augmented view used for CLIP, and two views with strong augmentation for SimCLR. When the vision encoder, such as a ResNet, includes batch normalization layers, strong image augmentations will significantly alter the mean and variance statistics in batch normalization layers, harming the intensity information critical for medical images. To address this issue, we propose to train without SimCLR for some epochs first, and then we freeze the mean and variance in the batch normalization layers while keeping the weight and bias as learnable parameters when incorporating the SimCLR loss (Eq. 2). This approach ensures that strong augmentations introduced by SimCLR will not alter the mean and variance in batch normalization layers, al-

lowing for better vision representation learning for medical images. Experiments in Section 5.2 support our analysis and prove the effectiveness of our proposed solution.

3.4   Fused-modal self-supervision

To enhance text representation, we integrate fused-modal self-supervision for the textual modality. Motivated by MMBERT [22] and PMC-CLIP [27], we utilize masked language modeling (MLM) self-supervision task for the input captions. In masked language modeling, the model predicts the masked text token by learning from the medical text features while using the image feature as a supplementary context. Before being fed to the text encoder, the captions are masked randomly with a probability of 15%. Following BERT [6], our masked tokens are replaced with a unique token [MASK]. After that, the visual and masked linguistic features obtained from the image and text encoder are fused by a multimodal transformer-based fusion module. The fusion module generates the full caption by replacing the [mask] token with the predicted text token. Hence the MLM loss is defined as:

$$\mathcal{L}_{MLM} = \mathbb{E}_{(V,C)\sim D}[CE(y^{mask}, p^{mask}(V,C)] \tag{3}$$

where y^{mask} denotes the ground truth and p^{mask} is the predicted masked token generated by the fusion module. The masked language modeling objective enables the model to focus on learning fine-grained details from the medical text to improve the representation and understanding of language-specific information.

3.5   Total training loss

The total loss is a weighted sum of all of the individual cross-modal, uni-modal and fused-modal losses and therefore is formulated as:

$$\mathcal{L}_{Total} = \lambda_{cm} \cdot (\mathcal{L}_{ITC} + \mathcal{L}_{\bar{I}TC} + \mathcal{L}_{I\bar{T}C}) + \lambda_{um} \cdot (\mathcal{L}_{I2I}) + \lambda_{fm} \cdot (\mathcal{L}_{MLM}) \tag{4}$$

where λ_{cm} , λ_{um}, and λ_{fm} refer to the weight value assigned to the cross-modal, uni-modal, and fused-modal losses respectively during training.

4   Experiment settings

We pre-train the Uni-Mlip on a large-scale dataset and then evaluate it on three downstream tasks. We provide the dataset and implementation details below.

4.1   Dataset

For pre-training, we use PMC-OA [27], a vast open-source medical image-caption dataset comprising 1.65M pairs sourced from PubMedCentral's OpenAccess.

For image-text retrieval evaluation, including both image-to-text (I2T) and text-to-image (T2I) retrieval tasks, we employ the Radiology Objects in COntext (ROCO) dataset [34], where 2000 samples are randomly selected in evaluation following [3, 27, 58]. As there is no overlapping between ROCO and PMC-OA, such a task also serves as a zero-shot evaluation for Uni-Mlip. For classification, we assess our model on three datasets: NIH [41], CheXpert [19] and MIMIC [21], consisting of 112,120, 224,316, and 377,110 chest X-ray images

respectively with various disease labels. Additionally, for the downstream task VQA, we use VQA-RAD [24] with 315 images and 3,515 questions, and SLAKE [29] with 642 images and 14,028 questions, both including closed-ended and open-ended questions on radiology images.

4.2 Implementation details

Implementation Details of Model Pre-training. Following [27], we adopt a modified ResNet50 [12] architecture as a vision encoder E_V, and PubMedBERT [11] as our text encoder E_T. Our fusion module employs four transformer layers to concatenate image and text embedding. Perturbation applied to image features uses DropBlock [9] with a 50% probability with a block size of 3x3, while perturbation on text features are carried out by using dropout [14] with a 75% probability. For the image-only self-supervision branch, the implementation of the SimCLR [2] augmentation includes color jittering, random gray-scale conversion, random Gaussian blurring, and horizontal flipping. All images are resized to 224x224 and trained with a batch size of 768 across 8 NVIDIA V100 GPUs for 100 epochs. The training employs a learning rate of 1e-4, with cosine scheduling and AdamW optimizer. For the training loss weights, we set $\lambda_{cm} = \frac{0.5}{3} = 0.167$, $\lambda_{um} = 0.5$, and $\lambda_{fm} = 0.5$ as default values, which are empirically determined.

