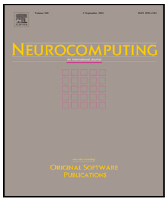
Neurocomputing 636 (2025) 130018

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Multimodal multitask similarity learning for vision language model on radiological images and reports

Yang Yu [a], Jiahao Wang [b], Weide Liu [a], Ivan Ho Mien [c,d], Pavitra Krishnaswamy [a,c], Xulei Yang [a,*], Jun Cheng [a]

[a] Machine Intellection Department, Institute for Infocomm Research (I^{2}R), Agency for Science, Technology and Research, (A*STAR), 1 Fusionopolis Way, #21-01 Connexis, Singapore, 138632, Singapore
[b] Mechanobiology Institute (MBI), National University of Singapore (NUS), 5A Engineering Drive 1, Singapore, 117411, Singapore
[c] Healthcare & Medtech Division, Institute for Infocomm Research (I^{2}R), Agency for Science, Technology and Research, (A*STAR), 1 Fusionopolis Way, #21-01 Connexis, Singapore, 138632, Singapore
[d] Department of Neuroradiology, National Neuroscience Institute (NNI), 11 Jln Tan Tock Seng, Singapore, 308433, Singapore

ARTICLE INFO

Communicated by D. Liu

Keywords:
Diagnostic imaging
Multimodal representation learning
Vision-language model

ABSTRACT

In recent years, large-scale Vision-Language Models (VLM) have shown promise in learning general representations for various medical image analysis tasks. However, current medical VLM methods typically employ contrastive learning approaches that have limited ability to capture nuanced yet crucial medical knowledge, particularly within similar medical images, and do not explicitly consider the uneven and complementary semantic information contained in different modalities. To address these challenges, we propose a novel Multimodal Multitask Similarity Learning (M2SL) method that learns joint representations of image–text pairs and captures the relational similarity between different modalities via a coupling network. Our method also notably leverages the rich information in the text inputs to construct a knowledge-driven semantic similarity matrix as the supervision signal. We conduct extensive experiments for cross-modal retrieval and zero-shot classification tasks on radiological images and reports and demonstrate substantial performance gains over existing methods. Our method also accommodates low-resource settings with limited training data availability and has significant implications for enhancing VLM development.

1. Introduction

The rapid growth of medical imaging datasets has accelerated the development of diverse deep-learning models to enhance clinical decision-making processes, especially for diagnostic radiology. However, annotating extensive medical imaging datasets demands specialized domain expertise and proves economically impractical at scale [1]. To address this challenge, a practical strategy entails leveraging insights from the associated medical reports containing comprehensive diagnoses of medical conditions as identified by radiologists [2]. Deep learning models that utilize multimodal data as inputs have drawn more attention in recent years, driven by the use of attention mechanisms or transformers structure [3–9]. However, image–text joint learning strategies are still needed for downstream tasks.

Accordingly, advances in Vision-Language Models (VLM) enable joint training of image and text on large-scale datasets to generate versatile and transferable representations for diverse downstream tasks such as cross-modal retrieval and zero-shot classification. Notably, these approaches bridge the gap between visual and linguistic understanding, as exemplified by the success of CLIP [30]. A summary of current medical VLM is shown in Fig. 1: (a) two encoders for images and reports followed by a contrastive learning-based similarity matrix [10–18]; (b) multiple encoders for augmented images and reports followed by both global and/or local contrastive learning-based similarity matrix [19–25,29]; (c) two encoders for images and reports followed by a matching loss-based similarity matrix with supervision signal from corresponding labels [26–28]. However, there are two fundamental issues with using paired medical images and reports. First, the contrastive learning-based approaches (a–b) attempt to draw together images and reports from the same patients while pushing apart those from different patients. But clinical images or reports unrelated to a given patient's studies may still display similar visual or textual patterns and hence encompass nuanced yet crucial medical knowledge, discounting such data from different patients may lead to

* Corresponding author.
E-mail addresses: yu_yang@i2r.a-star.edu.sg (Y. Yu), e0304997@u.nus.edu (J. Wang), weide001@e.ntu.edu.sg (W. Liu), Ivan_Ho@i2r.a-star.edu.sg (I.H. Mien), pavitrak@i2r.a-star.edu.sg (P. Krishnaswamy), yang_xulei@i2r.a-star.edu.sg (X. Yang), cheng_jun@i2r.a-star.edu.sg (J. Cheng).

https://doi.org/10.1016/j.neucom.2025.130018
Received 4 November 2024; Received in revised form 24 February 2025; Accepted 10 March 2025
Available online 18 March 2025

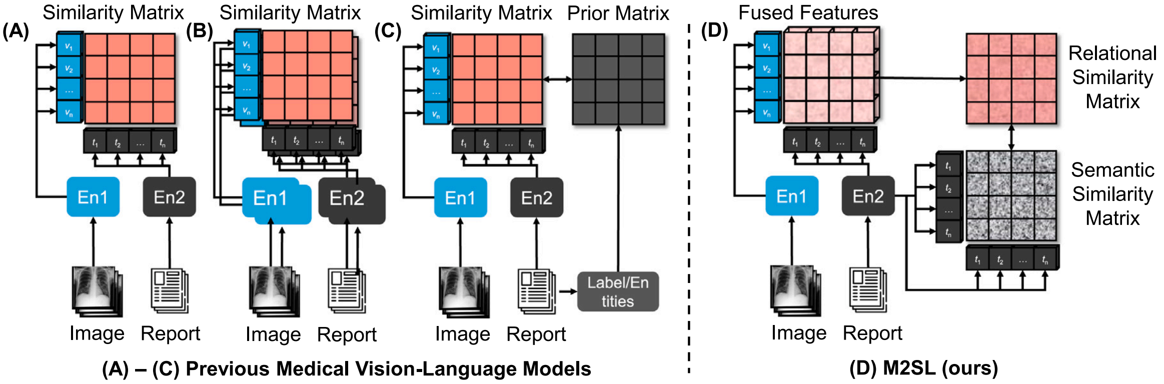


Fig. 1. Comparison of existing Vision-Language Models methods (a) [10–18], (b) [19–25], (c) [26–29], and (d) our proposed Multimodal Multitask Similarity Learning (M2SL) method. The proposed method leverages rich medical knowledge from the associated reports to build the knowledge-driven semantic similarity and exploits a coupling network to automatically assign weights to different modalities for learning the relational similarity.

false negatives [26,27,31]. Second, the assumption in the current VLM (a–c) that equal information can be extracted from various modalities during cross-modal learning does not generally hold as different modalities usually show complementary relationships, leading to an uneven distribution of information when representing the same scenario or semantics [32]. This means features specific to each modality cannot be perfectly matched across different modalities. Therefore, directly aligning modality-specific representations from various modalities in a unified space is inappropriate [33].

Recent developments in extracting semantic representations from diverse multimodal data [32–38] and using soft labels computed from text features instead of hard labels [39,40] also offer insights for soft cross-modal alignment. However, applying these in the context of radiological data requires overcoming significant disparity between the general image–text pairs and those found in the medical domain.

In this work, we introduce a novel method termed Multimodal Multitask Similarity Learning (M2SL) shown in Fig. 1d to address the false negatives and uneven and complementary semantic information challenges. We first leverage rich medical knowledge from the associated reports to build the knowledge-driven semantic similarity for fulfilling additional supervision and further exploit a coupling network to automatically assign various weights to different modalities for learning the relational similarity. Our method aims to address the aforementioned challenges by integrating and learning varying relationships between different modalities with a soft semantic matching loss, thereby enhancing the comprehensive understanding of medical images.

1. We present a novel multimodal multitask method that captures the relational similarity via a coupling network, allowing the model to implicitly account for the unequal distribution of information across distinct modalities.
2. We leverage the information in the text inputs to construct a knowledge-driven semantic similarity matrix as the supervision signal to equip the model with the ability to capture the subtle yet crucial medical knowledge.
3. We perform a series of experiments with various ablation studies on Chest X-rays to showcase the substantial performance enhancements over existing methods on cross-modal retrieval and zero-shot classification tasks. We also demonstrate the adaptability to practical scenarios characterized by limited access to training data.

