Towards All-in-One Medical Image Re-Identification

Yuan Tian[1] Kaiyuan Ji[2] Rongzhao Zhang[1] Yankai Jiang[1] Chunyi Li[3]
Xiaosong Wang[1] Guangtao Zhai[3]✉
[1]Shanghai AI Laboratory
[2]School of Communication and Electronic Engineering, East China Normal University
[3]Institute of Image Communication and Network Engineering, Shanghai Jiao Tong Unversity
tianyuan168326@outlook.com

Abstract

Medical image re-identification (MedReID) is underexplored so far, despite its critical applications in personalized healthcare and privacy protection. In this paper, we introduce a thorough benchmark and a unified model for this problem. First, to handle various medical modalities, we propose a novel Continuous Modality-based Parameter Adapter (ComPA). ComPA condenses medical content into a continuous modality representation and dynamically adjusts the modality-agnostic model with modality-specific parameters at runtime. This allows a single model to adaptively learn and process diverse modality data. Furthermore, we integrate medical priors into our model by aligning it with a bag of pre-trained medical foundation models, in terms of the differential features. Compared to single-image feature, modeling the inter-image difference better fits the re-identification problem, which involves discriminating multiple images. We evaluate the proposed model against 25 foundation models and 8 large multimodal language models across 11 image datasets, demonstrating consistently superior performance. Additionally, we deploy the proposed MedReID technique to two real-world applications, i.e., history-augmented personalized diagnosis and medical privacy protection. Codes and model is available at https://github.com/tianyuan168326/All-in-One-MedReID-Pytorch.

1. Introduction

Medical images [88], such as X-ray images and Computed Tomography (CT) scans, are essential for diagnosing and monitoring various health conditions. Up to 2020, images have accounted for about 90% of all medical data [114].

Despite the large-scale data advanced the computer-aided diagnosis tasks [6, 80], its privacy concern [62] is also serious. It is urgent to (1) efficiently manage patient historical images for personalized healthcare [2, 44, 58] and (2) effectively protect privacy before images are shared [22, 30, 41]. We argue that both sides call for the medical image re-identification (MedReID) technique.

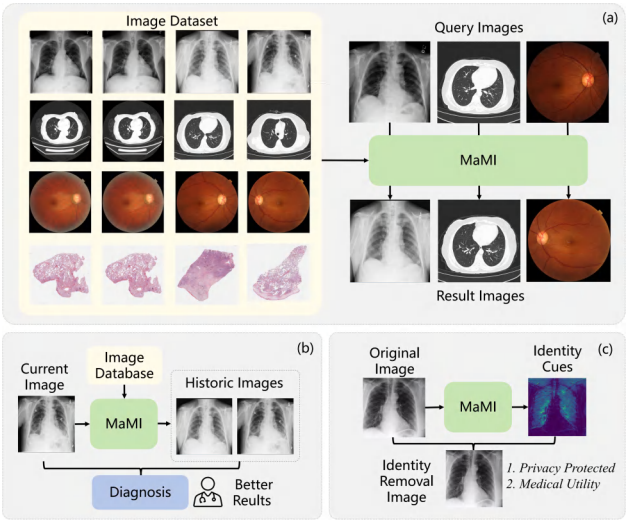

Figure 1. (a) We propose MaMI, an all-in-one modality-adaptive ReID model for medical images. (b) MaMI enhances personalized healthcare by integrating historical medical data. (c) MaMI detects identity cues and removes them from the original images, protecting privacy while maintaining medical utility.

As for historical image management, traditional methods [52, 68] manually pre-link images to patient metadata (e.g., name, medical record numbers), and retrieve images by querying the system with the metadata. However, the links are not always complete and accurate, especially when data are stored in different Picture Archiving and Communication System (PACS) platforms. This requires the MedReID technique to retrieve personal images from poorly organized data, providing accurate historical evidence for disease diagnosis [37].

As for medical image privacy protection, current methods only remove explicit information, such as the patient

✉ Corresponding Author

name [38]. However, some works [21, 56] have found that the identifiable visual information within the images can also breach privacy. A robust MedReID model can detect the identity-related regions of the image. By post-processing these regions, the images become unidentifiable, thereby enhancing their safety before data sharing.

Despite the importance of the MedReID problem, there are only few works investigating this. Fukuta *et al*. [20] and Singh *et al*. [67] exploit the low-level features for identifying fundus images. Packhäuser *et al*. [56] leverages neural networks to identify chest X-ray images. However, all these approaches are designed for one specific modality. They can not enjoy the mutual enhancement from various-modality data sources. Moreover, these models are with less medical priors, which limits their generalization.

