npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01339-7

A multimodal multidomain multilingual medical foundation model for zero shot clinical diagnosis

Fenglin Liu[1] ✉, Zheng Li[2], Qingyu Yin[2], Jinfa Huang[3], Jiebo Luo[3], Anshul Thakur[1], Kim Branson[4], Patrick Schwab[4], Bing Yin[2], Xian Wu[5] ✉, Yefeng Zheng[6] ✉ & David A. Clifton[1,7] ✉

Radiology images are one of the most commonly used in daily clinical diagnosis. Typically, clinical diagnosis using radiology images involves disease reporting and classification, where the former is a multimodal task whereby textual reports are generated to describe clinical findings in images, as are common in various domains, e.g., chest X-ray or computed tomography. Existing approaches are mainly supervised, the quality of which heavily depends on the volume and quality of available labeled data. However, for rarer or more novel diseases, enrolling patients to collect data is both time-consuming and expensive. For non-English languages, sufficient quantities of labeled data are typically not available. We propose the Multimodal Multidomain Multilingual Foundation Model. It is useful for rare diseases and non-English languages, where the labeled data are frequently much more scarce, and may even be absent. Our approach achieves encouraging performances on nine datasets, including 2 infectious and 14 non-infectious diseases.

Multimodal clinical diagnosis involves disease reporting and classification, aiming to understand clinical information from input medical images (e.g., from radiology) to (1) generate a coherent report describing clinical findings derived from the images, and (2) further diagnose the diseases present in the images[1–4]. Writing medical reports for different imaging domains, e.g., chest X-ray (CXR) and computed tomography (CT), is time-consuming and often described as being routine and tedious for clinicians[5,6], thereby increasing their workload and being associated with a corresponding human error rate[7–10]. Introducing a report generation system, which can automatically generate first-draft reports that clinicians can then review, modify, and approve, can partly simplify routine tasks and thus constitutes a valuable contribution to clinical practice by (partly) automating routine tasks to drive down error rates[11–13].

Recently, encoder-decoder-based deep learning methods have been widely used to advance the state-of-the-art[3,14,15]. These existing systems usually include a vision encoder to extract the representations of the input medical image and a language decoder to generate the target reports. However, such systems cannot typically deal with multidomain medical images and multilingual languages within a single framework and heavily rely on a large volume of medical training data annotated by clinicians, which is both expensive and time-consuming to collect. Therefore, existing methods have the following problems: (1) For rare diseases and novel diseases, including new disease variants, the availability of sufficient labeled data for training is typically unavailable at the early stages—arguably when tools are most urgently required. Taking COVID-19 as an example, the time to collect sufficient data to train systems was found to be longer than the duration of the first few waves of the pandemic (i.e., 6–12 months)[16–18]. By then, arguably, the worst of the pandemic had passed, including the immense early pressures that were placed on hospitals. (2) Additionally, for languages other than English, the labeled data are frequently much more scarce, and may even be unavailable[19]. Hence, the limited availability of labeled data presents a major challenge in training systems for non-English languages using existing methods. This challenge becomes even more demanding when dealing with even less commonly encountered languages, particularly those spoken by marginalized communities, making it harder to achieve "Fair AI" that can benefit underrepresented communities[20]. Meanwhile, many global healthcare systems involve different languages. We conclude that these problems often prevent the use of existing deep learning systems for analyzing medical data pertaining to new diseases or non-English languages.

To this end, we propose the Multimodal Multidomain Multilingual Foundation Model (M^3FM), which is pre-trained on public medical data

[1]Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK. [2]Amazon, Palo Alto, CA, USA. [3]Department of Computer Science, University of Rochester, Rochester, NY, USA. [4]GlaxoSmithKline, London, UK. [5]Jarvis Research Center, Tencent YouTu Lab, Beijing, China. [6]Medical Artificial Intelligence Laboratory, Westlake University, Hangzhou, China. [7]Oxford-Suzhou Centre for Advanced Research, Suzhou, China.
✉e-mail: fenglin.liu@eng.ox.ac.uk; kevinxwu@tencent.com; zhengyefeng@westlake.edu.cn; david.clifton@eng.ox.ac.uk

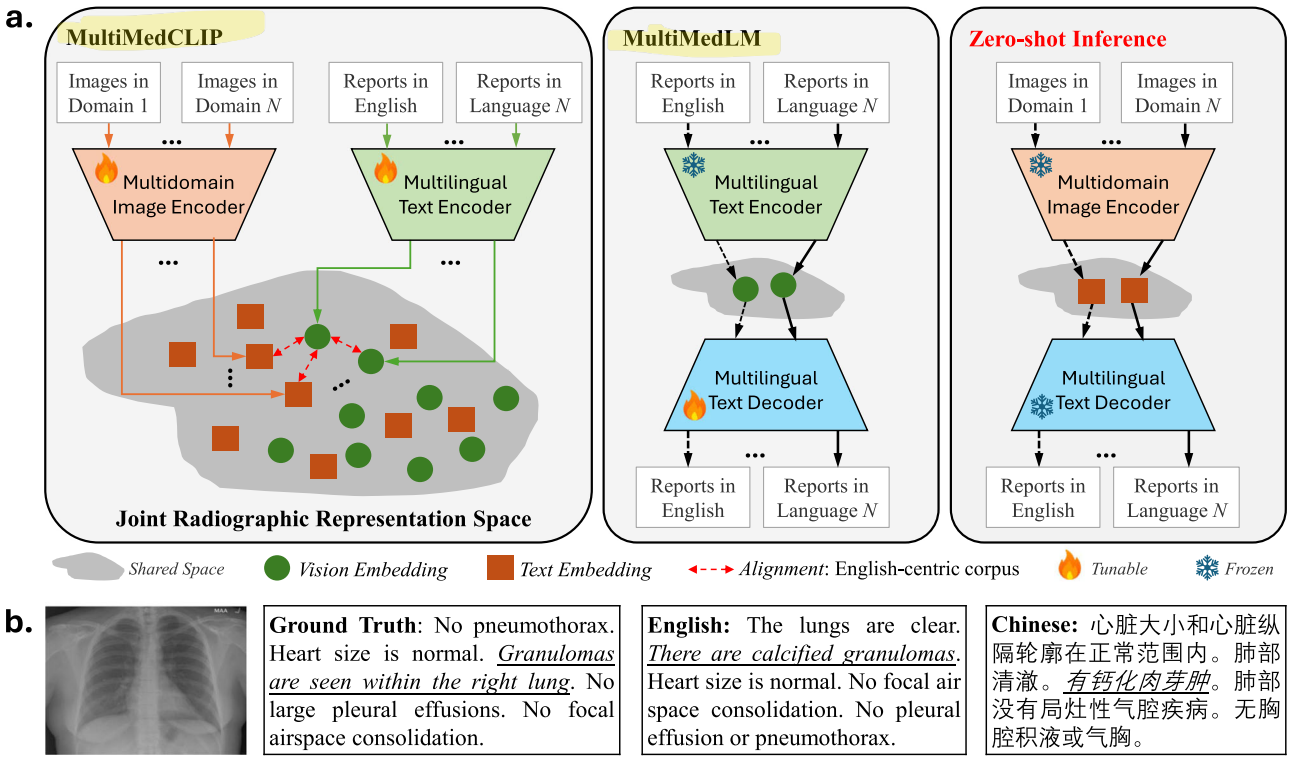

Fig. 1 | The M³FM and sample reports. a Illustration of the M3FM, comprising MultiMedCLIP and MultiMedLM. We first adopt the available English-centric corpora to pre-train the MultiMedCLIP module, which aligns visual and textual representations across domains and languages in a shared latent space. Then, we adopt the text data to pre-train the MultiMedLM by reconstructing input reports across languages based on the textual representations. After training, we can directly adopt the aligned visual representations in the shared latent space as input to perform zero-shot clinical diagnosis. **b** The examples of reports generated by our method for different languages, i.e., English and Chinese. The Underlined text denotes the generated correct important abnormalities.

across modalities, domains, and languages, to learn board knowledge. Our M³FM consists of two major modules: MultiMedCLIP and MultiMedLM, to perform zero-shot "multimodal multidomain multilingual" disease reporting and classification. In this paper, we use CXR and CT images with both English and Chinese reports as examples to illustrate our method. As shown in Fig. 1(a), (1) MultiMedCLIP creates a shared common latent space using the representations (a.k.a. embeddings) of public English-centric corpora, e.g., CXR-English pairs, CT-English pairs, Chinese-English pairs (We note that the English-centric corpora used for training are independent and have no overlap, i.e., the English text in CXR-English, the English text in CT-English, and the English text in Chinese-English are separate sets of text with no overlap. There are no labeled CXR-Chinese and CT-Chinese sets available). In this space, different image domains and non-English languages can be aligned to the English domain by minimizing the distance between their embeddings and English embeddings in the shared common latent space. As a result, our method can align different visual and textual representations across domains and languages *without training* on labeled data pairs, e.g., CXR-Chinese, and CT-Chinese. We note in passing that the use of existing resources, such as public English-centric corpora, can bring cost and energy savings, which is a critical step towards sustainable AI[21]. (2) Next, we adopt the text data to train a multilingual medical language model, i.e., MultiMedLM, by reconstructing the input text across languages, which enables the model to learn to understand and generate the medical text based on the textual representations. Such practice could provide opportunities to leverage any publicly available unlabeled text-only data, e.g., medical textbooks, to further improve performance. (3) At last, due to the visual and textual representations being aligned via MultiMedCLIP, we freeze the model parameters and can directly adopt the aligned visual representation of medical images from multiple domains as input to perform zero-shot inference for different languages, without training on any downstream labels. Our method could be easily extended to other languages/images by directly aligning the English and target languages/images.