2. Related work

2.1. Vision-language model on medical imaging data

Inspired by the recent advancements in CLIP [30] which jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (images, reports) training examples, ConVIRT [10] adopts an approach by jointly training vision and text encoders on paired images and reports, employing a bidirectional contrastive loss. GLoRIA [20] and MGCA [24] expand to encompass interactions at different levels between medical images and reports, enabling the extraction of pathological information from specific regions of the images. BioViL [12] utilizes a radiology-specific text encoder along with text augmentation and regularization techniques and thus preserves the quality of the language model by employing a masked language modeling loss on paired biomedical data. CXR-CLIP [21] and DeCLIP [22] expand image-label pairs into image–text pairs through a general prompt by leveraging multiple views and sections in reports to address the shortage of data. CARZero [29] employs cross-attention mechanisms to generate Similarity Representation (SimR) and utilizes a Large Language Model (LLM) to reformulate medical reports into a unified prompt template. MRT [41] focuses on the direct manipulation of multimodal representations within LMMs. Other studies such as MedViLL [18], MRM [11], REFERS [23], CMITM [13], and MPMA [14] have also further explored and extended self-supervised multimodal learning for fine-tuning and/or transfer learning using unmasked/masked image–report radiological data. Moreover, RadVLM [42] combines visual and textual data to perform tasks such as image captioning, question answering, and report generation, highlighting the potential of vision-language models in healthcare. Similarly, RadAlign [43] enhances radiology report generation by aligning visual and textual concepts, improving the model's comprehension of medical images and their descriptions for more accurate and contextually relevant reports. Despite these successes, conventional contrastive loss-based approaches may not have the capabilities to capture the nuanced yet crucial medical knowledge and simply treating the other reports as negative samples may introduce noise into the model learning process. MedCLIP [26] and SAT [27] address this issue through a customized matching loss and a reconstructed contrastive objective, using labels and extracted entities from reports as the supervisory signal. However, all those existing methods do not explicitly consider the uneven and complementary information contained in the two modalities and attempt to map distinct modalities onto a latent shared space with an equal amount of the information for direct comparing of representations across modalities by a widely used distance metric [44]. Recent research on prompt tuning in multimodal and visual models has also introduced several innovations. M2PT [45] integrates visual and textual prompts within vision encoders and language processors to enhance cross-modal feature alignment. E^{2}VPT [46] employs learnable key–value prompts in self-attention layers to improve fine-tuning efficiency. VFPT [47] leverages Fast Fourier Transform (FFT) in prompt embeddings, enabling feature extraction in both spatial and frequency domains. DPLNet [48] proposes a dual-prompt learning framework for efficient multimodal semantic segmentation by leveraging both visual

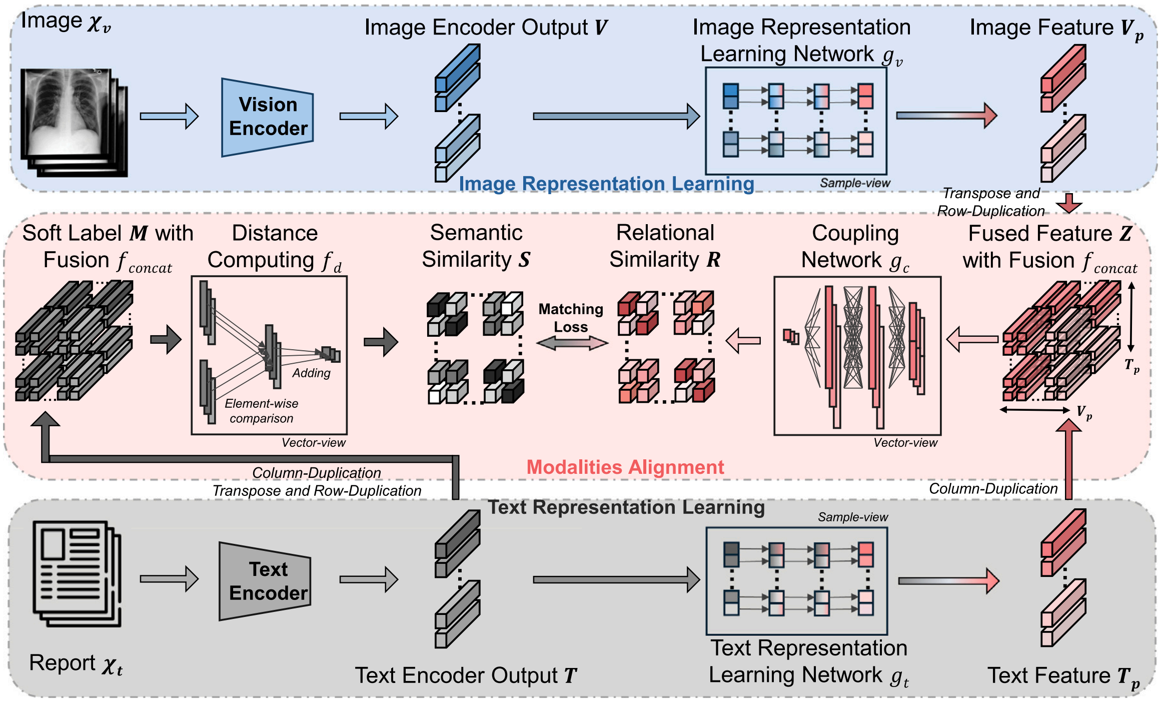


Image χ_v
Vision Encoder
Image Encoder Output $\boldsymbol{V}$
Image Representation Learning Network g_v
Image Feature $\boldsymbol{V}_p$
Image Representation Learning
Sample-view
Transpose and Row-Duplication
Soft Label $\boldsymbol{M}$ with Fusion f_{concat}
Distance Computing f_d
Adding
Element-wise comparison
Vector-view
Semantic Similarity $\boldsymbol{S}$
Matching Loss
Relational Similarity $\boldsymbol{R}$
Coupling Network g_c
Vector-view
Fused Feature $\boldsymbol{Z}$ with Fusion f_{concat}
$\boldsymbol{T}_p$
$\boldsymbol{V}_p$
Modalities Alignment
Column-Duplication
Transpose and Row-Duplication
Column-Duplication
Text Representation Learning
Sample-view
Text Encoder
Report χ_t
Text Encoder Output $\boldsymbol{T}$
Text Representation Learning Network g_t
Text Feature $\boldsymbol{T}_p$

Fig. 2. The workflow of the proposed Multimodal Multitask Similarity Learning (M2SL) method. The image and text representation learning modules first extract the corresponding features $\boldsymbol{V}_p$ and $\boldsymbol{T}_p$. Then, the relational similarity matrix $\boldsymbol{R}$ is computed via a coupling network using the further processed fuse feature $\boldsymbol{Z}$. Also, the semantic similarity matrix $\boldsymbol{S}$ is calculated by using the outputs $\boldsymbol{T}$ of the text encoder network as a soft label $\boldsymbol{M}$ for paired images and reports. It is expected that the relational similarity matrix $\boldsymbol{R}$ will closely approximate the semantic similarity matrix $\boldsymbol{S}$.

and textual prompts to understand and process multimodal inputs effectively. [49] examines when Visual Prompt Tuning (VPT) outperforms full fine-tuning for adapting large-scale vision models. FaST [50] introduces a dynamic system switch for task-dependent reasoning. While these advancements refine prompt tuning in VLM, their generalization capabilities across a wide range of tasks on medical datasets require further validation as medical datasets feature complex, domain-specific terminology that demands a deep understanding and context. Prompt tuning may struggle to capture the intricate relationships and specialized language in medical imaging, where precise comprehension of medical terms exceeds the capabilities of simple prompt-based learning.

2.2. Cross-modal retrieval for paired images and reports

Early retrieval methods in radiology depended on manually engineered characteristics, targeting specific regions highlighted by clinicians to locate database images sharing similar visual traits [51]. Recent progress has delved into broader applications of deep learning techniques for retrieval, focusing on similarities in modality or body parts [52]. Yet, these techniques remain unproven in complex clinical retrieval scenarios where user input is scarce, and they fail to leverage the potential of multimodal data within radiology databases. Conversely, recent literature has progressed deep learning techniques to effectively acquire semantic representations from diverse multimodal data sources, enhancing automated retrieval with access to large paired datasets for both images and reports. Significant strides have also been made in the radiology domain with cross-modal retrieval methods, where text has been employed to improve the retrieval performance [10,20,21,26,53–55]. Nonetheless, cross-modal retrieval in the broader general domain mainly focuses on global image regions and labels. In contrast, in the realm of medical imaging, where bodily or organ structures frequently exhibit resemblances among patients, nuanced details prove more crucial as markers of various diseases, yet they tend to be disregarded more easily [56].

2.3. Zero-shot classification

In zero-shot classification, the objective is to identify classes not encountered during training [57]. Generalized zero-shot classification expands to identify images across known and unknown domains. During the learning process, the model learns to link class attribute vectors with corresponding feature representations, establishing a dependable anchor for generating features of known and unknown classes alike. By utilizing the class attribute vector of the target class, it can create the corresponding feature representation, enabling effective cross-referencing [58]. However, medical images lack well-defined class attributes, as defining unambiguous attribute vectors for different disease classes requires significant clinical expertise and time, especially for unseen classes. Therefore, applying existing zero-shot methods to medical image classification is not straightforward. As an attempt to circumvent this problem, certain approaches leverage unlabeled data from unseen classes in a transductive manner [59,60]. However, the absence of supervised information from the unseen domain poses a notable hurdle in distinguishing between disease labels, particularly when many labels exhibit similar appearances.