As discussed in Section 3.3, freezing batch normalization is required to incorporate image-only self-supervision into the framework. Therefore, we initialize the three primary components—vision encoder, text encoder, and fusion module—with pre-trained weights from the 100th epoch of the PMC-CLIP model and freeze the mean and variance in the batch normalization layers during subsequent training. To ensure a fair comparison with PMC-CLIP, we re-implement PMC-CLIP by extending the model training to 200 epochs, thus doubling the original duration while retaining identical hyper-parameters.

Implementation Details of Downstream Tasks. For cross-modal retrieval, we randomly select 2000 test image-text pairs following [3, 27] and resize the images to 224x224. For image-to-text retrieval, we calculate the inner product between the image feature and each text feature in the test set. The text with the highest inner product value is deemed the predicted text label for the image. The recall is then computed for the top 1, 5, and 10 captions. The text-to-image retrieval process is similar.

Regarding the classification task, we follow [36] for dataset splits to mitigate comparison bias. With the pre-trained modified ResNet50 in Uni-Mlip, we add a linear classification layer on top for predictions. Input images are resized to 224x224, with batch sizes of 190 for CheXpert, and 150 for MIMIC and NIH. We perform end-to-end fine-tuning with binary cross-entropy loss and Adam optimizer, and the training is up to 64 epochs with early stopping criteria. The performance is evaluated with the macro Area Under the ROC Curve (AUC) metric.

For visual question answering (VQA), we cast the VQA as a classification task following the works in [3] and [7]. Instead of appending lots of new self-attention layers as in [3] and [7] to fuse the vision and text representations, we simplify the model framework with four components: a vision encoder, a text encoder, a one-layer transformer to fuse vision and text features, and a two-layer MLP to serve as the classification head. In addition, we append 20 learnable tokens in encoding questions, which can improve the representation power of text embedding effectively. The vision and text encoders are initialized with our pre-trained models and then trained with a learning rate of 5e-5. Meanwhile, the fusion and classification modules are randomly initialized and trained with a higher rate of 5e-4.
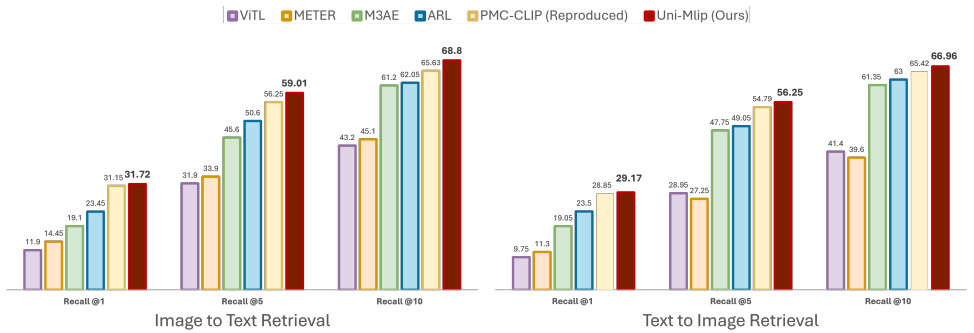
Figure 2: Evaluation Results on Image-to-Text (left) and Text-to-Image (right) retrieval tasks on ROCO dataset.

5   Empirical results

In this section, we quantitatively demonstrate Uni-Mlip's superiority in image-text retrieval, image classification, and VQA, and provide comprehensive ablation studies to highlight the effectiveness of each component in our unified self-supervised framework.