In this paper, we introduce a unified MedReID model, termed Modality-adaptive Medical Identifier (MaMI). To handle heterogeneous data from various modalities, MaMI introduces a Continuous Modality-based Parameter Adapter (ComPA). ComPA adapts a modality-agnostic model to modality-specific models at runtime. Given an input image, ComPA generates a continuous modality context, which dynamically produces modality-specific parameters. These parameters are then used to adjust the modality-agnostic model, enabling accurate re-identification of diverse medical modalities with a single model.

Furthermore, we integrate medical priors into our model by aligning it with pre-trained medical foundation models (MFMs), in terms of the inter-image key feature differences. The key features are obtained by attending to the local features using a group of learnable modality-specific query tokens. Compared to the single-image feature, the inter-image differences are more consistent with the ReID, which targets discriminating the identity relation of multiple images.

We compare our model, MaMI, against 25 foundation models and 8 large multi-modal language models across 11 medical image datasets, encompassing a wide range of modalities and body organs, establishing a thorough benchmark for the MedReID problem. Our model consistently outperforms the others. Additionally, we deploy our approach in real-world applications. First, historical data-augmented diagnosis, i.e., MaMI retrieves personalized historical patient data from unorganized datasets, significantly enhancing the accuracy of current medical examinations. Second, privacy protection, i.e., MaMI detects subtle visual cues that reveal patient identity and removes them from images before data sharing, ensuring privacy while preserving medical utility. Our contributions are:

- We propose the first all-in-one medical re-identification model, termed MaMI, capable of re-identifying medical images of various modalities using a single model. We build a thorough and fair benchmark for this novel problem.
- We propose a novel Continuous modality-based Parameter Adapter, which dynamically produces modality-specific parameters, and enables the model to adaptively re-identify different modalities.
- Our model inherits the medical priors from medical foundation models, while adapting them to the ReID problem by inter-image difference modeling.
- We showcase that MaMI can benefit real-world medical applications, e.g., history-augmented healthcare and medical privacy protection.

2. Related Work

Medical Image Re-Identification (MedReID). Numerous medical models focus on automatically diagnosing medical images [6, 90] or retrieving the images by disease features [17, 39]. There are few works focusing on the MedReID problem. Heinrich *et al*. [29] utilized low-level image descriptors such as Sobel [109] to detect patient identity from head CT images. Packhäuser *et al*. [56] and Ganz *et al*. [21] re-identify patients from chest X-ray and histopathology images, respectively. However, all these approaches are limited to a single modality and cannot benefit from large-scale data of various modalities.

Object Re-Identification. Most approaches [98] focus on identifying faces [15, 59, 78, 110], persons [9, 19, 27, 28, 31, 33, 42, 49, 50, 84, 95, 100], animals [36, 64], and vehicles [40, 111, 113]. However, there are few methods dedicated to medical images.

Medical Foundation Models. Early, there are amounts of dedicated models for independent tasks, such as video recognition [12, 18, 70, 72, 73, 75–77], low-level image processing [71, 74, 93, 99, 104, 106–108], and medical image analysis [7, 8, 69] tasks. Later, foundation models [43, 55] are becoming more and more popular, due to their strong generalization capability and strong performance. Recently, numerous medical foundation models, such as X-ray models [81, 83, 97], fundus image models [115][66], and CT models [25, 91], have been continuously proposed. We are the first to adapt their medical priors to the MedReID problem.

Medical Image Domain Adaptation. Medical image domain adaptation addresses domain shifts in imaging data, improving model generalization across different clinical settings [23, 101, 105]. However, these methods mainly focus on diagnosis tasks, how to devise a highly generalizable medical ReID model is left blank.