3. Method

Our M2SL method illustrated in Fig. 2 achieves a robust cross-modal similarity metric by utilizing the proposed pairwise relational similarity and report-driven semantic similarity. This method effectively addresses cross-modal discrepancies and also ensures that the distinct characteristics of each modality are preserved by learning an optimal cross-modal similarity metric with additional supervision from rich information contained in the reports.

3.1. Visual and textual embedding learning

We consider both the image and its corresponding report as inputs. We denote the set of n samples of the image modality as $\mathcal{X}_v = \{x_v^1, x_v^2,$

$\dots, \boldsymbol{x}_v^n\}$, where $\boldsymbol{x}_v^i$ denotes the ith data input for the image modality $\boldsymbol{v}$. We also denote the set of n samples of the text modality $\boldsymbol{t}$ as $\mathcal{X}_t = \{\boldsymbol{x}_t^1, \boldsymbol{x}_t^2, \dots, \boldsymbol{x}_t^n\}$, where $\boldsymbol{x}_t^i$ denotes the ith data input for the text modality. It is noteworthy that both image and text modalities encompass the same categories.

For image modality, an image input $\boldsymbol{x}_v^i$ is firstly fed into the pre-trained image encoder to obtain the output embeddings $\boldsymbol{v}^i$ and then fed into the corresponding modality-specific representation learning network, denoted as $g_v(\cdot)$, to learn the highly nonlinear features. To construct the representation learning network, we utilize three fully connected layers, each followed by a Rectified Linear Unit (ReLU) activation function. This architecture enables the network to capture and model complex patterns in the data, enhancing its ability to learn and generalize from the input embeddings with same dimensions. The objective of this setup is to reduce complexity by consolidating various data types and promoting the learning of semantic representations for each type of data. We could denote the derived modality-specific feature vector for the image modality as $\boldsymbol{v}_p^i$:

$$\boldsymbol{v}_p^i = g_v(\boldsymbol{v}^i). \tag{1}$$

For text modality, medical reports $\boldsymbol{x}_t^i$ often contain lengthy paragraphs that demand reasoning over multiple sentences. To address this issue, we utilize a self-attention-based language model as a text encoder to obtain the output embeddings $\boldsymbol{t}^i$. Such model is adept at comprehending long-range semantic dependencies within these reports, ensuring a more nuanced and accurate interpretation of the text inputs. Similarly, three additional connected layers followed by a ReLU activation function are stacked on the text encoder to construct the representation learning network. We could denote the obtained modality-specific feature vector for the text modality as $\boldsymbol{t}_p^i$:

$$\boldsymbol{t}_p^i = g_t(\boldsymbol{t}^i). \tag{2}$$

We also denote the image and text embedding sets as $\boldsymbol{V}$ and $\boldsymbol{T}$ where $\boldsymbol{V} = \{\boldsymbol{v}^i\}_{i=1}^n$ and $\boldsymbol{T} = \{\boldsymbol{t}^i\}_{i=1}^n$, and further denote the image and text feature set as $\boldsymbol{V}_p$ and $\boldsymbol{T}_p$ where $\boldsymbol{V}_p = \left\{\boldsymbol{v}_p^i\right\}_{i=1}^n$ and $\boldsymbol{T}_p = \left\{\boldsymbol{t}_p^i\right\}_{i=1}^n$.

3.2. Relational similarity computation

Unlike existing methods [10,12,20,21,23,26,30] to directly use the obtained image/text features to compute the pairwise similarities, we propose to use a coupling network trained in a Multilayer Perceptrons (MLP) fashion for capturing the nonlinear metric and computing the pairwise similarities between paired data. Prior work on VLM primarily employs fixed distance metrics, such as Euclidean or cosine distance, for cross-modal alignment. These metrics assume elementwise feature comparison and linear separability after embedding, making performance heavily dependent on the learned embedding network. Consequently, their effectiveness is limited when embeddings are not sufficiently discriminative. In contrast, an MLP-based coupling network, with non-linear activation, jointly learns both a deep embedding and a flexible, non-linear similarity metric. This enables more robust identification of matching and mismatching pairs. We first employ concatenation operation as a fusion mechanism denoted as $f_{concat}(\cdot)$ to get the fused feature set $\boldsymbol{Z}$ for any pairwise cross-modal samples from image feature set $\boldsymbol{V}_p$ and the text feature set $\boldsymbol{T}_p$. We denote the fused feature set:

$$\boldsymbol{Z} = f_{concat}(\boldsymbol{V}_p^T, \boldsymbol{T}_p), \tag{3}$$

where the superscript ^{T} denotes the transpose of a matrix.

Finally, we exploit a four-layer MLP-based and trainable coupling network denoted as $g_c(\cdot)$ to get the pairwise relational similarity matrix $\boldsymbol{R}$ using the obtained fused feature set $\boldsymbol{Z}$:

$$\boldsymbol{R} = g_c(\boldsymbol{Z}). \tag{4}$$

The coupling network $g_c(\cdot)$ transforms each fused feature vector into a singular predicted similarity for the corresponding image–text pair, by directly modeling the pairwise similarity between different modalities and therefore bypasses the need to account for the unequal distribution of information across the distinct modalities. This approach also ensures that various weights could be automatically assigned to different modalities to exploit the uneven and complementary semantic information, instead of mapping distinct modalities into a latent shared space with an equal amount of the information.

3.3. Semantic similarity computation

To fulfill the additional supervision, we then leverage the rich information contained in the $\boldsymbol{T}$ to construct a knowledge-driven soft label set denoted as $\boldsymbol{M}$ for any pairwise cross-modal samples using the same fusion mechanism $f_{concat}(\cdot)$ as in Eq. (3), with the images and text feature sets replaced with the text embedding set:

$$\boldsymbol{M} = f_{concat}(\boldsymbol{T}^T, \boldsymbol{T}), \tag{5}$$

Instead of directly using the labels shared by paired images and texts, we measure their semantic similarity from the shared reports, via the direct element-wise distance calculation $f_d(\cdot)$ from the soft label set $\boldsymbol{M}$. We denoted the pairwise semantic similarity matrix as $\boldsymbol{S}$:

$$\boldsymbol{S} = f_d(\boldsymbol{M}). \tag{6}$$

3.4. Loss calculation

In the proposed method, the pairwise relational similarity matrix $\boldsymbol{R}$ is expected to approximate the semantic similarity matrix $\boldsymbol{S}$, and the objective function is formulated as follows:

$$\mathcal{L} = \|\boldsymbol{R} - \boldsymbol{S}\|_F^2, \tag{7}$$

where $\|.\|_F$ is the Frobenius norm. This loss function facilitates the training of the M2SL method through a back-propagation mechanism and a stochastic gradient descent-based optimization algorithm.

3.5. Performance evaluation

In the testing phase, we utilize the computed relational similarity to represent the distance between a query image $\boldsymbol{x}_v^{\boldsymbol{Q}}$ and every sample within the text database $\mathcal{X}_t^{\boldsymbol{D}}$. For cross-modal retrieval, we rank the database entries based on this distance, thereby obtaining the desired outputs. In zero-shot classification, building on the approach proposed in [20,30], we reformulate the image classification task as an image–text similarity measurement problem by converting classification labels into textual descriptions. Specifically, we collaborated with a radiologist to incorporate medical domain knowledge in generating meaningful textual representations for each classification category. These descriptions encompass relevant subtypes, severities, and anatomical locations associated with the medical conditions. To construct the textual prompts, we systematically generate class representations by randomly combining appropriate terms for subtypes, severities, and locations, ensuring comprehensive and clinically relevant descriptions.

4. Experiments and results

4.1. Datasets

We utilized multiple datasets and performed a range of experiments to evaluate the effectiveness of the proposed method across different tasks.

4.1.1. MIMIC-CXR dataset

MIMIC-CXR dataset is a comprehensive 2D chest X-ray repository with associated free-text radiology reports [61]. We used the stratified training split (157 392 paired image–text) of this dataset for feature representation learning and modality alignment.

Table 1
Performance comparison between proposed method (M2SL) and other state-of-the-art methods [10,12,20,21,23,24,26,29,30] for cross-modal retrieval (mAP: mean average precision) and zero-shot classification (Acc (Std): Accuracy with Standard Deviation) tasks on the MIMIC-5 × 200, CheXpert-5 × 200, and RSNA datasets.