We describe extensive experiments on nine public benchmark datasets across 2 infectious and 14 non-infectious diseases (i.e., tuberculosis and COVID-19); these show that our approach significantly outperforms previous methods and particularly deals with CXR report generation, CT report generation, and disease diagnosis across English and Chinese within a single framework for the first time. Overall, our method could be useful for rare diseases, new diseases (or variants of disease), and non-English languages, for which labeled data for training could be absent, and where our approach is promising.

Overall, the contributions of our work are as follows: (1) We propose an effective M³FM approach to make the first attempt to conduct zero-shot multimodal multidomain multilingual clinical diagnosis where the labeled data for training are scarce or even absent. (2) M³FM introduces two proposed modules, i.e., MultiMedCLIP and MultiMedLM, where the former can learn the joint radiographic representations of multidomain images and multilingual languages, and the latter can efficiently and accurately generate multilingual medical reports given the textual or visual representations extracted by MultiMedCLIP. (3) We verify the effectiveness of our approach on nine datasets, including (i) two domains of medical imaging data, i.e., CXR and CT; (ii) two different languages, i.e., English and Chinese; (iii) two kinds of clinical diagnosis tasks, i.e., disease reporting and disease classification; (iv) diverse diseases, including 14 non-infectious and 2 infectious diseases, i.e., tuberculosis and COVID-19; and (5) three experimental settings, i.e., zero-shot, few-shot, and fully supervised learning.

Results

In this section, we will introduce the datasets, metrics, and settings used for experiments; and then we illustrate the results of our proposed method.

Datasets, metrics, and settings

We pre-train our method on English-centric corpora, i.e., CXR-English pairs, CT-English pairs, and Chinese-English pairs. It is worth noting that these three English corpora are independent and can have no overlap. We evaluate our approach on four disease reporting tasks across domains and languages, i.e., CXR-to-English, CXR-to-Chinese, CT-to-English, and CT-to-Chinese report generation tasks. As there are no human-annotated datasets available for the CXR-to-Chinese task, we report the qualitative results in Fig. 1b. For CXR-to-English and CT-to-English tasks, as the pre-training data (the data to pre-train M^{3}FM) and evaluation data are from different institutions and we also do not use any labeled downstream data for training, we can still consider these two evaluations as zero-shot report generation. For the disease classification task, we follow previous works[2] to directly combine the radiology images and the generated reports to make predictions.

We adopt the MIMC-CXR[22] and COVID-19-CT-CXR[23] datasets for pre-training.

1) *MIMC-CXR*[22] is the recently released largest dataset to date and consists of 377,110 CXR images associated with 227,835 English radiographic reports.

2) *COVID-19-CT-CXR*[23] is a public dataset including 1k CT/CXR images and corresponding English reports.

In our work, for non-English corpora, considering that only Chinese reports are publicly available, we use Chinese as an example of a non-English language to evaluate our approach. To this end, we extract half of the English corpora from each of the MIMIC-CXR and COVID-19-CT-CXR datasets, and then construct Chinese-English training pairs using Google Translator[24]. Our experiments show that our approach can achieve improved results with machine-translated text. Nevertheless, our model is not limited to the presently utilized pre-trained dataset and has the potential for significant improvement through the utilization of human-annotated high-quality datasets.

The downstream tasks include disease reporting and disease diagnosis.

1. *IU-Xray*[25] is a widely used benchmark dataset containing 7470 CXR images associated with 3955 English radiographic reports. For a fair comparison, we follow common practice[14,26–28] to pre-process the dataset, which is randomly split into 80%-10%-10% training-validation-testing sets.
2. *COVID-19 CT*[29] consists of 1104 CT images associated with 368 radiographic reports in Chinese. For a fair comparison, we randomly split 80% of the samples for training, 10% for validation, and the remaining 10% for testing.
3. *COV-CTR*[30] is constructed by The First Affiliated Hospital of Harbin Medical University and comprises a total of 726 COVID-19 CT images (335 for COVID-19 and 391 for Non-COVID) associated with English and Chinese reports.
4. *Shenzhen Tuberculosis*[31] contains a total of 662 CXR images with 336 abnormal tuberculosis images and 326 normal images, used for the binary classification task. We follow[32] to split the dataset into training, validation and test sets with a ratio of 7:1:2, respectively.
5. *COVID-CXR*[33,34] is composed of over 900 CXR images used for the COVID-19 diagnosis task, which requires the models to distinguish COVID-19 from non-COVID-19 cases. The dataset is split into 80%-10%-10% training-validation-test sets.
6. *NIH ChestX-ray*[35] contains 112,120 CXR images, with each image labeled with occurrences of 14 common radiographic diseases. We follow previous works[32,36,37] to split the dataset into 70%, 10%, and 20% for training, validation, and test, respectively.
7. *CheXpert*[38] contains over 220,000 CXR images used for disease diagnosis (i.e., multi-label disease classification) tasks. We follow common practice[37,39,40] to pre-process the dataset, resulting in 218,414 images in the training set, 5000 images in the validation set, and 234 images in the test set.
8. *RSNA Pneumonia*[41] consists of around 30k radiographic images. We adopt the official split, acquiring 85%-5%-10% training-validation-test sets.
9. *SIIM-ACR Pneumothorax*[42] includes 12,047 CXR images. We follow[37,39] to split the dataset into 70%/15%/15% for train/valid/test.

For the disease reporting task, we adopt the standard evaluation toolkit[43] to report the widely used language generation metrics, i.e., BLEU[44], ROUGE-L[45], METEOR[46], and CIDEr[47], to report our results. These language generation metrics measure the match between the generated reports and ground truth reports and can be used for both English and Chinese texts. All values are reported in percentage (%). For the disease diagnosis task, to calculate the performance of disease diagnosis, we employ the widely used area under the curve (AUC) of the receiver operating characteristic.

Results of disease reporting

In Table 1, we report the performances of disease reporting across languages, domains, and modalities under three experimental settings. For comparison, we re-implement four state-of-the-art methods[6,15,28,48]. As we can see, our method can achieve the best results under different settings.

Under the zero-shot setting, all previous works can not deal with disease reporting. In contrast, our M^{3}FM is able to simultaneously perform multilingual multidomain disease reporting in a single unified framework. More encouragingly, our zero-shot M^{3}FM outperforms previous few-shot learning methods, e.g., R2Gen on CXR-to-English and all baselines on CT-to-Chinese. Besides, without any downstream data for training, on COVID-19 clinical diagnosis, our method achieves competitive results w.r.t. the existing methods trained on the full training set across English and Chinese. It shows that our method could provide a solid basis for disease reporting, which can be easily improved by further using a small amount of downstream labeled data for training. It is validated in the following few-shot learning setting.

From Table 1, we can find that our method with 10% downstream labeled data for training achieves the best results, and even outperforms the previous fully supervised method R2Gen by up to 1.5% CIDEr and 1.2% ROUGE-L scores in CT-to-Chinese report generation. It shows that our method M^{3}FM can be efficiently trained with limited labels to generate accurate and robust multilingual reports. In contrast, existing efforts usually require more than a thousand, or even more, labeled data. Such advantages of our method are particularly useful for rare and new diseases, where the labeled data are scarce.

To further prove the effectiveness of our method, we propose to continuously fine-tune our method using the full training set as in existing works. It means that we make the model tunable to learn better representations. Besides, it will allow us to simulate whether our model can be substantially improved by incorporating more labeled data as the disease evolves during new diseases. Table 1 shows that our method significantly surpasses previous methods and achieves encouraging performances across most metrics.