3. Approach

3.1. Overview

As outlined in Figure 2, we introduce two key ideas to enable a single model to identify various-modality medical images, in an all-in-one manner. First, we achieve modality-

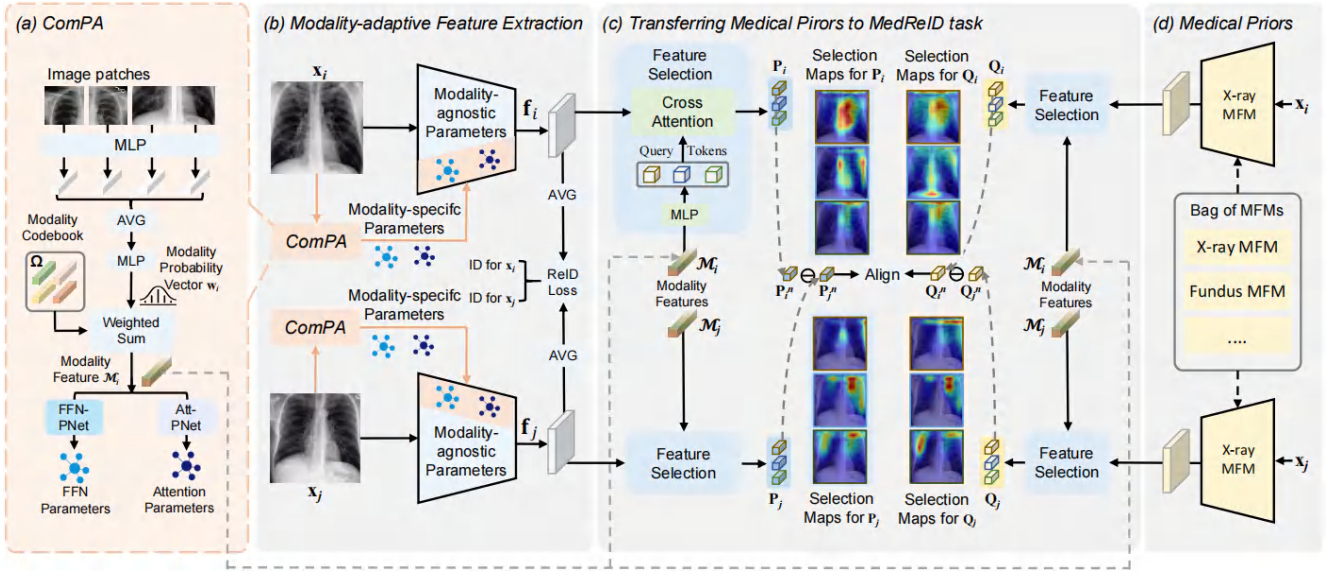

Figure 2. Overview of the proposed all-in-one MedReID framework, namely Modality-adaptive Medical Identifier (MaMI). (a) We introduce a Continuous Modality-based Parameter Adapter (ComPA) to dynamically adjust a modality-agnostic model into an input modality-specific model at runtime. (b) The adjusted model extracts the identity-related visual features from the input medical images. (c) During the optimization, we also transfer the rich medical priors from the (d) medical foundation models (MFMs) to the MedReID task, by aligning the inter-image key differences. We illustrate with X-ray images, though our method also supports other modalities.

adaptive feature extraction, by upgrading a modality-agnostic model to a modality-specific model at runtime. Second, we optimize the model to focus more on medically-relevant regions, by transferring the medical priors within medical foundation models to the MedReID task.

3.2. Modality-Adaptive Feature Extraction

We leverage a typical Transformer network, ViT-Base [16], as the backbone for feature extraction. ViT consists of several attention blocks and feed-forward networks (FFNs). During runtime, we dynamically adjust the network to cater to the current input image.

Motivation. We try to fine-tune a pre-trained ViT model, namely, CLIP [63], towards the MedReID task with two strategies, (1) Single-modality, which separately fine-tunes a specialized model for each modality, and (2) Multiple-modality, which combines the data of all modalities and fine-tunes a unified model. The results are shown in Table 1. Compared to single-modality, the multiple-modality strategy shows improvement in eye fundus modality (76.88% $\rightarrow$ 82.48%), while demonstrating a decrease in X-ray modality (94.21% $\rightarrow$ 92.30%). This indicates that using combined data to learn a unified model benefits some modalities due to more training data, while also limiting the upper bound of some other modalities. We argue that the reason is that, naively putting multiple modalities into a single model, mostly learns the modality-agnostic knowledge, neglecting the modality-specific knowledge.

Continuous modality-based Parameter Adapter (ComPA). To address the above challenge, we propose

Method	X-ray (%)	Fundus (%)
CLIP baseline	33.10	41.14
Single-modality	94.21	76.88
Multiple-modality	92.30	82.48
Continuous-modality (Ours)	96.89	85.71

Table 1. Comparison of different modality handling strategies. We adopt the MIMIC-X [38] and Mess2 [14] datasets to evaluate the performances on X-ray and eye fundus images.

the ComPA to amend the modality-agnostic model with input-modality-specific model parameters, as shown in Figure 2 (b). This effectively decouples the learning of modality-agnostic and modality-specific knowledge.