Methods and tasks	Cross-modal retrieval			Zero-shot classification		
	MIMIC-5 × 200			MIMIC-5 × 200	CheXpert-5 × 200	RSNA
	mAP@1	mAP@5	mAP@10	Acc (Std)	Acc (Std)	Acc (Std)
CLIP [30]	0.169	0.412	0.388	0.206 (<0.01)	0.197 (<0.01)	0.457 (0.011)
ConVIRT [10]	0.465	0.539	0.538	0.438 (<0.01)	0.352 (<0.01)	0.774 (0.027)
GLoRIA [20]	0.467	0.564	0.550	0.475 (<0.01)	**0.450** (<0.01)	0.683 (<0.01)
MGCA [24]	0.471	0.574	0.554	0.480 (<0.01)	0.409 (<0.01)	0.762 (0.012)
BioViL [12]	0.473	0.577	0.556	0.485 (<0.01)	0.422 (<0.01)	0.771 (0.013)
MedCLIP [26]	0.476	0.580	0.559	0.471 (<0.01)	0.411 (<0.01)	0.818 (0.016)
REFERS [23]	0.524	0.599	0.586	0.495 (<0.01)	0.418 (<0.01)	0.780 (<0.01)
CXR-CLIP [21]	0.518	0.612	0.585	0.497 (<0.01)	0.359 (0.016)	0.769 (0.017)
CARZero [29]	0.521	0.610	0.587	0.502 (<0.01)	0.405 (<0.01)	0.783 (<0.01)
M2SL (ours)	**0.556**	**0.619**	**0.608**	**0.511** (<0.01)	0.407 (<0.01)	**0.831** (<0.01)

4.1.2. MIMIC-5 × 200 dataset

For evaluation purposes, we sampled a multi-class classification subset derived from the stratified testing split from MIMIC-CXR, following similar settings in [20]. This group consists of 1000 images that are exclusively positive for the five tasks: Atelectasis, Cardiomegaly, Edema, Pleural effusion, and Pneumonia.

4.1.3. CheXpert-5 × 200 dataset

We included the CheXpert dataset, which is a comprehensive compilation of 2D chest X-rays [62]. A CheXpert-5 × 200 subset was also sampled for the same five tasks as mentioned earlier for evaluation purposes.

4.1.4. RSNA pneumonia dataset

We also exploited the RSNA Pneumonia dataset, which encompasses pneumonia cases extracted from a public database of 2D chest X-rays [63]. The dataset is structured as a binary classification task, distinguishing between pneumonia and normal cases. We extracted a balanced subset of 1000 images with approximately a 1:1 ratio of positive and negative instances.

Our model is trained using the training images of MIMIC-CXR dataset, and tested on MIMIC-5 × 200, CheXpert-5 × 200, and RSNA datasets. The performance of our model on MIMIC-5 × 200 indicates the capability of the proposed approach for cross-modal retrieval using the image to search for similar reports. The performance of our model on MIMIC-5 × 200, CheXpert-5 × 200, and RSNA indicates the generalization ability and robustness of the proposed approach for prompt-based zero-shot classification.

4.2. Implementation details

The proposed M2SL method contains two parts, feature representation learning and modalities alignment. For feature representation learning, we used the DenseNet-121 [64] fine-tuned with selected training datasets as the backbone for the vision encoder, and CheXbert [65] with weights drawn from the transformer library as the backbone for the text encoder, followed by representation learning network to extract high-level representations from various modalities. We provided ResNet-50 [66] and Swin Transformer [67], which is in line with previous works [10,12,20,21,23,24,26,29,30] as the alternative vision encoders. We also provided BioClinicalBert [68], which is in line with previous works [10,20,21,23,24,26,29] as the alternative text encoders. For modalities alignment, we froze the encoders and utilized the resulting modality-specific representations to create pairwise samples via a fusion mechanism. These pairwise samples were then fed into the coupling network, generating a pairwise relational similarity. We utilized image augmentations to first scale the raw images to 224 × 224 and then applied a random horizontal flip with 0.5 probability. For other hyperparameters, optimizer selection was set to Adam, learning rate was initially set to 1e–4 with a beta of [0.5, 0.99], weight decay of 1e–4, and cosine annealing scheduler. The batch size was set to 128 for experiments with iterations set to 60 000. We trained our model with 14 million of trainable parameters on 160,000 paired images and reports using a single Nvidia RTX A5000 GPU, completing the process in approximately 4 h. For inference across various downstream tasks, it processes 1000 images in just less than 5 s on the same GPU.

4.3. Evaluation metrics

We investigated the performance of the proposed method for cross-modal retrieval and zero-shot image classification tasks. For the cross-modal retrieval, we reported top K mean Average Precision (mAP) scores for $K = 1/5/10$ (K is the number of retrieved cases). For the zero-shot classification using text prompts, we reported Accuracy (Acc) with the mean and Standard Deviation (Std) across three runs for the random prompt generation process.

4.4. Comparison with state-of-the-art methods

We compared the performance of our proposed M2SL method against other existing vision-language joint representation learning state-of-the-art (SOTA) methods, including CLIP [30], ConVIRT [10], GLoRIA [20], MGCA [24], BioViL [12], MedCLIP [26], REFERS [23], CXR-CLIP [21], and CARZero [29].

We first assessed the semantic richness of learned relational similarity using MIMIC-5 × 200 dataset within the context of the cross-modal retrieval task. Using the image as a query input, we evaluated the similarity between the queried image and all potential reports. We employed precision@K based on the alignment of the report category with that of the query image to compute the mean average precision of the top K retrieved reports. The obtained results highlight the better performance of our proposed M2SL method (Table 1), suggesting that our approach supports the necessary complementary semantic data for cross-modal retrieval.

We then conducted evaluation on three datasets: MIMIC-5 × 200, CheXpert-5 × 200, and RSNA for zero-shot classification, with results also shown in Table 1. We utilized trained image–text encoders and coupling networks to make zero-shot predictions by computing the relational similarity of aligned image features and the features of generated prompts for each disease category. Notably, for MIMIC-5 × 200, our approach demonstrates consistent performance improvements compared to all SOTA methods. The results indicate the effectiveness of employing prompt ensembles within the proposed M2SL method, leading to enhanced overall performance improvements. We observed that the zero-shot classification performance on CheXpert-5 × 200 is lower than that of GLoRIA. This discrepancy may be attributed to GLoRIA's explicit incorporation of a localized image–text attention mechanism, which likely enhances its ability to generalize to CheXpert's classification setup, given its distinct curation process. For RSNA, our model outperforms all the SOTA methods.

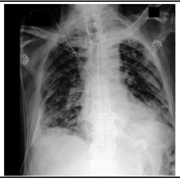
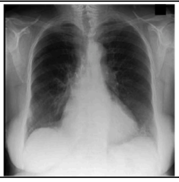
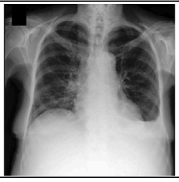
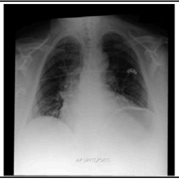
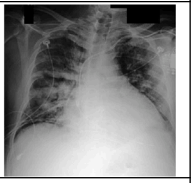
Dataset	MIMIC-5x200				
Classes	**Edema**	**Cardiomegaly**	**Pleural Effusion**	**Atelectasis**	**Pneumonia**
Image Examples					
Ground Truth (Report)	Severe pulmonary **edema** as seen on the recent CT scan.	Mild **cardiomegaly**. No acute cardiopulmonary process.	Slight interval increase in the left-sided **pleural effusion**.	Low lung volumes with mild bibasilar **atelectasis**.	Worsening pulmonary vascular congestion and Multifocal **pneumonia**.
REFERS – 1st Retrieved Report (Cross-modal Retrieval)	Right lower lobe and lingular **pneumonia**.	Stable **cardiomegaly**. Otherwise, unremarkable.	Reoccurrence of left-sided **pleural effusion**. No other new abnormalities.	Mild **cardiomegaly** with central pulmonary vascular congestion without frank interstitial edema.	Worsening multifocal **pneumonia**.
REFERS – 1st Retrieved Prompt (Zero-shot Classification	Right lung bases **pneumonia**.	**Cardiomegaly** which is unchanged.	Stable left **pleural effusion**.	Moderate **cardiomegaly** with no acute chest abnormality.	Lung bases multifocal **pneumonia**.
M2SL – 1st Retrieved Report (Cross-modal Retrieval)	Right upper lobe and possibly right lower lobe **pneumonia**.	Moderate **cardiomegaly** with no acute chest abnormality.	Interval increase in moderate sized left **pleural effusion**.	Low lung volumes with bibasilar **atelectasis**.	Worsening multifocal **pneumonia**.
M2SL – 1st Retrieved Prompt (Zero-shot Classification	Lung bases **pneumonia**.	**Cardiomegaly** without acute cardiopulmonary process.	Increased left **pleural effusion**.	Lower lobe with mild bibasilar **atelectasis**.	Bilateral lung bases multifocal **pneumonia**.

Fig. 3. The performance visualization of the proposed method (M2SL) and selected state-of-the-art (SOTA) method (REFERS [23]) for cross-modal retrieval and zero-shot classification tasks on different classes from MIMIC-5 × 200 dataset. If the label from the retrieved text or prompt matches with the label from the ground truth report, it is considered as a correct (green) retrieval or classification; otherwise, it is considered as an incorrect (red) case. Graphs are mosaicked for confidential purposes.