At last, we randomly select 50 samples from the test sets and invite two clinicians to compare our approach and baselines independently. They are unaware of which model generates these reports. In detail, we require them to evaluate how often the generated reports are "helpful" versus "unhelpful" in assisting the clinicians, that is, the reports describe the clinical findings accurately and the clinicians only need to slightly modify the machine-generated reports, rather than writing new reports from scratch. The left part of Table 2 shows that (1) without any labeled data for training, our method can generate desirable multilingual and multidomain reports; (2) with only 10% labeled data for training, our method M^{3}FM outperforms fully supervised method R2Gen (trained on 100% labeled data), by 6%, 8%, and 8% points on the CXR-to-English, CT-to-Chinese, and CT-to-English tasks, respectively; 3) with full training data, our method generates more 'helpful' results than previous works. These results demonstrate the effectiveness of our method in relieving clinicians from the time-consuming and laborious report writing.

Results of disease classification

In this section, we further apply the proposed M^{3}FM to disease classification (i.e., disease diagnosis) tasks[36,37,39]. The results of the diagnosis performance AUC score are reported in Tables 3 and 4.

Table 1 | Performances of disease reporting

Settings	Methods	Year	Ratio of data	CXR-to-English (*IU-Xray*[25])					CT-to-Chinese (*COVID-19 CT*[28])				CT-to-English (*COV-CTR*[30])		
				BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	BLEU-2	BLEU-3	BLEU-4	ROUGE	CIDEr	BLEU-4	ROUGE	CIDEr
Fully supervised	R2Gen[28]	2020	100%	30.4	21.9	16.5	18.7	37.1	53.3	45.1	39.4	54.5	80.4	53.2	73.5	66.4
	CMN[32]	2021	100%	33.4	24.2	16.8	20.3	38.7	56.9	49.2	44.7	61.0	85.1	56.2	75.6	69.0
	XProNet[15]	2022	100%	35.7	26.2	19.9	22.0	41.1	57.7	49.0	44.4	59.4	84.5	59.5	76.1	69.7
	PMRG[48]	2024	100%	36.3	27.8	19.7	22.4	41.5	59.3	51.5	45.0	62.9	87.2	61.3	77.4	72.6
	M^{3}FM	Ours	100%	$36.1_{(0.4)}$	$27.5_{(0.3)}$	$20.3_{(0.3)}$	$24.2_{(0.2)}$	$42.7_{(0.5)}$	$61.9_{(0.6)}$	$52.6_{(0.4)}$	$46.3_{(0.3)}$	$64.0_{(0.5)}$	$89.3_{(0.8)}$	$64.8_{(0.4)}$	$79.6_{(0.6)}$	$75.1_{(1.0)}$
Few-shot	R2Gen[28]	2020	10%	23.6	17.7	11.4	13.8	28.5	35.9	33.2	31.3	41.7	57.6	45.7	66.1	56.4
	CMN[32]	2021	10%	29.3	20.5	15.2	16.9	34.8	44.3	39.5	37.3	50.8	72.9	46.2	66.8	58.3
	XProNet[15]	2022	10%	27.5	19.2	14.0	16.3	34.7	38.4	35.0	33.5	44.8	60.7	47.8	67.2	59.5
	PMRG[48]	2024	10%	29.7	20.9	15.5	18.0	35.2	43.1	38.8	36.9	48.0	75.4	50.5	69.2	62.3
	M^{3}FM	Ours	10%	$31.0_{(1.3)}$	$21.5_{(1.2)}$	$16.3_{(0.9)}$	$18.9_{(0.8)}$	$36.3_{(1.5)}$	$51.2_{(1.4)}$	$45.4_{(1.1)}$	$40.3_{(0.9)}$	$55.7_{(1.6)}$	$81.9_{(2.0)}$	$54.7_{(1.0)}$	$73.3_{(1.2)}$	$67.7_{(1.7)}$
Zero-shot	M^{3}FM	Ours	0%	$28.1_{(0.7)}$	$19.8_{(0.9)}$	$13.7_{(0.5)}$	$15.8_{(0.5)}$	$34.2_{(0.6)}$	$47.4_{(0.8)}$	$42.3_{(0.6)}$	$38.8_{(0.4)}$	$51.6_{(1.0)}$	$75.1_{(1.5)}$	$44.5_{(0.8)}$	$62.9_{(1.3)}$	$54.6_{(1.4)}$

We conducted multiple runs with different seeds and reported the mean and standard deviation$_{(STD)}$ of results. The ratio of data indicates the ratio of training splits of downstream datasets used by the methods for model training.

Table 4 reports the AUC score of two infectious diseases, i.e., Tuberculosis (TB) from the Shenzhen Tuberculosis dataset and COVID-19 from the COVID-CXR dataset. As we can see, with 10% data for training, our M^{3}FM significantly outperforms existing best results by up to 5.1% and 3.9% AUC scores on Tuberculosis and COVID-19, respectively. With 100% data for training, our method achieves the best results among the two infectious diseases.

Table 3 further reports the results on 14 non-infectious diseases from the NIH ChestX-ray dataset[35]. It shows that, with extremely limited labels (1%) for training, our approach achieves competitive results with the fully supervised method Model Genesis; with 10% labeled data for training, our approach outperforms the strong baselines, i.e., MRM and REFERS, on Consolidation, Fibrosis, Pleural, and Pneumonia.; The promising results prove the generalization capabilities and the effectiveness of our approach in relaxing the reliance on the labeled data to provide a solid basis for disease diagnosis.

Discussion

In this section, we provide the discussions from different aspects to better understand our approach.

M^{3}FM can beat previous state-of-the-art methods on widely used benchmarks. Table 5 illustrates the performances of our method and existing state-of-the-art performances on the four benchmark datasets. We follow previous works to investigate the performances under the few-shot and fully supervised training setting. As we can see, our approach consistently achieves the best results across all settings and benchmarks. In particular, under the few-shot learning setting, i.e., using 10% labeled data for training, our method can outperform previous fully supervised methods. On the CheXpert dataset, our method with 1% training data achieves an 88.8 AUC score, which surpasses the 88.7 AUC score achieved by previous fully supervised methods. It further proves the effectiveness of our approach for diagnosing rare or new diseases, where the labels are limited. With full training data, our method significantly achieves encouraging results on all benchmarks, which proves the effectiveness of our approach.

M^{3}FM can transfer well to different patient characteristics. In this section, we assess the gender and age sensitivity of our method. To this end, to ensure an even distribution, we further randomly sample and divide the COVID-19 CT dataset[29] into two fine-grained balanced subsets according to Gender and Age, resulting in two gender groups, i.e., Female and Male, and two age groups, i.e., Age <= 50 and Age > 50. Table 6 shows that our method can consistently outperform the existing state-of-the-art models across most metrics. In particular, our method achieves similar results with XProNet (100%) on 'Age > 50'. These results prove the robustness and the generalization capability of our proposed model to different genders and ages.

M^{3}FM can perform well on diseases that are entirely unseen with respect to the more common diseases used in model training. We have further performed a prospective experiment to evaluate the performance of our method on diseases that are unseen with respect to the more common diseases used in model training. In implementations, we pre-train the model on common thorax disease data ((1) Pleural: pleural effusion, pleural other, pneumothorax; (2) Heart: cardiomegaly; (3) Lung: atelectasis, edema, lung lesion, lung opacity, pneumonia; and (4) Others: fracture, consolidation, enlarged cardiomediastinum)) collected before 2019 and evaluated the model on COVID-19 data with different diagnostic characteristics. Therefore, we adopt the prospective experiment to evaluate the ability of our method to "notice" the potential small but important radiological differences between common thorax diseases and COVID-19. We report the performance of disease reporting in Table 7. As we can observe, under limited label settings, our M^{3}FM outperforms previous methods by 11.7% in BLEU-4, 9.6% in ROUGE, and 9.7% in CIDEr scores, and achieves close results compared to previous fully supervised methods, such as XProNet. With full training data as in previous works, our approach consistently achieves the improved performance across all metrics, which demonstrate

Table 2 | Human evaluation of how many times clinicians would have deemed the generated report as "helpful" vs. "unhelpful" in terms of assisting them in writing reports

vs. Methods	Ratio of data	CXR-to-English		CT-to-Chinese		CT-to-English	
		Helpful ↑	Unhelpful ↓	Helpful ↑	Unhelpful ↓	Helpful ↑	Unhelpful ↓
R2Gen[28]	100%	52%	48%	60%	40%	66%	34%
XProNet[15]	100%	64%	36%	74%	26%	80%	20%
M^{3}FM	0%	44%	56%	54%	46%	62%	38%
M^{3}FM	10%	58%	42%	68%	32%	74%	26%
M^{3}FM	100%	76%	24%	84%	16%	88%	12%

the effectiveness of our approach in better capturing the small but important radiological differences between common and rare diseases, resulting in more accurate predictions for rare diseases compared to previous methods.