Rather than employing categorical modality labels, such as 0/1/2 for X-ray/Fundus/CT, ComPA introduces a novel continuous modality representation to handle the modality specificity, as shown in Figure 2 (a). Specifically, given an input image $\mathbf{x}_i \in \mathbb{R}^{3\times H\times W}$, where H and W denote its spatial scales, we convert each 16×16 patch into local modality contexts by a Multilayer Perceptron (MLP), which are averaged to obtain the global modality context.

To improve generalization for images outside the training domain, instead of directly employing the above unconstrained modality context, we constrain the underlying modality representations to be derived from a set of basis centers. Specifically, another MLP transforms the global context into a modality probability vector $\mathbf{w} \in \mathbb{R}^L$, where $L = 32$ denotes the number of all pseudo modalities. Note that this number significantly exceeds the typical number of medical modalities, such as CT and X-ray, due to the diverse imaging styles within a single modality class. For

example, variations in X-ray machines and settings can result in numerous imaging styles [5]. $\mathbf{w}_i$ is then used to compute a weighted sum of learnable modality bases $\mathbf{\Omega} \in \mathbb{R}^{L\times 768}$, producing the ultimate continuous modality feature $\mathcal{M}_i = \mathbf{w}_i\mathbf{\Omega} \in \mathbb{R}^{768}$. $\mathbf{\Omega}$ is randomly initialized and learned with other components in an end-to-end manner.

Given $\mathcal{M}_i$, two MLPs, named Att-PNet and FFN-PNet, generate parameters for the attention and FFN layers of the ViT model, respectively. Nevertheless, directly predicting these parameters would require an infeasibly large number of parameters. For instance, for a ViT-Base model with 86M parameters, given the dimension of $\mathcal{M}_i$ is 768, the last layer of the above two PNets would include $86 \times 768\text{M} \approx 66\text{G}$ parameters, which is intractable. To mitigate this issue, we predict low-rank parameters [32], instead of the full parameters. Meanwhile, we implement the last linear layer of two PNets in a group-wise manner [94], for further reducing the parameter number and computational cost.

The above-generated parameters are merged into the modality-agnostic network in a layer-wise manner. Following LORA [32], we expand the generated low-rank parameters to match the shape of the ViT layers and add them to the corresponding layer parameters.

Our approach shares similarities with recent Mixture-of-Expert (MOE)-LORA paradigms [92, 96], which dynamically weights a series of LORA modules. However, there are two fundamental differences. *Goal Difference*: We aim to perceive the input medical image modality by operating on low-level patch features, whereas MOE-LORA methods utilize high-level semantic features to select different LORAs for various semantic tasks. *Mechanism Difference:* MOE-LORA weights a series of LORA modules fixed in runtime, while our approach directly generates LORA parameters at runtime. This makes our approach fitting the current input image more precisely.

Feature Extraction. The input image $\mathbf{x}_i$ is fed into the above merged network to produce the feature $\mathbf{f}_i \in \mathbb{R}^{768\times h\times w}$, where $h = H/16$ and $w = W/16$ denote the feature resolution. $\mathbf{f}_i$ is then averaged into the global identity feature for identity comparison. For multi-slice modalities, such as CT/MRI scans, we extract feature maps from each slice in the scan and further average them as the scan-wise feature. While more advanced inter-slice operations [57][45] could be employed, we opt for the average operation to maintain the simplicity and efficiency.

3.3. Learning Rich Medical Priors from MFMs

Motivation. With the MedReID loss alone as the learning objective, the model may be biased towards the trivial textures, such as machine noises. In contrast, medical foundation models (MFMs) pre-trained on massive medical images focus on anatomical characteristics, which is more related to the patient intrinsic identity. This motivates us to transfer the rich medical priors within MFMs to our model.

Considering that local features contain more fine-grained information than global features, we use the local feature map of MFMs to guide our model. Furthermore, to close the domain gap between the pre-training task of MFMs and our MedReID task, we propose two strategies, (1) selecting the identity-related key features from the local features, (2) learning the inter-image differential features, instead of the single-image features, as shown in Figure 2 (c).