Table 2
Ablation study on image (Img) and text (Ttx) encoder selection for cross-modal retrieval (mAP: mean average precision) and zero-shot classification (Acc(Std): Accuracy with Standard Deviation) tasks on the MIMIC-5 × 200, CheXpert-5 × 200, and RSNA datasets. R_{50}: ResNet-50, SwinT: Swin-Transformer, D_{121}: DenseNet-121.

Methods and tasks	Cross-modal retrieval			Zero-shot classification		
Encoder (Img+Txt)	MIMIC-5 × 200			MIMIC-5 × 200	CheXpert-5 × 200	RSNA
	mAP@1	mAP@5	mAP@10	Acc (Std)	Acc (Std)	Acc (Std)
R_{50}+CheXbert	0.540	0.610	0.594	0.496 (<0.01)	0.434 (<0.01)	0.811 (<0.01)
SwinT+CheXbert	0.530	0.593	0.583	0.499 (<0.01)	**0.446** (<0.01)	**0.844** (<0.01)
D_{121}+BioClinicalBert	0.526	0.602	0.583	0.502 (<0.01)	0.403 (<0.01)	0.815 (<0.01)
D_{121}+CheXbert	**0.556**	**0.619**	**0.608**	**0.511** (<0.01)	0.407 (<0.01)	0.831 (<0.01)

4.5. Modality alignment visualization

Retrieval results for the proposed method (M2SL) and selected SOTA method (REFERS [23]) of example X-ray images in the MIMIC-5 × 200 dataset are illustrated in Fig. 3 for both cross-modal retrieval and zero-shot classification tasks. For each query, the test image and its associated report are displayed alongside the retrieved report and prompt. In these cases incorrectly retrieved by the REFERS method yet correctly retrieved by the M2SL method, the proposed method captures the correct label and accurately identifies the severity and location of diseased areas. An example is shown in the fourth image from the left, where our proposed method could successfully identify mild atelectasis in the lower lungs, usually caused by a blockage of the air passages or pressure on the lung. For incorrect retrieval (e.g., the first image from the left), M2SL still identifies semantically relevant airspace shadowing, as pneumonia remains a plausible differential diagnosis based on imaging alone. Addtionally, M2SL overlooks pulmonary edema in the left lung. However, it is important to note that both pulmonary edema and pneumonia can lead to fluid accumulation in the lungs, making differentiation challenging. While a clinician interpreting the image would typically have access to complementary scans and clinical information not available to the model. Integrating clinical information and health records into textual reports could potentially enhance the M2SL model's ability to generate more precise diagnoses.

4.6. Ablation study

A series of experiments on ablation studies were conducted to verify the performance of the proposed M2SL method, which was explored in the following four aspects.

4.6.1. Ablation study on encoder selection

To thoroughly investigate the encoding capabilities of different image and text encoders, we conducted two comprehensive sets of experiments in Table 2. The experimental results revealed that while fixing the text encoders, the consistency of having better performance using our proposed M2SL method is observed across all chosen image encoders (1st and 2nd row). Interestingly, we also noticed that when being evaluated on the CheXpert-5 × 200 and RSNA datasets for the zero-shot classification task, the Swin-Transformer model (2nd row) tends to perform better for capturing more semantic meaningful representations when facing domain shifts (with performance close to the highest accuracy of 0.450 from GLoRIA [20] on the CheXpert-5 × 200 dataset), as it retains the efficiency processing image patches while utilizing Transformers to capture multi-scale features for enhancing both local and global feature learning. This finding underscores the importance of selecting an appropriate image encoder as a potential strategy for handling datasets where strong local image–text alignment plays a crucial role in classification. However, it is crucial to recognize that these advantages come with significant computational expenses, rendering the approach less cost-effective. Moreover, when switching the text encoders from CheXbert to BioClinicalBert while fixing the

Table 3
Ablation study on the effectiveness of proposed components: report-based semantic similarity ($\boldsymbol{S}_j$) and coupling network (g_c) with relational similarity ($\boldsymbol{R}$) for cross-modal retrieval (mAP: mean average precision) and zero-shot classification (Acc (Std): Accuracy with Standard Deviation) tasks on the MIMIC-5 × 200 dataset.

Methods and tasks		Cross-modal retrieval			Zero-shot classification
Report-based	Coupling network (g_c) and	MIMIC-5 × 200			MIMIC-5 × 200
Semantic similarity ($\boldsymbol{S}$)	Relational similarity ($\boldsymbol{R}$)	mAP@1	mAP@5	mAP@10	Acc (Std)
×	×	0.460	0.524	0.508	0.465 (<0.01)
✓	×	0.467	0.560	0.537	0.467 (<0.01)
×	✓	0.542	0.603	0.595	0.496 (0.015)
✓	✓	**0.556**	**0.619**	**0.608**	**0.511** (<0.01)

Table 4
Ablation study on coupling network (g_c) structure for cross-modal retrieval (mAP: mean average precision) and zero-shot classification (Acc(Std): Accuracy with Standard Deviation) tasks on the MIMIC-5 × 200 dataset. MLP: MultiLayer Perceptron.

Methods and tasks	Cross-modal retrieval			Zero-shot classification
Coupling network	MIMIC-5 × 200			MIMIC-5 × 200
(g_c)	mAP@1	mAP@5	mAP@10	Acc (Std)
3-Layer MLP	0.439	0.523	0.504	0.502 (<0.01)
4-Layer MLP	**0.556**	**0.619**	**0.608**	**0.511** (<0.01)
5-Layer MLP	0.549	0.593	0.587	0.509 (<0.01)

Table 5
Ablation study on feature vector dimensions for representation learning for cross-modal retrieval (mAP: mean average precision) and zero-shot classification (Acc(Std): Accuracy with Standard Deviation) tasks on the MIMIC-5 × 200 datasets.

Methods and tasks	Cross-modal retrieval			Zero-shot classification
Feature vector	MIMIC-5 × 200			MIMIC-5 × 200
Dimensions	mAP@1	mAP@5	mAP@10	Acc (Std)
200	0.531	0.608	0.597	0.497 (<0.01)
300	**0.556**	**0.619**	**0.608**	0.511 (<0.01)
400	0.536	0.601	0.587	**0.515** (<0.01)
1000	0.530	0.583	0.548	0.489 (<0.01)

image encoders, there was a slight drop in model performance, but still better than most SOTA models (3rd row). CheXbert was initially trained with annotations derived from a rule-based labeler and subsequently fine-tuned using a smaller dataset of expert annotations, which were further augmented through automated back-translation employing “uncased” text. In contrast, BioClinicalBERT was initialized with the weights of BioBERT and further trained on clinical notes from the MIMIC-III dataset, using “cased” text. Consequently, text encoders that are more specifically tailored to radiology reports demonstrate incrementally superior performance. This observation is also consistent with the findings in [69].

4.6.2. Effectiveness of proposed components

To fairly evaluate the benefits of various components for our proposed M2SL method, we separately investigated its impact in Table 3 for cross-modal retrieval and zero-shot classification tasks on MIMIC-5 × 200 dataset. When replacing the label-based prior with a knowledge-driven semantic similarity leveraging the rich information from the reports as the additional supervision (1st to 2nd row), there is a performance improvement of up to 0.036 for mAP for cross-modal retrieval task compared to the conventional method of constructing similarity using the labels, and such improvements are even more obvious with coupling network and relational similarity added (3rd to 4th row). These results are also consistent with our hypothesis that a knowledge-driven semantic similarity matrix could serve as an extra supervision signal to better guide the learning process by shedding light on alleviating false negatives. Additionally, adding the relational similarity computed from a coupling network with fused features (1st to 3rd row and 2nd to 4th row) could also contribute to the performance increment of the retrieval task by up to 0.089 for mAP, in comparison with directly using the obtained image/text features to compute the pairwise similarities. This approach efficiently bridges the gap between different modalities without the need to explicitly learn a shared feature space. Moreover, when both semantic similarity and relational similarity are used (1st to 4th row), there is a performance increment up to 0.100 for mAP. Similar performance improvements could also be observed for zero-shot classification tasks when evaluating the contributions of those two components. These findings further demonstrate that our M2SL method effectively computes pairwise similarity while addressing the challenge posed by the varying levels of information across different modalities.

4.6.3. Ablation study on coupling network structure

Ablation experiments were also conducted on the coupling network structure: number of MLP layers in Table 4. For the MLP-layer number, we set it to 3, 4 (default), and 5 on the MIMIC-5 × 200 dataset. According to the result, M2SL is optimal for MLP-layer number of 4 for both cross-modal retrieval and zero-shot classification tasks. Furthermore, the findings suggest that a more intricate coupling network does not necessarily ensure superior performance compared to a simpler model. We attribute this phenomenon to the characteristics of the dataset’s distribution. Specifically, when the data distribution is relatively uniform, expanding the coupling network’s complexity may not consistently enhance performance. This implies that the relationship between network complexity and performance improvement is contingent upon the inherent properties of the dataset.