M^{3}FM could align different domains across modalities and languages. The improved results demonstrates our arguments and shows the effectiveness of our approach in bridging the gaps between different domains. Note that in our zero-shot learning settings, we freeze the text encoder during MultiMedLM training. As we have discussed, a key challenge of our multimodal multidomain multilingual zero-shot clinical diagnosis is to align different domains across modalities and languages. Taking CXR and English as an example, once we have aligned them, i.e., the embeddings of CXR and English are aligned, making the text encoder tunable will cause the model to shift the distribution of English embeddings, which disrupts the alignment between CXR and English domains. It results in impairing the model's performance. Our approach with the tunable encoder achieves 7.5/25.3 BLEU-4 scores on zero-shot CXR-to-English/CT-to-Chinese experiments, respectively, underperforming the zero-shot performance of our full model with the frozen encoder, i.e., 13.7/38.8 BLEU-4 scores.

M^{3}FM could generate desirable reports to relieve the burden of writing medical reports. To better understand our method in low-resource settings, we provide an intuitive example. Figure 2 shows that our method can generate decent and informative reports across different languages, supported by captured important abnormalities, e.g., '*There are calcified granuloma*s.' in the example. The generated reports could relieve the heavy burden of clinicians in writing reports and the captured abnormalities could provide beneficial diagnostic information for supporting them in clinical decision-making. Specifically, for rare/novel diseases and non-English languages, expert-annotated labels for training are extremely limited. The ability of our method to generate zero/few-shot multidomain multilingual reports is encouraging.

Different components introduced in M^{3}FM can contribute to performance. In this section, we report the ablation results in Table 8, which shows that each proposed module can contribute to the improvements. The Base model denotes the encoder-decoder model directly trained on the downstream dataset. In particular, we can find that both MultiMedCLIP and MultiMedLM can bring significant improvements. As a result, under the few-shot learning setting, combining them can enable our method to substantially outperform the Base model by up to 11.3% in the ROUGE-L score and 32.0% in the CIDEr score on the CXR-to-English and CT-to-Chinese, respectively. Under the zero-shot learning setting, by comparing the results of M^{3}FM and (a)/(b), we can notice that the lack of either MultiMedCLIP or MultiMedLM leads to a dramatic drop in performance, which further demonstrates our arguments and the effectiveness of our approach.

M^{3}FM can achieve better performance by using human-annotated translation datasets for training. In our approach, we train the model on the Chinese-English pairs obtained by Google Translate, but the evaluations are conducted on real-world human-written text collected from hospitals. Using machine-translated texts for training would lead to inaccuracies and loss of critical context. Here, we adopt the human-annotated Chinese-English training pairs from COV-CTR[30] for model training. The results are reported in Table 9. As we can see, when using human-annotated translation datasets for training, our method can significantly outperform previous methods. In particular, our method M^{3}FM-Human outperforms several previous methods (R2Gen[28], XProNet[15], PMRG[48]) trained on full data. As a result, when we try to apply our model to underrepresented languages, we can collect a small amount (<100 samples) of human-annotated data, which does not consume significant costs. In contrast, previous efforts usually require tens of thousands (100x), or even more, of annotated data to achieve similar performance. Even in the worst case, when there is no available human-annotated training text for underrepresented languages, our results show that our method can still achieve desirable performance using Google Translate. Meanwhile, we believe that with the rapid development of translators and large language models, e.g., GPT-4[50], the inaccuracies in machine-translated text will be significantly mitigated. It shows the huge potential of our approach to provide a solid basis for non-English languages.

Incorporating multilingual language leads to better language understanding. In this section, we analyze the effect of introducing multilingual language in our approach under the few-shot learning setting. As shown in Table 10, removing either English or Chinese slightly decreases the performance, which indicates that introducing multilingual language improves the performance of the model in each language application scenario, which could be attributed to the fact that our approach could unify different linguistic knowledge from multiple languages, leading to better language understanding and overall improvement.

To better understand the effectiveness of our method, we have further provided a comparison between our method and previous state-of-the-art methods using translation software or LLMs designed for translations. In Table 11, we report the non-English (i.e., Chinese) report generation performance of three medical English report generation models equipped with Google Translate[24] and an English-to-Chinese translation model NLLB-200[51], resulting in six baseline models. As we can see, all baseline models underperform our method of Chinese report generation. The reason is that baseline models suffer from inaccuracies introduced by the translators, including visual irrelevancy and disfluency errors[52,53]. (1) Visual irrelevancy: The visual information in the medical images cannot reach the translators, generating visually irrelevant Chinese reports; (2) Disfluency: The errors in the generated English reports will be propagated and accumulated by the translators, affecting the quality of generated Chinese reports, especially when the translators are trained in a general domain different from medical reports. More importantly, using translators would cause the domain shift problem. For example, if the English report generation model is trained using Hospital A's data, when we apply the model to Hospital B for Chinese report generation using translators, the writing style of the generated Chinese report would be aligned with the original Hospital A, inconsistent with the expectations of Hospital B. Using a unified model, we can not only avoid visual irrelevancy and disfluency errors but also learn both domain knowledge and linguistic knowledge from different languages to achieve improved performances. It further demonstrates the effectiveness of our method in introducing multilingual language within a single framework.

Table 3 | The classification performance AUC score of various methods on 14 non-infectious diseases

Settings	Methods	Year	Ratio of data	Atelectasis	Cardiomegaly	Consolidation	Edema	Effusion	Emphysema	Fibrosis	Hernia	Infiltration	Mass	Nodule	Pleural	Pneumonia	Pneumothorax
Fully supervised	Model Genesis[83]	2021	100%	0.788	0.845	0.792	0.878	0.866	0.897	0.810	0.852	0.711	0.819	0.732	0.758	0.730	0.856
	REFERS[32]	2022	100%	0.830	0.923	0.821	0.902	0.887	0.914	0.839	0.933	0.741	0.855	0.767	0.785	0.770	0.891
	MRM[37]	2023	100%	0.842	0.930	0.822	0.910	0.896	0.943	0.867	0.944	0.718	0.882	0.785	0.814	0.773	0.902
	M^{3}FM	Ours	100%	0.850	0.936	0.831	0.906	0.895	0.945	0.872	0.948	0.749	0.884	0.787	0.819	0.781	0.893
Few-shot	REFERS[32]	2022	10%	0.801	0.898	0.795	0.878	0.875	0.882	0.772	0.861	0.696	0.820	0.728	0.742	0.722	0.856
	MRM[37]	2023	10%	0.823	0.909	0.811	0.890	0.888	0.922	0.848	0.940	0.701	0.866	0.751	0.786	0.743	0.884
	M^{3}FM	Ours	10%	0.841	0.927	0.828	0.899	0.889	0.941	0.870	0.941	0.734	0.878	0.782	0.816	0.775	0.887
	REFERS[32]	2022	1%	0.775	0.856	0.786	0.849	0.854	0.795	0.723	0.771	0.675	0.762	0.665	0.716	0.693	0.817
	MRM[37]	2023	1%	0.788	0.903	0.800	0.865	0.869	0.820	0.719	0.900	0.672	0.823	0.696	0.723	0.696	0.840
	M^{3}FM	Ours	1%	0.793	0.906	0.811	0.887	0.876	0.856	0.744	0.911	0.708	0.825	0.714	0.748	0.728	0.859

Our approach surpasses baselines across most diseases and settings, verifying its generalization capability.

Table 4 | The disease classification performance AUC score on infectious diseases, i.e., Tuberculosis (TB) and COVID-19

Methods	Year	*TB*[31]		*COVID*[34]	
		10%	100%	10%	100%
ConVIRT[40]	2022	0.814	0.970	0.759	0.805
BioViL[81]	2022	0.826	0.975	0.773	0.812
REFERS[32]	2022	-	0.980	-	0.821
MRM[37]	2023	-	-	-	0.858
M^{3}FM	Ours	$0.877_{(0.006)}$	$0.983_{(0.002)}$	$0.812_{(0.007)}$	$0.873_{(0.004)}$

We conducted multiple runs with different seeds and reported the mean and standard deviation$_{(STD)}$ of performance.

Methods

In this section, we introduce our proposed model in detail.