5.1   Downstream task results

Image-Text Retrieval For image-text retrieval, we compare our results with ViTL[23], and METER [7], M3AE [3], ARL [4], and our own reproduction of PMC-CLIP [27]. As shown in Figure 2, Uni-Mlip consistently outperforms the others under both image-to-text and text-to-image tasks across Recall@1, Recall@5, and Recall@10 metrics. In particular, in image-to-text retrieval, Uni-Mlip surpasses the previous state-of-the-art by 3.17% in Recall@10 and achieves a 2.76% improvement in Recall@5. Similarly, in text-to-image retrieval, the model outperforms PMC-CLIP with gains of 1.46% in Recall@5 and 1.54% in Recall@10.This task, regarded as a zero-shot evaluation, shows that our method with cross-modality, uni-modality, and fused-modality SSL objectives at both the input and feature levels achieves superior alignment between medical image and text, facilitating better retrieval performance.
Medical Image Classification For image classification, we compare the performance of our solution with previous state-of-the-art medical vision-language models. As shown from Table 1, the AUC scores indicate that Uni-Mlip achieves the best results for different data splits (1%, 10%, and 100%) on MIMIC [21], CXP [19] and NIH [41] datasets. Notably, when using 10% of the data, Uni-Mlip achieves an AUC of 79.1% on MIMIC, a gain of 1.7% over the previous best performance of 77.41% by PMC-CLIP. Our model consistently outperforms baselines that do not employ image uni-modal self-supervision, which further validates the importance of incorporating uni-modal image self-supervision to improve the visual representations, resulting in a better performance on medical image classification tasks.
Medical Visual Question Answering For the VQA task, we report each dataset's open-ended, close-ended, and overall accuracy. Table 2 demonstrates the superiority of Uni-Mlip in the overall evaluations. Moreover, Uni-Mlip gains consistently on closed-ended VQA with substantial improvement of 1.48% on VQA-RAD and 1.44 % on Slake over PMC-CLIP. Hence, Uni-Mlip excels at leveraging medical text and image data to answer visual questions accurately.

Method	MIMIC			CXP			NIH		
	1%	10%	100%	1%	10%	100%	1%	10%	100%
Random	53.6	66.5	78.2	62.6	69.0	76.9	56.4	67.1	76.9
ImageNet	67.8	70.5	79.3	63.7	70.7	77.7	59.7	68.9	78.1
ConVIRT [45]	67.8	73.4	80.1	63.2	71.3	77.7	60.0	69.0	76.6
GLoRIA [17]	67.5	72.6	80.1	62.9	69.0	77.8	60.1	71.2	77.7
MGCA [40]	68.4	74.4	80.2	63.4	72.1	78.1	61.1	67.8	77.3
M-FLAG [50]	69.5	74.8	80.2	64.4	71.4	78.1	62.2	71.6	78.7
PMC-CLIP [27] (reproduced)	73.1	77.4	81.8	69.1	74.9	79.1	64.9	76.3	82.3
Uni-Mlip (ours)	**73.2**	**79.1**	**82.0**	**69.1**	**75.3**	**79.8**	**65.6**	**76.4**	**82.9**

Table 1: Classification results on MIMIC, CXP and NIH dataset.

Methods	VQA-RAD			Slake		
	Open	Closed	Overall	Open	Closed	Overall
MEVF-BAN [53]	49.20	77.20	66.10	77.80	79.80	78.60
CPRD-BAN [28]	52.50	77.90	67.80	79.50	83.40	81.10
PubMedCLIP [8]	60.10	80.00	72.10	78.40	82.50	80.10
PMC-CLIP [27] (reproduced)	**61.45**	80.14	72.73	**80.16**	84.38	81.81
Uni-Mlip (ours)	60.43	**81.62**	**73.17**	79.38	**85.82**	**81.90**

Table 2: VQA results on VQA-RAD and Slake

While cross-modal and fused-modal interactions are crucial for integrating visual and textual information, uni-modal supervision provides a foundation for robust modality-specific representations. Our Uni-Mlip effectively leverages the benefits of cross-modal, uni-modal, and fused-modal self-supervision, enhancing the overall representation quality and task-specific performance. In addition, the superior performances of Uni-Mlip in the tasks of image classification and VQA, where finetuning is performed after initializing with the pre-trained Uni-Mlip, further indicate that the proposed framework also offers a robust starting point in the optimization space suitable for other task-oriented fine-tuning.

5.2 Ablation studies

To accelerate our analysis and reduce computational demands, we conducted ablation experiments using the ROCO dataset [54] instead of PMC-OA for pre-training, focusing on image-text retrieval tasks. Performance evaluation was conducted on both image-to-text (I2T) and text-to-image (T2I) retrieval tasks, reporting the top 1 recall value.