4.6.4. Ablation study on feature vector dimensions

We further conducted ablation experiments on the derived modality-specific feature vector after representation learning networks with different dimension numbers from 200, 300 (default), 400, and up to 1000 in Table 5. The results indicate that the model demonstrates optimal performance when the feature vector dimensions are set to 300. It is noteworthy that the influence of feature vector dimensions on the model’s performance is relatively less compared to the significance of the coupling network structure. This observation is particularly pronounced in the context of cross-modal retrieval tasks. Like the projection module in contrastive learning, the representation learning network transforms high-dimensional embeddings into low-dimensional features. This dimensionality reduction not only facilitates the identification of patterns and anomalies but also enhances our overall understanding of the data’s underlying structure. To simulate scenarios involving higher-dimensional feature inputs for the coupling network, we extended our ablation study by increasing the feature vector dimensions to 1000. The results indicate even with these high-dimensional inputs, the model’s performance remains consistently comparable, demonstrating its robustness in handling increased feature complexity.

4.6.5. Ablation study on fusion mechanisms

We also investigated the effects of various fusion mechanisms for relational similarity calculation including common-used methods such as adding, multiplication, and concatenation in Table 6. The results indicate that element-wise addition or multiplication only partially

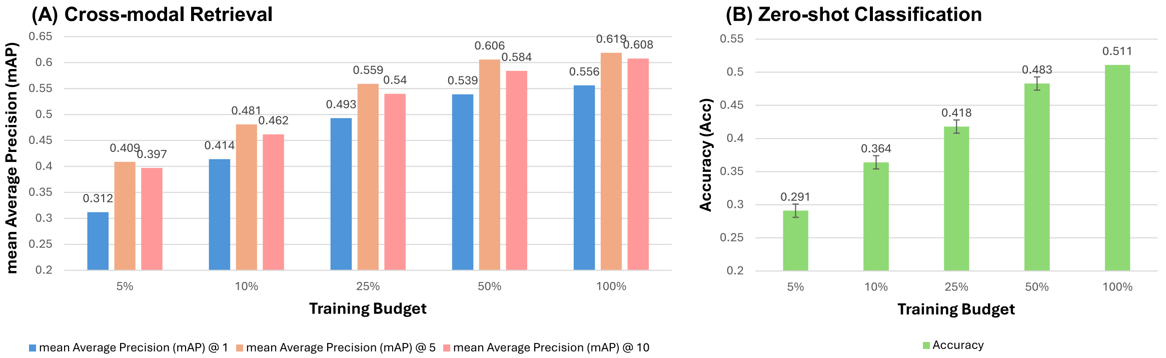


Fig. 4. The performance of the proposed Multimodal Multitask Similarity Learning (M2SL) method for (a) cross-modal retrieval and (b) zero-shot classification tasks on MIMIC-5 × 200 dataset using different amounts of training data (5%, 10%, 25%, 50%, 100%).

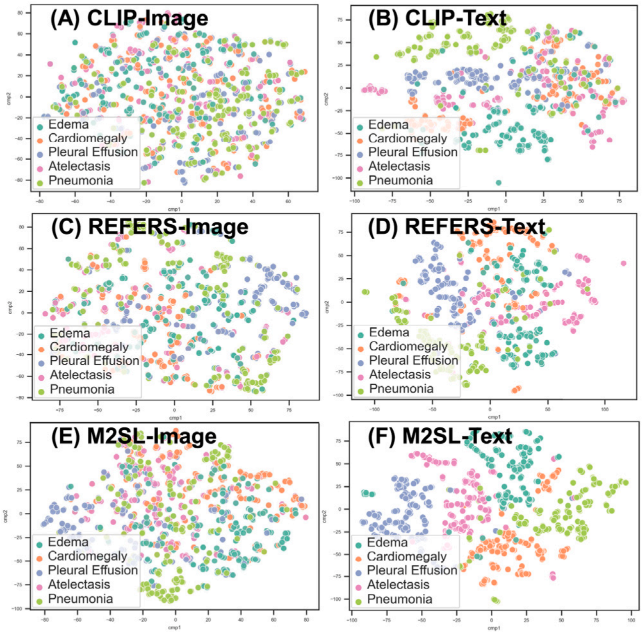


Fig. 5. Embeddings visualization of MIMIC-5 × 200 images and reports by (a)–(b) CLIP, (c)–(d) REFERS and (e)–(f) proposed (M2SL) method. Dimension reduced by t-SNE [70].

Table 6
Ablation study on fusion mechanisms for cross-modal retrieval (mAP: mean average precision) and zero-shot classification (Acc(Std): Accuracy with Standard Deviation) tasks on the MIMIC-5 × 200 datasets.

Methods and tasks	Cross-modal retrieval			Zero-shot classification
Fusion	MIMIC-5 × 200			MIMIC-5 × 200
Mechanisms	mAP@1	mAP@5	mAP@10	Acc (Std)
Adding	0.465	0.545	0.529	0.483 (<0.01)
Multiplication	0.526	0.590	0.578	0.476 (<0.01)
Concatenation	**0.556**	**0.619**	**0.608**	**0.511** (<0.01)

captures the interactions and correlations among multi-modal features, potentially hindering fusion performance. In contrast, concatenation introduces new positional dimensions to the processed features, integrating them as additional components without disrupting the original features, leading to a more coherent and comprehensive outcome.

4.7. Data efficiency

Notably, CLIP's utilization of 400M image–text pairs during training [30] reduces its versatility, particularly within medical applications, where such data abundance is often not commonly available. To address the challenge of low resources in the medical domain, we also investigated M2SL's performance when facing a limited training budget. We systematically subsampled the training data to 5%, 10% (15K), 25% (40K), and 50% (80K) of the original corpus for both feature representation learning and modalities alignment. The resulting model's performance across various tasks on MIMIC-5 × 200 data is depicted in Fig. 4. Surprisingly, even with a mere 25% (40K) data, M2SL exhibits only an 11.3% (mAP@1)/9.7% (mAP@5)/11.2% (mAP@10) drop in the performance of cross-modal retrieval, comparing to using 100% (160K) of the training budget (Fig. 4a). Similarly, when utilizing 50% (80K) data for zero-shot prediction, M2SL could still maintain a performance of 94.5% of the one using 100% (160K) of the training budget (Fig. 4b). While performance declines significantly when only 5% of the data is available for training. However, the primary objective of this evaluation is not only to demonstrate the effectiveness of our proposed method in mitigating the semantic gap between visual and textual modalities under low-resource settings (e.g., 25% and 50%). Furthermore, our results indicate that as more data is incorporated, our method continues to improve, underscoring its capacity to effectively learn from multimodal datasets.

4.8. Embedding visualization

We validated the efficacy of our method by visualizing t-SNE plots [70] for both image and text embeddings generated from MIMIC-5 × 200 images, through a comparative analysis with CLIP [30] and REFERS [23] model embeddings in Fig. 5. Our model demonstrates better clustering. Conversely, the t-SNE plot of the CLIP model appears homogeneous due to the substantial overlap presented in most medical X-rays, with only minor variations in diseased regions. Although the overlapping regions in the t-SNE plot of the REFERS model have been reduced, distinct cluster formations remain unclear. Notably, the different distribution patterns in the images and reports are also consistent with our hypothesis that there is a difference in information density from various modalities.

5. Conclusion

In this study, we introduce a multimodal multitask method for learning pairwise relational similarity and knowledge-driven semantic similarity through joint representation learning, utilizing both radiological images and reports. It allows the model to implicitly account for the unequal distribution of information across distinct modalities and equips the model with the ability to capture the subtle yet crucial medical knowledge. Extensive experiments conducted on various datasets showcase the efficacy of our method in enhancing both cross-modal retrieval and zero-shot classification performance. Our work also has significant implications for enhancing the VLM development for radiological applications, especially when multimodal image-report data contains uneven distributions of information and variations in quality. Despite the computational demands of multimodal training, our framework achieves competitive or superior performance in cross-modal alignment while maintaining efficiency, making it suitable for real-time applications. Future work could explore optimizations like model distillation to enhance efficiency. While M2SL has demonstrated strong performance in cross-modal retrieval and zero-shot prediction, several future works could further enhance practical applicability. To assess the effectiveness of its pre-training process, fine-tuning methods through linear probing could be further explored in tasks such as classification or segmentation. Moreover, medical datasets often exhibit imbalances in demographic attributes (e.g., age, sex, race) and disease prevalence. These biases can influence model performance and fairness, potentially leading to disparities in clinical decision support applications. Future work should incorporate bias assessment frameworks, such as evaluating performance across different demographic subgroups and disease severities, to ensure equitable model behavior. Techniques like dataset rebalancing, domain adaptation, and bias-aware training strategies could help mitigate these biases. Lastly, our current evaluation is limited to chest X-rays, which may not directly translate to other imaging modalities, such as CT scans, MRIs, or ultrasound. Differences in image characteristics, disease patterns, and reporting conventions pose challenges when applying VLMs across modalities. To enhance generalizability, future research should explore multi-modal training strategies, including pretraining on diverse imaging datasets and fine-tuning on modality-specific annotations. Additionally, zero-shot and few-shot adaptation techniques could be investigated to extend model applicability beyond chest X-rays.