Framework

We will introduce the proposed M^{3}FM for zero-shot disease reporting. Our method includes the MultiMedCLIP and MultiMedLM, where the former introduces a multidomain vision encoder $\mathbb{E}_v(\cdot)$ and a multilingual text encoder $\mathbb{E}_m(\cdot)$, and the latter further introduces a multilingual text decoder $\mathbb{D}_m(\cdot)$. In this paper, we take CT images and CXR images, and English and Chinese as examples. We denote the CXR images, CT images, the English (EN) text, and the Chinese (ZH) text as V_{CXR} and V_{CT}, T_{EN}, and T_{ZH}, respectively. As shown in Fig. 1, MultiMedCLIP aligns and bridges different languages and images in a shared common latent space; and then MultiMedLM reconstructs the text based on the textual representations in the shared latent space; and finally M^{3}FM generates the multilingual reports directly based on the visual representations of input images from different domains in the same latent space. Therefore, our M^{3}FM is defined as follows:

$$
\begin{aligned}
&\text{MultiMedCLIP}\begin{cases}\text{CXR}-\text{English} & :\mathbb{E}_v(V_{\mathrm{CXR}})\overset{\text{Align}}{\longleftrightarrow}\mathbb{E}_m(T_{\mathrm{EN}})\\ \text{CT}-\text{English} & :\mathbb{E}_v(V_{\mathrm{CT}})\overset{\text{Align}}{\longleftrightarrow}\mathbb{E}_m(T_{\mathrm{EN}})\\ \text{Chinese}-\text{English} & :\mathbb{E}_m(T_{\mathrm{ZH}})\overset{\text{Align}}{\longleftrightarrow}\mathbb{E}_m(T_{\mathrm{EN}})\end{cases}\\
&\text{MultiMedLM}\begin{cases}\text{English}-\text{English} & :\mathbb{E}_m(T_{\mathrm{EN}})\overset{\text{Reconstruct}}{\overrightarrow{\mathbb{D}_m(\cdot)}}T_{\mathrm{EN}}\\ \text{Chinese}-\text{Chinese} & :\mathbb{E}_m(T_{\mathrm{ZH}})\overset{\text{Reconstruct}}{\overrightarrow{\mathbb{D}_m(\cdot)}}T_{\mathrm{ZH}}\end{cases}\\
&\text{Zero}-\text{shot}\begin{cases}\text{CXR}-\text{to}-\text{English} & :\mathbb{E}_v(V_{\mathrm{CXR}})\overset{\text{Generate}}{\overrightarrow{\mathbb{D}_m(\cdot)}}T_{\mathrm{EN}}\\ \text{CXR}-\text{to}-\text{Chinese} & :\mathbb{E}_v(V_{\mathrm{CXR}})\overset{\text{Generate}}{\overrightarrow{\mathbb{D}_m(\cdot)}}T_{\mathrm{ZH}}\\ \text{CT}-\text{to}-\text{English} & :\mathbb{E}_v(V_{\mathrm{CT}})\overset{\text{Generate}}{\overrightarrow{\mathbb{D}_m(\cdot)}}T_{\mathrm{EN}}\\ \text{CT}-\text{to}-\text{Chinese} & :\mathbb{E}_v(V_{\mathrm{CT}})\overset{\text{Generate}}{\overrightarrow{\mathbb{D}_m(\cdot)}}T_{\mathrm{ZH}}\end{cases}
\end{aligned}
\tag{1}
$$

As we can see, our method performs zero-shot disease reporting for different types of images in the $\mathbb{D}_m(\mathbb{E}_v(V))\rightarrow T$ pipeline. For the CXR-to-English and CT-to-English tasks, we use evaluation data from a different domain than the training data. Besides, we also do not use any labeled downstream data for training, therefore, these two tasks can still be considered as zero-shot disease reporting.

MultiMedCLIP

We propose MultiMedCLIP to manage distribution drifts between various modalities and languages. Inspired by the great success of contrastive learning methods[54–56], e.g., CLIP[57], in aligning and bridging the visual and textual modalities[57,58], we adopt the Info Noise Contrastive Estimation (InfoNCE)[59] as one of the training objectives. Besides, we further

Table 5 | The disease classification performance AUC score of state-of-the-art methods on benchmark datasets

Methods	Year	CheXpert[38]			NIH[35]			RSNA[41]			SIIM-ACR[42]		
		1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%
GLoRIA[39]	2021	0.865	0.875	0.878	0.671	0.764	0.818	0.860	0.867	0.885	0.744	0.854	0.901
ConVIRT[40]	2022	0.870	0.881	0.881	0.662	0.766	0.813	0.840	0.856	0.876	0.713	0.783	0.900
BioViL[81]	2022	-	-	-	0.695	0.753	0.825	0.823	0.854	0.884	0.695	0.778	0.869
REFERS[32]	2022	0.872	0.881	0.882	0.767	0.809	0.847	0.894	0.916	0.927	-	-	-
MedKLIP[36]	2023	-	-	-	0.772	0.789	0.832	0.873	0.880	0.893	0.853	0.907	0.919
MRM[37]	2023	0.885	0.885	0.887	0.794	0.840	0.859	0.913	0.927	0.933	-	-	-
M^{3}FM	Ours	$0.888_{(0.009)}$	$0.890_{(0.007)}$	$0.893_{(0.004)}$	$0.822_{(0.006)}$	$0.860_{(0.004)}$	$0.863_{(0.003)}$	$0.925_{(0.005)}$	$0.934_{(0.004)}$	$0.940_{(0.004)}$	$0.864_{(0.011)}$	$0.920_{(0.008)}$	$0.932_{(0.005)}$

We conducted multiple runs with different seeds and reported the mean and standard deviation (STD) of performance.

introduce the mean square error (MSE) loss for training, which has shown success in multilingual language modeling[60]. In implementations, given two representations A and B with different distributions, the InfoNCE loss and MSE loss, i.e., InfoNCE(A, B) and MSE(A, B), can minimize the distance between two different distributions, as in the existing CLIP model[57]. As a result, we can exploit the CXR-English, CT-English, and English-non-English pairs to align the CXR and English domains, CT and English domains, and English and non-English domains, respectively. Due to various domains, i.e., CXR, CT, and non-English domains, being aligned with the English domain, they are aligned and bridged with each other[61].

We first sample a batch of N English-centric training pairs (D, T_{EN}), where the data D could be the CXR images, the CT images, and the Chinese text. Then, we denote the (d_i, t_i) as the representations of i^{th} input training pair extracted by MultiMedCLIP. The InfoNCE and MSE losses are defined as:

$$\begin{aligned}\mathcal{L}_{\mathrm{InfoNCE}} &= -\frac{1}{N}\left(\sum_{i=1}^{N}\log\frac{\exp(\langle d_i, t_i\rangle/\tau)}{\sum_{j=1}^{N}\exp(\langle d_i, t_j\rangle/\tau)}\right.\\ &\quad\left.+\sum_{i=1}^{N}\log\frac{\exp(\langle t_i, d_i\rangle/\tau)}{\sum_{j=1}^{N}\exp(\langle t_i, d_j\rangle/\tau)}\right),\\ \mathcal{L}_{\mathrm{MSE}} &= -\tfrac{1}{N}\sum_{i=1}^{N}\left(d_i - t_i\right)^2,\end{aligned} \tag{2}$$

where the $\langle\cdot,\cdot\rangle$, and τ denote the cosine similarity and the temperature parameter[54], respectively.

Therefore, the overall training objective of MultiMedCLIP is:

$$\mathcal{L}_{\mathrm{MultiMedCLIP}} = \lambda_1\cdot\mathcal{L}_{\mathrm{InfoNCE}} + \lambda_2\cdot\mathcal{L}_{\mathrm{MSE}}, \tag{3}$$

where the hyper-parameters $\lambda_1, \lambda_2 \in [0, 1]$ controls the strength of each loss item. As a result, our M^{3}FM can well align visual representations of different domains and textual representations of different languages in a shared latent space, which provides a solid basis for downstream zero-shot inference.

MultiMedLM

MultiMedLM aims to learn to generate the final reports based on the representations extracted by MultiMedCLIP. To this end, we propose to train the model by reconstructing the reports given the textual representations $\mathbb{E}_m(T)$ of input text T, which can be Chinese text and English text. In detail, we adopt the widely used natural language generation loss, i.e., cross-entropy (XE) loss, as the training objective, which is defined as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{MultiMedLM}} &= \mathcal{L}_{\mathrm{XE}}(\mathbb{D}_m(\mathbb{E}_m(T)), T)\\ &= -\sum_{i=1}^{|T|}\log p(y_i|y_{0:i-1}; \mathbb{E}_m(T)),\end{aligned} \tag{4}$$

where $T = \{y_0, y_1, y_2, \ldots, y_{|T|}\}$ (y_0: the begin-of-sentence token; $|T|$: the number of tokens). In particular, y_0 is implemented as the language-specific token (i.e., [EN] and [CH]), enabling the model to be aware of which language to be generated[62,63]. It is worth noting that the introduced reconstructing training can be viewed as unsupervised training and only requires unlabeled text-only data for training. Therefore, our method could be further improved by using more large-scale unlabeled medical texts, e.g., PubMed[64] and MIMIC-III clinical notes[65] used in refs. 66–69.