In Table 3, we illustrate the impact of freezing the mean and variance in the batch normalization layer when integrating SimCLR into the CLIP-based model for medical data. Consistently, V represents the input image, while V^a and V^b depict strongly augmented images used in SimCLR. Notably, the second row reveals a significant performance drop upon the naive addition of SimCLR, consistent with the observation in [15]. We further analyze this by feeding the augmented images in the forward pass while setting the weight of L_{I2I} loss to zero (third row). The still low performance reveals that the drop is due to feeding strongly augmented images in the forward pass. Consequently, we propose freezing batch statistics in batch normalization layers when incorporating the image-only SSL loss L_{I2I}, resulting in a noticeable improvement over the baseline as shown in the last row. This study emphasizes the importance of freezing batch normalization when integrating the SimCLR

Vision Encoder Input	Training Objectives	Freeze BN	I2T	T2I
V	L_{ITC}, L_{MLM}	✗	22.2	21.7
V, V^a, V^b	$L_{ITC}, L_{MLM}, L_{I2I}$	✗	5.0	1.0
V, V^a, V^b	L_{ITC}, L_{MLM}	✗	4.0	6.9
V, V^a, V^b	$L_{ITC}, L_{MLM}, L_{I2I}$	✓	25.7	24.0

Table 3: Ablation study on ROCO dataset showing the importance of freezing batch normalization when incorporating image-to-image self-supervised learning loss L_{I2I}.

L_{ITC}	L_{MLM}	$L_{I\tilde{T}C}$	$L_{\tilde{I}TC}$	L_{I2I}	I2T	T2I	Performance Gain (I2T/T2I)
✓	✓	✗	✗	✗	22.2	21.7	0.0 / 0.0
✓	✓	✓	✗	✗	23.2	23.3	+1.0 / +1.6
✓	✓	✗	✓	✗	23.8	23.4	+1.6 / +1.7
✓	✓	✗	✗	✓	25.7	24.0	+3.5 / +2.3
✓	✓	✓	✓	✓	**25.9**	**24.5**	+3.7 / + 2.8

Table 4: Ablation study on different pre-training objectives on ROCO dataset.

framework for Med-VLP.

Moreover, we demonstrate the effectiveness of our method by evaluating the significance of each learning objective. As demonstrated in Table 4, incorporating cross-modal self-supervision at the feature level yields notable performance gains of 1.0 and 1.6 in the I2T retrieval task, and 1.6 and 1.7 in the T2I retrieval task for $L_{I\tilde{T}C}$ and $L_{\tilde{I}TC}$, respectively. Additionally, integrating the L_{I2I} objective significantly boosts performance by 3.5 and 2.3 in the I2T and T2I tasks, respectively. Ultimately, the combination of all five training objectives results in superior performance across both tasks, highlighting the significance of our unified framework with cross-modal self-supervisions, both input-level and feature-level, as well as uni-modal and fused-modal self-supervisions.

6 Conclusion

We introduce Uni-Mlip in this paper, a unified self-supervision framework tailored for enhancing medical vision-language pre-training. Our simple yet effective approach proves the advantage of integrating cross-modality, uni-modality, and fused-modality self-supervision techniques. Additionally, our framework also unifies self-supervision from both the input level and the feature level. Moreover, our technique of freezing the mean and variance in the batch normalization layer is shown to be essential for integrating image-only contrastive learning in VLP on medical data.

Extensive experiments validate the effectiveness of our method, as it achieves SOTA performance across various downstream medical tasks. Uni-Mlip not only holds promise for advancing vision-language pre-training in clinical applications but also presents potential benefits for domains characterized by limited multimodal data, such as satellite and agriculture imagery analysis.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[3] Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 679–689, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-16443-9.

[4] Zhihong Chen, Guanbin Li, and Xiang Wan. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5152–5161, 2022.

[5] Pujin Cheng, Li Lin, Junyan Lyu, Yijin Huang, Wenhan Luo, and Xiaoying Tang. Prior: Prototype representation joint learning from medical images and reports. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21361–21371, 2023.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019. URL https://api.semanticscholar.org/CorpusID:52967399.

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

[8] Sedigheh Eslami, Gerard de Melo, and Christoph Meinel. Does clip benefit visual question answering in the medical domain as much as it does in the general domain?, 2021.

[9] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. *Advances in neural information processing systems*, 31, 2018.

[10] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

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

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[14] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *ArXiv*, abs/1207.0580, 2012. URL https://api.semanticscholar.org/CorpusID:14832074.

[15] Haoxu Huang, Samyak Rawlekar, Sumit Chopra, and Cem M Deniz. Radiology reports improve visual representations learned from radiographs. In *Medical Imaging with Deep Learning*, pages 1385–1405. PMLR, 2024.

[16] Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3922–3931, 2021. doi: 10.1109/ICCV48922.2021.00391.