CRediT authorship contribution statement

Yang Yu: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Jiahao Wang:** Writing – review & editing, Visualization, Validation, Methodology, Investigation, Formal analysis. **Weide Liu:** Writing – review & editing, Methodology, Investigation, Conceptualization. **Ivan Ho Mien:** Writing – review & editing, Validation, Investigation, Data curation. **Pavitra Krishnaswamy:** Writing – review & editing, Writing – original draft, Validation, Investigation, Formal analysis. **Xulei Yang:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization. **Jun Cheng:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research work is supported in part by the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021).

Data availability

All the selected datasets used in this work are publicly available. However, the authors do not have permission to share code due to the institute's policy.

References

[1] Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M Summers, Roland Wiest, On the interpretability of artificial intelligence in radiology: challenges and opportunities, Radiol.: Artif. Intell. 2 (3) (2020) e190043.

[2] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, Pranav Rajpurkar, Foundation models for generalist medical artificial intelligence, Nature 616 (7956) (2023) 259–265.

[3] Zihan Li, Yunxiang Li, Qingde Li, Puyang Wang, Dazhou Guo, Le Lu, Dakai Jin, You Zhang, Qingqi Hong, Lvit: language meets vision transformer in medical image segmentation, IEEE Trans. Med. Imaging (2023).

[4] Yi Zhong, Mengqiu Xu, Kongming Liang, Kaixin Chen, Ming Wu, Ariadne's thread: Using text prompts to improve segmentation of infected areas from chest X-ray images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 724–733.

[5] Go-Eun Lee, Seon Ho Kim, Jungchan Cho, Sang Tae Choi, Sang-Il Choi, Text-guided cross-position attention for segmentation: Case of medical image, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 537–546.

[6] Jaeyoung Huh, Sangjoon Park, Jeong Eun Lee, Jong Chul Ye, Improving medical speech-to-text accuracy using vision-language pre-training models, IEEE J. Biomed. Heal. Inform. (2023).

[7] Wenjing Zhang, Quange Tan, Pengxin Li, Qi Zhang, Rong Wang, Cross-modal transformer with language query for referring image segmentation, Neurocomputing 536 (2023) 191–205.

[8] Gianluca Moro, Stefano Salvatori, Giacomo Frisoni, Efficient text-image semantic search: A multi-modal vision-language approach for fashion retrieval, Neurocomputing 538 (2023) 126196.

[9] Yi Zhang, Ce Zhang, Yushun Tang, Zhihai He, Cross-modal concept learning and inference for vision-language models, Neurocomputing 583 (2024) 127530.

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

[11] Hong-Yu Zhou, Chenyu Lian, Liansheng Wang, Yizhou Yu, Advancing radiograph representation learning with masked record modeling, in: The Eleventh International Conference on Learning Representations, 2022.

[12] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al., Making the most of text semantics to improve biomedical vision–language processing, in: European Conference on Computer Vision, Springer, 2022, pp. 1–21.

[13] Cheng Chen, Aoxiao Zhong, Dufan Wu, Jie Luo, Quanzheng Li, Contrastive masked image-text modeling for medical visual representation learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 493–503.

[14] Ke Zhang, Yan Yang, Jun Yu, Hanliang Jiang, Jianping Fan, Qingming Huang, Weidong Han, Multi-task paired masking with alignment modeling for medical vision-language pre-training, IEEE Trans. Multimed. (2023).

[15] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, Pranav Rajpurkar, Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning, Nat. Biomed. Eng. 6 (12) (2022) 1399–1406.

[16] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al., A visual-language foundation model for computational pathology, Nature Med. 30 (3) (2024) 863–874.

[17] Matthew Christensen, Milos Vukadinovic, Neal Yuan, David Ouyang, Vision–language foundation model for echocardiogram interpretation, Nature Med. (2024) 1–8.

[18] Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, Edward Choi, Multi-modal understanding and generation for medical images and text via vision-language pre-training, IEEE J. Biomed. Heal. Inform. 26 (12) (2022) 6070–6080.

[19] Xiaosong Wang, Ziyue Xu, Leo Tam, Dong Yang, Daguang Xu, Self-supervised image-text pre-training with mixed data in chest x-rays, 2021, arXiv preprint arXiv:2103.16022.

[20] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, 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, 2021, pp. 3942–3951.

[21] Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K Hong, Woonhyuk Baek, Byungseok Roh, Cxr-clip: Toward large scale chest x-ray language-image pre-training, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 101–111.

[22] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, Junjie Yan, Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm, in: International Conference on Learning Representations, 2021.

[23] Hong-Yu Zhou, Xiaoyu Chen, Yinghao Zhang, Ruibang Luo, Liansheng Wang, Yizhou Yu, Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports, Nat. Mach. Intell. 4 (1) (2022) 32–40.

[24] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, Lequan Yu, Multi-granularity cross-modal alignment for generalized medical visual representation learning, Adv. Neural Inf. Process. Syst. 35 (2022) 33536–33549.

[25] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, Weidi Xie, Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 21372–21383.

[26] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, Jimeng Sun, MedCLIP: Contrastive learning from unpaired medical images and text, in: 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, 2022.

[27] Bo Liu, Donghuan Lu, Dong Wei, Xian Wu, Yan Wang, Yu Zhang, Yefeng Zheng, Improving medical vision-language contrastive pretraining with semantics-aware triage, IEEE Trans. Med. Imaging (2023).

[28] Chong Ma, Hanqi Jiang, Wenting Chen, Zihao Wu, Xiaowei Yu, Fang Zeng, Lei Guo, Dajiang Zhu, Tuo Zhang, Dinggang Shen, et al., Eye-gaze guided multi-modal alignment framework for radiology, 2024, arXiv preprint arXiv:2403.12416.

[29] Haoran Lai, Qingsong Yao, Zihang Jiang, Rongsheng Wang, Zhiyang He, Xiaodong Tao, S Kevin Zhou, Carzero: Cross-attention alignment for radiology zero-shot classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 11137–11146.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[31] Bo Liu, Li-Ming Zhan, Li Xu, Xiao-Ming Wu, Medical visual question answering via conditional reasoning and contrastive learning, IEEE Trans. Med. Imaging 42 (5) (2022) 1532–1545.

[32] Yuxin Peng, Jinwei Qi, Yuxin Yuan, Modality-specific cross-modal similarity measurement with recurrent attention network, IEEE Trans. Image Process. 27 (11) (2018) 5585–5599.

[33] Xu Wang, Peng Hu, Liangli Zhen, Dezhong Peng, Drsl: Deep relational similarity learning for cross-modal retrieval, Inform. Sci. 546 (2021) 298–311.

[34] Peng Hu, Liangli Zhen, Dezhong Peng, Pei Liu, Scalable deep multimodal learning for cross-modal retrieval, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 635–644.

[35] Liangli Zhen, Peng Hu, Xu Wang, Dezhong Peng, Deep supervised cross-modal retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10394–10403.

[36] Liangli Zhen, Peng Hu, Xi Peng, Rick Siow Mong Goh, Joey Tianyi Zhou, Deep multimodal transfer learning for cross-modal retrieval, IEEE Trans. Neural Netw. Learn. Syst. 33 (2) (2020) 798–810.

[37] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, Timothy M Hospedales, Learning to compare: Relation network for few-shot learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1199–1208.

[38] Binyuan Hui, Pengfei Zhu, Qinghua Hu, Qilong Wang, Self-attention relation network for few-shot learning, in: 2019 IEEE International Conference on Multimedia & Expo Workshops, ICMEW, IEEE, 2019, pp. 198–203.

[39] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, Rongrong Ji, Chunhua Shen, Pyramidclip: Hierarchical feature alignment for vision-language model pretraining, Adv. Neural Inf. Process. Syst. 35 (2022) 35959–35970.

[40] Yuting Gao, Jinfeng Liu, Zihan Xu, Tong Wu, Enwei Zhang, Ke Li, Jie Yang, Wei Liu, Xing Sun, Softclip: Softer cross-modal alignment makes clip stronger, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, 2024, pp. 1860–1868, 3.

[41] Liu Yiyang, Liang James Chenhao, Tang Ruixiang, Lee Yugyung, RABBANI MAJID, Dianat Sohail, Rao Raghuveer, Huang Lifu, Liu Dongfang, Wang Qifan, Han Cheng, Re-imagining multimodal instruction tuning: A representation view, in: Proceedings of the International Conference on Learning Representations, ICLR, 2025.