Training Settings

For a fair comparison[32,37], we adopt the CLIP and ViT pre-trained from existing works[37,70,71] as the backbone to implement the multidomain vision encoder. For the multilingual encoder and multilingual decoder,

Table 6 | Robustness analysis of our approach on the two fine-grained test sets: Gender and Age

Methods	Ratio of data	Gender: Female			Gender: Male			Age <= 50			Age > 50		
		BLEU-4	ROUGE-L	CIDEr	BLEU-4	ROUGE-L	CIDEr	BLEU-4	ROUGE-L	CIDEr	BLEU-4	ROUGE-L	CIDEr
XProNet[15]	100%	44.7	59.7	85.0	44.0	58.9	83.8	43.2	58.5	82.6	45.1	59.9	86.0
M^{3}FM	100%	47.0	64.3	90.9	45.8	63.5	88.1	43.9	62.8	86.7	48.2	64.7	91.4
XProNet[15]	10%	33.8	45.3	61.9	33.0	44.1	59.0	31.9	44.0	58.1	34.7	45.4	62.8
M^{3}FM	10%	41.5	57.7	83.4	38.9	53.3	80.1	36.8	51.3	76.5	43.2	59.1	86.0
M^{3}FM	0%	40.1	52.3	78.2	35.8	50.5	71.3	34.7	45.6	63.5	42.4	57.1	84.7

Table 7 | Prospective evaluations of different methods

Methods	Year	Ratio of Data	BLEU-4	ROUGE	CIDEr
R2Gen[28]	2020	100%	51.4	72.3	65.8
XProNet[15]	2022	100%	57.3	74.9	69.2
PMRG[48]	2024	100%	60.1	76.0	71.3
M^{3}FM	Ours	100%	63.9	78.4	74.2
R2Gen[28]	2020	10%	42.5	63.8	55.6
XProNet[15]	2022	10%	44.4	65.9	58.7
PMRG[48]	2024	10%	47.1	66.4	59.5
M^{3}FM	Ours	10%	52.6	72.8	65.3

The ratio of data indicates the proportion of downstream data used by the methods for fine-tuning. Higher is better in all metrics BLEU-4, ROUGE, and CIDEr.

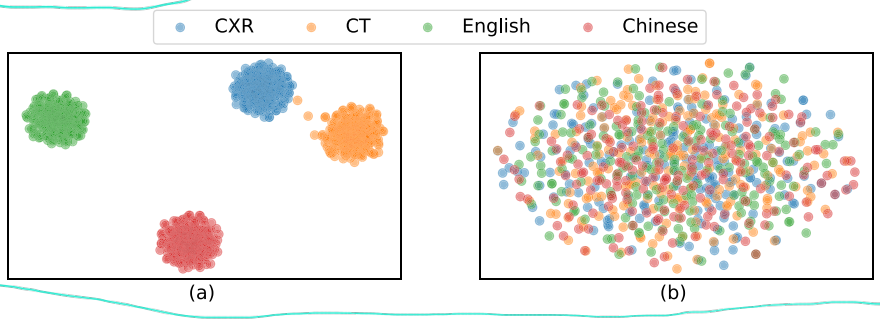

Fig. 2 | We show the t-SNE visualization[49] of CXR, CT, Chinese, and English embeddings. We plot the scatter diagrams with 200 samples for each domain, modality, and language. For comparison, we show the embeddings learned by **a** the Base model (i.e., without our proposal), and **b** our full model M^{3}FM.

Table 8 | Quantitative analysis of the proposed M^{3}FM, which includes MultiMedCLIP and MultiMedLM

Settings	Methods	Components		CXR-to-English (*IU-Xray*[25])			CT-to-Chinese (*COVID-19 CT*[29])		
		MultiMedCLIP	MultiMedLM	BLEU-4	METEOR	ROUGE-L	BLEU-4	ROUGE-L	CIDEr
Few-shot	Base	-	-	10.1	12.5	25.0	27.0	35.8	49.9
	(a)	✓	-	16.0	17.2	37.4	34.9	48.4	68.5
	(b)	-	✓	14.2	17.8	34.7	35.5	48.6	73.2
	M^{3}FM	✓	✓	16.3	18.9	36.3	40.3	55.7	81.9
Zero-shot	(a)	✓	-	8.6	9.2	18.7	28.7	34.3	46.8
	(b)	-	✓	5.2	4.9	11.4	21.5	26.0	34.6
	M^{3}FM	✓	✓	13.7	15.8	34.2	38.8	51.6	75.1

we employ the BERT-Base model[72] as the backbone. In particular, we use six encoder layers and three decoder layers to implement the multilingual encoder and decoder, respectively. For the Chinese text, we adopt the Jieba (https://github.com/fxsjy/jieba) toolkit to process the Chinese text; For the English text, we convert all tokens of reports to lowercase and remove non-alpha tokens. Based on the average performances on the validation set, we set $\lambda_1 = 1$ and $\lambda_2 = 0$ for CXR-English alignment and CT-English alignment and set $\lambda_1 = 0$ and $\lambda_2 = 1$ for Chinese-English alignment. To stabilize the training of the MultiMedLM, we further introduce corruption (i.e., dropout) and add Gaussian noise, as in previous work[61], to the input text for robust text reconstruction. Meanwhile, as our MultiMedLM training can be viewed as unsupervised training, we

can adopt the text-only data from the training splits of downstream datasets for further fine-tuning. During training, we adopt the AdamW optimizer[73] with a learning rate of 1e-4 and a batch size of 32. In detail, we first pre-train the MultiMedCLIP and then pre-train the MultiMedLM. In detail, during fine-tuning, we follow[40] to use a batch size of 64 and a learning rate of 1e-3/5e-4 for parameter optimization on the COVID-CXR dataset and the other datasets. We adopt the PyTorch[74] and V100 GPUs using mixed-precision training[75] for experiments. As mentioned, in our work, we have conducted zero-shot, few-shot, and fully supervised evaluations. For the zero-shot evaluation, we directly freeze the pre-trained model to perform the evaluation on the test splits of downstream evaluation datasets. For few-shot and fully supervised evaluations, following previous works[36,37,48], we further adopt the training splits of downstream evaluation datasets to fine-tune our pre-trained model. During evaluations, we concat the visual and textual embeddings to boost the performance.

We note that the original public datasets only provide 2D slices. For a fair comparison, following existing works[29,30], we directly use the 2D slices for evaluation, which could limit its generalizability to more complex imaging modalities, e.g., the full 3D volumes from CT and MRI. Although our method is designed for 2D images, our work provides a solution to accurately understand different types of medical data across modalities (i.e., images and texts), domains (i.e., CT and CXR), and languages (i.e., English and Chinese). The experiments show that our solution is valid for multiple types of medical data. Therefore, we believe our method is not limited to 2D images and has the potential to be applied to 3D images to deliver zero-shot clinical diagnoses, which is particularly useful in low-data domains. In the future, it would be helpful to explore the effectiveness of our methods on 3D images.

Related works

As shown in Table 12, the related works are introduced from disease reporting and disease classification models. (1) Disease reporting, a.k.a. medical report generation[14,15,26,27,76,77], is similar to the task of image-based language generation[43,78], which aims to generate descriptive text to describe the input image. Conventional foundation methods[15,48,79,80] typically rely on a large amount of labeled data, which, however, are not easy to obtain, especially in the case of rare diseases and non-English languages. (2) To efficiently perform disease classification, inspired by the great success of CLIP[57], different CLIP-based medical foundation models[39,40,70], e.g., BioViL[81], REFERS[32], MedKLIP[36], and MRM[37], have been developed for a better understanding of medical multimodal data. In implementation, they leverage contrastive learning to pre-train the CLIP model[57] using medical data. However, most existing models are CXR-specific, and none of them can deal with multidomain images and multilingual texts in a single framework. Meanwhile, previous works cannot perform zero-shot disease reporting for different domains of languages and images. It is useful for rare diseases and non-English languages, where labeled data for training are frequently much more scarce. Our proposed M^{3}FM model can deal with medical data across modalities (i.e., images and texts), domains (i.e., CT and CXR), and languages (i.e., English and Chinese), and enable zero-shot clinical diagnosis, supporting both disease reporting and classification.

Ethical considerations

We only use public data as secondary and do not recruit any human research participants for this study. Our model was trained and evaluated on public datasets, in which all protected health information (e.g., patient name, sex, gender, and date of birth) is officially de-identified for all datasets used in our experiments.

Recruitment statement

We only secondary use the public data and do not recruit any human research participants for this study. For the public data, all necessary patient/participant consent has been obtained, and the appropriate institutional forms have been officially archived.