[17] Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3942–3951, October 2021.

[18] Shih-Cheng Huang, Anuj Pareek, Malte Jensen, Matthew Lungren, Serena Yeung, and Akshay Chaudhari. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *npj Digital Medicine*, 6, 04 2023. doi: 10.1038/s41746-023-00811-0.

[19] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David Mong, Safwan Halabi, Jesse Sandberg, Ricky Jones, David Larson, Curtis Langlotz, Bhavik Patel, Matthew Lungren, and Andrew Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:590–597, 07 2019. doi: 10.1609/aaai.v33i01.3301590.

[20] Jiho Jang, Chaerin Kong, DongHyeon Jeon, Seonhoon Kim, and Nojun Kwak. Unifying vision-language representation space with single-tower transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37:980–988, 06 2023. doi: 10.1609/aaai.v37i1.25178.

[21] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1), December 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0322-0. URL http://dx.doi.org/10.1038/s41597-019-0322-0.

[22] Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. Mmbert: Multimodal bert pretraining for improved medical vqa. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1033–1036. IEEE, 2021.

[23] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021.

[24] Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1), November 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.251. URL http://dx.doi.org/10.1038/sdata.2018.251.

[25] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10955–10965, 2022. doi: 10.1109/CVPR52688.2022.01069.

[26] Pengfei Li, Gang Liu, Jinlong He, Zixu Zhao, and Shenjun Zhong. Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 374–383, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43907-0.

[27] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer, 2023.

[28] Bo Liu, Li-Ming Zhan, and Xiao-Ming Wu. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 210–220, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87196-3.

[29] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, April 2021. doi: 10.1109/isbi48211.2021.9434010. URL http://dx.doi.org/10.1109/ISBI48211.2021.9434010.

[30] Che Liu, Sibo Cheng, Chen Chen, Mengyun Qiao, Weitong Zhang, Anand Shah, Wen-jia Bai, and Rossella Arcucci. M-flag: Medical vision-language pre-training with frozen language models and latent space geometry optimization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 637–647. Springer, 2023.

[31] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022 - 17th European Conference, 2022, Proceedings*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 529–544, Germany, 2022. Springer Science and Business Media Deutschland GmbH. ISBN 9783031198083. doi: 10.1007/978-3-031-19809-0_30. Publisher Copyright: © 2022, The Author(s), under exclusive license to Springer Nature Switzerland AG.; 17th European Conference on Computer Vision, ECCV 2022 ; Conference date: 23-10-2022 Through 27-10-2022.

[32] Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rueckert. Joint learning of localized representations from medical images and reports. In *European Conference on Computer Vision*, pages 685–701. Springer, 2022.

[33] Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. Overcoming data limitation in medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 522–530. Springer, 2019.

[34] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M. Friedrich. Radiology objects in context (roco): A multimodal image dataset. In Danail Stoyanov, Zeike Taylor, Simone Balocco, Raphael Sznitman, Anne Martel, Lena Maier-Hein, Luc Duong, Guillaume Zahnd, Stefanie Demirci, Shadi Albarqouni, Su-Lin Lee, Stefano Moriconi, Veronika Cheplygina, Diana Mateus, Emanuele Trucco, Eric Granger, and Pierre Jannin, editors, *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 180–189, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01364-6.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

[36] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: proceedings of the Pacific symposium*, pages 232–243. World Scientific, 2020.

[37] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.,

2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/6b180037abbebea991d8b1232f8a8ca9-Paper.pdf.

[38] Sanjay Subramanian, Lucy Lu Wang, Sachin Mehta, Ben Bogin, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. Medicat: A dataset of medical images, captions, and textual references, 2020.

[39] Michael Tschannen, Basil Mustafa, and Neil Houlsby. Clippo: Image-and-language understanding from pixels only. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11006–11017, 2023. doi: 10.1109/CVPR52729.2023.01059.

[40] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning, 2022.

[41] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017. doi: 10.1109/cvpr.2017.369. URL http://dx.doi.org/10.1109/CVPR.2017.369.

[42] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19141–19151, 2022. doi: 10.1109/CVPR52688.2022.01857.

[43] Lewei Yao, Runhu Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *ArXiv*, abs/2111.07783, 2021. URL https://api.semanticscholar.org/CorpusID:244117525.

[44] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing XU, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 9125–9138. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/3ba960559212691be13fa81d9e5e0047-Paper-Conference.pdf.

[45] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022.