[42] Nicolas Deperrois, Hidetoshi Matsuo, Samuel Ruipérez-Campillo, Moritz Vandenhirtz, Sonia Laguna, Alain Ryser, Koji Fujimoto, Mizuho Nishio, Thomas M Sutter, Julia E Vogt, et al., RadVLM: A multitask conversational vision-language model for radiology, 2025, arXiv preprint arXiv:2502.03333.

[43] Difei Gu, Yunhe Gao, Yang Zhou, Mu Zhou, Dimitris Metaxas, RadAlign: Advancing radiology report generation with vision-language concept alignment, 2025, arXiv preprint arXiv:2501.07525.

[44] Gijs van Tulder, Marleen de Bruijne, Learning cross-modality representations from multi-modal images, IEEE Trans. Med. Imaging 38 (2) (2018) 638–648.

[45] Taowen Wang, Yiyang Liu, James Liang, Junhan Zhao, Yiming Cui, Yuning Mao, Shaoliang Nie, Jiahao Liu, Fuli Feng, Zenglin Xu, et al., M2PT: Multimodal prompt tuning for zero-shot instruction learning, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 3723–3740.

[46] Cheng Han, Qifan Wang, Yiming Cui, Zhiwen Cao, Wenguan Wang, Siyuan Qi, Dongfang Liu, E 2 VPT: An effective and efficient approach for visual prompt tuning, in: 2023 IEEE/CVF International Conference on Computer Vision, ICCV, IEEE, 2023, pp. 17445–17456.

[47] Runjia Zeng, Cheng Han, Qifan Wang, Chunshu Wu, Tong Geng, Lifu Huang, Ying Nian Wu, Dongfang Liu, Visual Fourier prompt tuning, in: The Thirty-Eighth Annual Conference on Neural Information Processing Systems, 2024.

[48] Shaohua Dong, Yunhe Feng, Qing Yang, Yan Huang, Dongfang Liu, Heng Fan, Efficient multimodal semantic segmentation via dual-prompt learning, in: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, IEEE, 2024, pp. 14196–14203.

[49] Cheng Han, Qifan Wang, Yiming Cui, Wenguan Wang, Lifu Huang, Siyuan Qi, Dongfang Liu, Facing the elephant in the room: Visual prompt tuning or full fine-tuning? in: The Twelfth International Conference on Learning Representations, 2024.

[50] Guangyan Sun, Mingyu Jin, Zhenting Wang, Cheng-Long Wang, Siqi Ma, Qifan Wang, Ying Nian Wu, Yongfeng Zhang, Dongfang Liu, Visual agents as fast and slow thinkers, 2024, CoRR.

[51] Abdol Hamid Pilevar, CBMIR: Content-based image retrieval algorithm for medical image databases, J. Med. Signals Sens. 1 (1) (2011) 12–18.

[52] Adnan Qayyum, Syed Muhammad Anwar, Muhammad Awais, Muhammad Majid, Medical image retrieval using deep convolutional neural network, Neurocomputing 266 (2017) 8–20.

[53] Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y Ng, Pranav Rajpurkar, Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model, in: Machine Learning for Health, PMLR, 2021, pp. 209–219.

[54] Yang Yu, Peng Hu, Jie Lin, Pavitra Krishnaswamy, Multimodal multitask deep learning for X-ray image retrieval, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24, Springer, 2021, pp. 603–613.

[55] Yong Zhang, Weihua Ou, Jiacheng Zhang, Jiaxin Deng, Category supervised cross-modal hashing retrieval for chest x-ray and radiology reports, Comput. Electr. Eng. 98 (2022) 107673.

[56] Tom van Sonsbeek, Marcel Worring, X-tra: Improving chest x-ray tasks with cross-modal retrieval augmentation, in: International Conference on Information Processing in Medical Imaging, Springer, 2023, pp. 471–482.

[57] Angshuman Paul, Thomas C Shen, Sungwon Lee, Niranjan Balachandar, Yifan Peng, Zhiyong Lu, Ronald M Summers, Generalized zero-shot chest x-ray diagnosis through trait-guided multi-view semantic embedding with self-training, IEEE Trans. Med. Imaging 40 (10) (2021) 2642–2655.

[58] Dwarikanath Mahapatra, Behzad Bozorgtabar, Zongyuan Ge, Medical image classification using generalized zero shot learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3344–3353.

[59] Devraj Mandal, Sanath Narayan, Sai Kumar Dwivedi, Vikram Gupta, Shuaib Ahmed, Fahad Shahbaz Khan, Ling Shao, Out-of-distribution detection for generalized zero-shot action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9985–9993.

[60] Akanksha Paul, Narayanan C. Krishnan, Prateek Munjal, Semantically aligned bias reducing zero shot learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7056–7065.

[61] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, Steven Horng, MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports, Sci. Data 6 (1) (2019) 317.

[62] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al., Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 590–597, 01.

[63] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al., Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia, Radiol.: Artif. Intell. 1 (1) (2019) e180041.

[64] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, Kilian Q Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[65] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, Matthew Lungren, Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2020, pp. 1500–1519.

[66] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[67] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

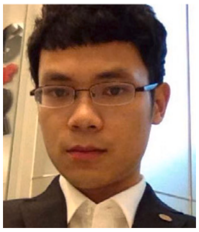
[68] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, Matthew McDermott, Publicly available clinical BERT embeddings, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019, pp. 72–78.

[69] Hyun Gi Lee, Evan Sholle, Ashley Beecy, Subhi Al'Aref, Yifan Peng, Leveraging deep representations of radiology reports in survival analysis for predicting heart failure patient mortality, in: Proceedings of the Conference. Association for Computational Linguistics. North American Chapter. Meeting, vol. 2021, NIH Public Access, 2021, p. 4533.

[70] Laurens Van der Maaten, Geoffrey Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008).

Dr. Yang Yu is a senior researcher at the Institute for Infocomm Research, A*STAR, Singapore, with a focus on AI-aided diagnostics and analytics for healthcare data. He holds a Ph.D. from the National University of Singapore and a bachelor's degree from Nanyang Technological University. Yu's research aims to advance artificial intelligence applications in healthcare and data interpretation. Yu has a strong background in deep learning and computer vision and is committed to developing innovative solutions that enhance data analysis and diagnostic processes.

Dr. Jiahao Wang is a researcher specializing in medical image analysis, few-shot learning, and dataefficient training. He holds a Ph.D. from the National University of Singapore and an a bachelor's degree from Zhejiang University. Currently as an AI Researcher, he leads the video multimodal foundation model team, focusing on long and short video understanding, token compression, and long-context training. Jiahao has co-authored multiple academic papers and has contributed to AI applications in healthcare.



Dr. Weide Liu is currently a Research Fellow at Boston Children's Hospital, Harvard Medical School. Before that, he was a Research Scientist at A*STAR in Singapore. Weide received his Ph.D. and bachelor's degrees from Nanyang Technological University. His research interests include computer vision, language, machine learning, and medical image analysis.

Dr. Ivan Ho Mien is a principal scientist at the Institute for Infocomm Research, A*STAR, Singapore . He also holds joint appointments as a consultant neuroradiologist at the National Neuroscience Institute and an adjunct assistant professor with the Duke–NUS Medical School. He obtained his Beng (Hons), MBBS, and Ph.D. from the National University of Singapore and is a fellow of the Royal College of Radiologists (UK). His research focuses on applications of artificial intelligence and machine learning for clinical imaging and decision support.

Dr. Pavitra Krishnaswamy received her Ph.D. in Electrical and Medical Engineering from the Massachusetts Institute of Technology and Harvard Medical School, USA. She is currently a Principal Scientist and Deputy Division Head at the Institute for Infocomm Research, A*STAR, Singapore. Her research interests include statistical learning and inference, representation learning, and multimodal learning leveraging real-world data for a range of healthcare applications. She is on the Singapore 100 Women in Technology List and is a Fellow of the American Medical Informatics Association.

Dr. Xulei Yang received his Ph.D. degree from Nanyang Technological University in 2007. He is currently a principal scientist and group leader at Institute for Infocomm Research, A*STAR, Singapore, previously the research head at YITU Technology Singapore, with more than 16 years of R&D experience in deep/machine learning for computer vision and healthcare. He has published more than 100 scientific papers and international patents in the fields of deep learning, 3D Vision and medical imaging. He is currently an IEEE Senior Member and Kaggle Competition Master.

Dr. Jun Cheng received the B.E. degree from the University of Science and Technology of China, and the Ph.D. degree from Nanyang Technological University, Singapore. He is now a principal research scientist in the Institute for Infocomm Research, A*STAR, Singapore, working on AI for medical imaging, robust machine vision and perception. He is an Associate Editor for IEEE Transactions on Image Processing and IEEE Transactions on Medical Imaging.