Table 9 | Effect of our method using human-annotated text for training

Methods	Year	Ratio of Data	BLEU-4	ROUGE	CIDEr
R2Gen[28]	2020	100%	53.2	73.5	66.4
CMN[82]	2021	100%	56.2	75.6	69.0
XProNet[15]	2022	100%	59.5	76.1	69.7
PMRG[48]	2024	100%	61.3	77.4	72.6
M^{3}FM-Machine	Ours	100%	64.8	79.6	75.1
M^{3}FM-Human	Ours	100%	66.8 (↑2.0)	80.8 (↑1.2)	76.5 (↑1.4)
R2Gen[28]	2020	10%	45.7	66.1	56.4
CMN[82]	2021	10%	46.2	66.8	58.3
XProNet[15]	2022	10%	47.8	67.2	59.5
PMRG[48]	2024	10%	50.5	69.2	62.3
M^{3}FM-Machine	Ours	10%	54.7	73.3	67.7
M^{3}FM-Human	Ours	10%	58.9 (↑4.2)	76.2 (↑2.9)	70.9 (↑3.2)

M^{3}FM-Machine and M^{3}FM-Human denote our method using machine-translated and human-annotated texts for training, respectively. Higher is better in all columns. The (↑ Number) denotes the increased performance compared to our method trained on machine-translated text.

Table 11 | Effect of using translators to generate non-English (i.e., Chinese) medical reports

Methods	CT-to-Chinese (*COVID-19 CT*[29])		
	BLEU-4	ROUGE-L	CIDEr
XProNet[15] + Google Translator[24]	27.8	36.7	47.9
XProNet[15] + NLLB-200[51]	24.6	32.0	40.5
PMRG[48] + Google Translator[24]	30.1	39.3	53.2
PMRG[48] + NLLB-200[51]	26.4	35.5	49.1
M^{3}FM	40.3	55.7	81.9

Table 10 | Effect of introducing multilingual language (English and Chinese) in our method

Methods	Modules		CXR-to-English (*IU-Xray*[25])			CT-to-Chinese (*COVID-19 CT*[29])		
	English	Chinese	BLEU-4	METEOR	ROUGE-L	BLEU-4	ROUGE-L	CIDEr
M^{3}FM	✓	✓	16.3	18.9	36.3	40.3	55.7	81.9
w/o Chinese	✓	-	16.0	17.2	37.4	-	-	-
w/o English	-	✓	-	-	-	34.9	48.4	68.5

Table 12 | Comparison with existing works

Methods	Year	Image domains		Text languages		Target diseases		Experimental settings			Evaluation tasks	
		CXR	CT	English	Chinese	Non-infectious	Infectious	Fully supervised	Few-shot	Zero-shot	Classification	Reporting
XProNet[15]	2022	✓	-	✓	-	✓	-	✓	-	-	-	✓
FS-Gen[84]	2022	✓	-	✓	-	✓	-	-	✓	-	-	✓
DeltaNet[85]	2022	✓	-	✓	-	-	✓	✓	-	-	-	✓
CN-X2RG[80]	2023	✓	-	-	✓	✓	-	✓	-	-	-	✓
PMRG[48]	2024	✓	-	✓	-	✓	-	✓	-	-	-	✓
Dia-LLaMA[86]	2024	-	✓	✓	-	✓	-	✓	-	-	-	✓
BioViL[81]	2022	✓	-	✓	-	✓	-	✓	✓	✓	✓	-
REFERS[32]	2022	✓	-	✓	-	✓	✓	✓	✓	-	✓	-
MedKLIP[36]	2023	✓	-	✓	-	✓	✓	✓	✓	✓	✓	-
MRM[37]	2023	✓	-	✓	-	✓	✓	✓	✓	-	✓	-
M^{3}FM	Ours	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

As we can see, our approach can conduct zero-shot multimodal multidomain multilingual clinical diagnosis, supporting both disease reporting and classification tasks.

Data availability

The data used in our work are all available: (1) MIMIC-CXR is available at https://physionet.org/content/mimic-cxr/2.0.0/. (2) COVID-19-CT-CXR is available at https://github.com/ncbi-nlp/COVID-19-CT-CXR. (3) IU-Xray is available at https://openi.nlm.nih.gov/. (4) COVID-19 CT is available at https://covid19ct.github.io/. (5) COV-CTR dataset is available at https://github.com/mlii0117/COV-CTR. (6) Shenzhen Tuberculosis: https://www.kaggle.com/raddar/tuberculosis-chest-xrays-shenzhen. (7) COVID-CXR is available at https://github.com/ieee8023/covid-chestxray-dataset. (8) NIH ChestX-ray is available at https://nihcc.app.box.com/v/ChestXray-NIHCC. (9) CheXpert is available at https://stanfordmlgroup.github.io/competitions/chexpert/. (10) RSNA Pneumonia is available at https://www.kaggle.com/c/rsna-pneumonia-detection-challenge. (11) SIIM-ACR is available at https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation.

Code availability

Our code is available at: https://github.com/ai-in-health/M3FM.

Received: 18 May 2024; Accepted: 11 November 2024;
Published online: 06 February 2025

References

1. Carlile, M. et al. Deployment of artificial intelligence for radiographic diagnosis of Covid-19 pneumonia in the emergency department. *J. Am. Coll. Emerg. Phys. Open* **1**, 1459–1464 (2020).
2. Wang, X., Peng, Y., Lu, L., Lu, Z. & Summers, R. M. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2018).
3. Liu, F., Wu, X., Ge, S., Fan, W. & Zou, Y. Exploring and distilling posterior and prior knowledge for radiology report generation. In *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2021).
4. Jing, B., Xie, P. & Xing, E. P. On the automatic generation of medical imaging reports. In *Annual Meeting of the Association for Computational Linguistics* (*Association for Computational Linguistics*, 2018).
5. Brady, A., Laoide, R. Ó., Mccarthy, P. & Mcdermott, R. Discrepancy and error in radiology: concepts, causes and consequences. *Ulst. Med. J.* **81**, 3–9 (2012).
6. Liu, F. et al. Auto-encoding knowledge graph for unsupervised medical report generation. In *Annual Conference on Neural Information Processing Systems* (NeurIPS, 2021).
7. Sinsky, C. et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Ann. Intern. Med.* **165**, 753–760 (2016).
8. Weiner, M. & Biondich, P. The influence of information technology on patient-physician relationships. *J. Gen. Intern. Med.* **21**, 35–39 (2006).
9. Tawfik, D. S. et al. Physician burnout, well-being, and work unit safety grades in relationship to reported medical errors. In *Mayo Clinic Proceedings*. 1571–1580 (Elsevier, 2018).
10. West, C. P., Dyrbye, L. N. & Shanafelt, T. D. Physician burnout: contributors, consequences and solutions. *J. Intern. Med.* **283**, 516–529 (2018).
11. Roberts, M. et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for Covid-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **3**, 199–217 (2021).
12. Driggs, D. et al. Machine learning for covid-19 diagnosis and prognostication: lessons for amplifying the signal while reducing the noise. *Radiol. Artif. Intell.* **3**, e210011 (2021).
13. Zhou, S. K. et al. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proc. IEEE* **109**, 820–838 (2021).
14. Jing, B., Wang, Z. & Xing, E. P. Show, describe and conclude: On exploiting the structure information of chest x-ray reports. In *Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, 2019).
15. Wang, J., Bhalerao, A. & He, Y. Cross-modal prototype driven network for radiology report generation. In *European Conference on Computer Vision* (IEEE, 2022).
16. Bhattacharya, S. et al. Deep learning and medical image processing for coronavirus (Covid-19) pandemic: a survey. *Sustain. Cities Soc.* **65**, 102589 (2021).
17. Soomro, T. A. et al. Artificial intelligence (AI) for medical imaging to combat coronavirus disease (Covid-19): a detailed review with direction for future research. *Artif. Intell. Rev.* **55**, 1409–1439 (2022).
18. Liu, F. et al. A medical multimodal large language model for future pandemics. *NPJ Digit. Med.* **6**, 226 (2023).
19. Galimova, R. M., Buzaev, I. V., Ramilevich, K. A., Yuldybaev, L. K. & Shaykhulova, A. F. Artificial intelligence-developments in medicine in the last two years. *Chronic Dis. Transl. Med.* **5**, 64–68 (2019).

20. Chen, A. et al. Inclusion of non–English-speaking participants in pediatric health research: a review. *JAMA Pediatr.* **177**, 81–88 (2023).
21. Budennyy, S. et al. Eco2AI: carbon emissions tracking of machine learning models as the first step towards sustainable AI. In *Doklady Mathematics*. 1–11 (Springer, 2023).
22. Johnson, A. E. W. et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **6**, 317 (2019).
23. Peng, Y. et al. COVID-19-CT-CXR: a freely accessible and weakly labeled chest x-ray and CT image collection on COVID-19 from biomedical literature. *IEEE Trans. Big Data* **7**, 3–12 (2021).
24. Wu, Y. et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Trans. Assoc Comput Linguist* **5**, 339–351 (2016).
25. Demner-Fushman, D. et al. Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Med. Inform. Assoc.* **23**, 304–310 (2016).
26. Li, C. Y., Liang, X., Hu, Z. & Xing, E. P. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *AAAI Conference on Artificial Intelligence* (AAAI, 2019).
27. Li, Y., Liang, X., Hu, Z. & Xing, E. P. Hybrid retrieval-generation reinforced agent for medical image report generation. In *Annual Conference on Neural Information Processing Systems* (NeurIPS, 2018).
28. Chen, Z., Song, Y., Chang, T. & Wan, X. Generating radiology reports via memory-driven transformer. In *Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2020).
29. Liu, G. et al. Medical-vlbert: Medical visual language BERT for COVID-19 CT report generation with alternate learning. *IEEE Trans. Neural Netw. Learn. Syst.* **32**, 3786–3797 (2021).
30. Li, M., Liu, R., Wang, F., Chang, X. & Liang, X. Auxiliary signal-guided knowledge encoder-decoder for medical report generation. *World Wide Web* **26**, 253–270 (2023).
31. Jaeger, S. et al. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quant. imaging Med. Surg.* **4**, 475 (2014).
32. Zhou, H. et al. Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. *Nat. Mach. Intell.* **4**, 32–40 (2022).
33. Cohen, J. P., Morrison, P. & Dao, L. Covid-19 image data collection: Prospective predictions are the future. *Mach Learn Biomed Imaging* **1**, 1–38 (2020).
34. Cohen, J. P. et al. Covid-19 image data collection: Prospective predictions are the future. *Mach. Learn. Biomed. Imaging* **1**, 1–10 (2020).
35. Wang, X. et al. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2017).
36. Wu, C., Zhang, X., Zhang, Y., Wang, Y. & Xie, W. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 21372–21383 (IEEE, 2023).
37. Zhou, H.-Y., Lian, C., Wang, L. & Yu, Y. Advancing radiograph representation learning with masked record modeling. In *The Eleventh International Conference on Learning Representations* (ICLR, 2023).
38. Irvin, J. et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI Conference on Artificial Intelligence* (AAAI, 2019).
39. Huang, S., Shen, L., Lungren, M. P. & Yeung, S. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *International Conference on Computer Vision*. 3922–3931 (IEEE, 2021).
40. Zhang, Y., Jiang, H., Miura, Y., Manning, C. D. & Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. In *Proceedings of Machine Learning for Healthcare* (PMLR, 2022).
41. Shih, G. et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiol. Artif. Intell.* **1**, e180041 (2019).
42. Society for Imaging Informatics in Medicine (SIIM). Siim-acr pneumothorax segmentation. In *Kaggle* (https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation) (2019).
43. Chen, X. et al. Microsoft COCO captions: Data collection and evaluation server. Preprint at https://arxiv.org/abs/1504.00325 (2015).
44. Papineni, K., Roukos, S., Ward, T. & Zhu, W. BLEU: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, 2002).
45. Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, 2004).
46. Banerjee, S. & Lavie, A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *IEEvaluation@ACL* (Association for Computational Linguistics, 2005).
47. Vedantam, R., Zitnick, C. L. & Parikh, D. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2015).
48. Jin, H., Che, H., Lin, Y. & Chen, H. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 2607–2615 (AAAI, 2024).
49. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *JMLR* **9**, 2579–2605 (2008).
50. OpenAI. Gpt-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).
51. Costa-jussà, M. R. et al. No language left behind: Scaling human-centered machine translation. Preprint at https://arxiv.org/abs/2207.04672 (2022).
52. Song, Y., Chen, S., Zhao, Y. & Jin, Q. Unpaired cross-lingual image caption generation with self-supervised rewards. In *Proceedings of the 27th ACM International Conference on Multimedia*. 784–792 (ACM, 2019).
53. Liu, F. et al. Aligning source visual and target language domains for unpaired video captioning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 9255–9268 (2021).
54. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. E. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning* (PMLR, 2020).
55. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. B. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2020).
56. Chen, X., Fan, H., Girshick, R. B. & He, K. Improved baselines with momentum contrastive learning. Preprint at https://arxiv.org/abs/2003.04297 (2020).
57. Radford, A. et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (PMLR, 2021).
58. Jia, C. et al. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning* (PMLR, 2021).
59. Oord, A. V. D., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. Preprint at https://arxiv.org/abs/1807.03748 (2018).
60. Reimers, N. & Gurevych, I. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2020).

61. Yang, B. et al. Zeronlg: Aligning and autoencoding domains for zero-shot multimodal and multilingual natural language generation. *IEEE Trans. Pattern Anal. Mach. Intell.* **46**, 5712–5724 (2024).
62. Tang, Y. et al. Multilingual translation with extensible multilingual pretraining and finetuning. Preprint at https://arxiv.org/abs/2008.00401 (2020).
63. Fan, A. et al. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.* **22**, 4839–4886 (2021).
64. National Institutes of Health. PubMed Corpora (https://pubmed.ncbi.nlm.nih.gov/download/). (National Library of Medicine, 2022).
65. Johnson, A. E. W. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 1–9 (2016).
66. Lee, J. et al. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform* **36**, 1234–1240 (2020).
67. Gu, Y. et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Heal.* **3**, 2:1–2:23 (2022).
68. Alsentzer, E. et al. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop* (Association for Computational Linguistics, 2019).
69. Peng, Y., Yan, S. & Lu, Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMO on ten benchmarking datasets. In *BioNLP@ACL*. 58–65 (Association for Computational Linguistics, 2019).
70. Wang, Z., Wu, Z., Agarwal, D. & Sun, J. Medclip: Contrastive learning from unpaired medical images and text. In *Conference on Empirical Methods in Natural Language Processing*. 3876–3887 (Association for Computational Linguistics, 2022).
71. Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations* (PMLR, 2021).
72. Vaswani, A. et al. Attention is all you need. In *Annual Conference on Neural Information Processing Systems* (NeurIPS, 2017).
73. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations* (ICLR, 2019).
74. Paszke, A. et al. Pytorch: An imperative style, high-performance deep learning library. In *Annual Conference on Neural Information Processing Systems* (NeurIPS, 2019).
75. Micikevicius, P. et al. Mixed precision training. In *International Conference on Learning Representations* (ICLR, 2018).
76. Liu, F. et al. Contrastive attention for automatic chest x-ray report generation. In *Findings of the Association for Computational Linguistics* (Association for Computational Linguistics, 2021).
77. You, D. et al. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In *International Conference on Medical Image Computing and Computer Assisted Intervention* (Springer, 2021).
78. Xu, K. et al. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning* (PMLR, 2015).
79. Zhou, H. et al. A survey of large language models in medicine: Progress, application, and challenge. Preprint at https://arxiv.org/abs/2311.05112 (2023).
80. Tang, W. et al. Generating Chinese radiology reports from X-ray images: a public dataset and an X-ray-to-reports generation method. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 79–88 (Springer, 2023).
81. Boecking, B. et al. Making the most of text semantics to improve biomedical vision–language processing. In *European Conference on Computer Vision*, 1–21 (Springer, 2022).
82. Chen, Z., Shen, Y., Song, Y. & Wan, X. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Association for Computational Linguistics, 2021).
83. Zhou, Z., Sodha, V., Pang, J., Gotway, M. B. & Liang, J. Models genesis. *Med. Image Anal.* **67**, 101840 (2021).
84. Jia, X. et al. Few-shot radiology report generation via knowledge transfer and multi-modal alignment. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1574–1579 (IEEE, 2022).
85. Wu, X. et al. Deltanet: Conditional medical report generation for COVID-19 diagnosis. In *International Conference on Computational Linguistics* (COLING, 2022).
86. Chen, Z., Luo, L., Bie, Y. & Chen, H. Dia-LLaMA: Towards large language model-driven ct report generation. Preprint at https://arxiv.org/abs/2403.16386 (2024).

Acknowledgements

This work was supported in part by the Pandemic Sciences Institute at the University of Oxford; the National Institute for Health Research (NIHR) Oxford Biomedical Research Center (BRC); an NIHR Research Professorship; a Royal Academy of Engineering Research Chair; the Well-come Trust funded VITAL project; the UK Research and Innovation (UKRI); the Engineering and Physical Sciences Research Council (EPSRC); and the InnoHK Hong Kong Center for Cerebro-cardiovascular Engineering (COCHE), and the Clarendon Fund. Xian Wu, Yefeng Zheng, and David A. Clifton are the corresponding authors of this paper. We sincerely thank all the reviewers and editors for their constructive comments and suggestions that substantially improved this paper.

Author contributions

D.C. conceived the project. F.L., X.W., Y.Z. conceived and designed the study, performed the data analysis, and prepared the manuscript. All authors contributed to experiments, results interpretation, and final manuscript preparation. All authors have read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Fenglin Liu, Xian Wu, Yefeng Zheng or David A. Clifton.