Key Feature Selection. Given the modality feature $\mathcal{M}_i$ of the image $\mathbf{x}_i$, we use a three-layer MLP to map it into N query tokens $\mathbf{O}_i = \{\mathbf{O}_i^1, \mathbf{O}_i^2, \ldots, \mathbf{O}_i^N\}$, where N denotes the query number. The above tokens are modality-specific, enabling flexible handling of key structures in different modalities. For example, key features for Chest X-ray images include ribbon shape, heart size, and clavicle shape, while key features for fundus images include optic disc shape and vessel patterns, etc.

For the nth query token $\mathbf{O}_i^n \in \mathbb{R}^d$, we calculate its attention map $\mathbf{A}_i^n$ with the image feature map $\mathbf{f}_i$,

$$\mathbf{A}_i^n = \text{Softmax}\left(\frac{\mathbf{O}_i^n \,\text{Linear}(\mathbf{f}_i)}{\sqrt{d}}\right) \in \mathbb{R}^{h\times w}, \tag{1}$$

where the feature dimension d is 768, Linear denotes a linear transformation. Then, $\mathbf{A}_i^n$ attentively pools the feature map $\mathbf{f}_i$, producing the nth key feature $\mathbf{P}_i^n = \sum_{j=o}^{h\times w} \mathbf{A}_i^n[o] \cdot \mathbf{f}_i[o]$, where o denote the spatial position index. For the features from the MFM, we first choose the MFM from the MFM sets, based on the modality of $\mathbf{x}_i$. Then, the nth key feature is selected in a similar manner, denoted as $\mathbf{Q}_i^n$.

Feature Difference Alignment. Considering that the MedReID task requires modeling the subtle differences between different images, we propose to align the inter-image feature difference from our model to those of MFM, instead of directly aligning singe-image feature. Given two medical images, $\mathbf{x}_i$ and $\mathbf{x}_j$, after performing the above feature selection procedures, the nth key features from our model are denoted as $\mathbf{P}_i^n$ and $\mathbf{P}_j^n$, while those from the MFM are denoted as $\mathbf{Q}_i^n$ and $\mathbf{Q}_j^n$. Then, we could use a simple subtraction operation to calculate the nth feature differences, which are given by $\mathbf{u}^n = \mathbf{P}_i^n - \mathbf{P}_j^n$ and $\mathbf{v}^n = \mathbf{Q}_i^n - \mathbf{Q}_j^n$, respectively, for our model and the MFM, respectively. Then, we adopt the contrastive loss to align the above features,

$$\mathcal{L}_{med-align} = \frac{1}{N}\sum_{n=1}^{N} -\log(S(\mathbf{u}^n, \mathbf{v}^n)), \tag{2}$$

where

$$S(u^n, v^n) = \frac{\exp(\mathbf{u}^n \cdot \mathbf{v}^n/\tau)}{\exp(\mathbf{u}^n \cdot \mathbf{v}^n/\tau) + \sum_{k\in\mathcal{N}} \exp(\mathbf{u}^n \cdot \mathbf{v}^k/\tau)}, \tag{3}$$

where $\mathcal{N}$ denotes negative samples, which include non-nth feature differences of the image pair ($\mathbf{x}_i$,$\mathbf{x}_j$), as well as all feature differences from other image pairs. τ denotes the temperature, which is set to 0.07, following MoCo [10].

3.4. Framework Training

To enable our model to discriminate the medical images from different patients, while also of rich medical priors, we adopt the following loss function:

$$\mathcal{L} = \underbrace{\mathcal{L}_{id-classify} + \mathcal{L}_{tri}}_{\text{Identity terms}} + \underbrace{\lambda \mathcal{L}_{\text{med-align}}}_{\text{Medical term}}, \quad (4)$$

where $\mathcal{L}_{id-classify}$ is the cross-entropy loss for patient ID classification, $\mathcal{L}_{tri}$ denotes the triplet loss with soft margin, following [27]. λ denotes the balancing weight.

4. Experiments

4.1. Model Details

Implementation Details. During training, we apply random flipping, random cropping, random erasing [112], and random slice sampling for data augmentation. Specifically, random flipping involves horizontal and vertical flips, while random cropping randomly crops the patches of size 224 $\times$ 224 from the original image. Random slice sampling denotes randomly selecting 8 slices of the CT scans. For each training batch, all images belong to the same modality. λ is set to 0.01. The rank number of the generated parameters is set to 16. The group number of the last linear layer of FFN-PNet and Att-PNet is set to 64. At test time, we resize the shorter side of the images to 256 and then center-crop the middle 224$\times$224 region. For multiple-slice scans, we uniformly sample 8 slices. The initial learning rate is set to 1e-5 and is gradually decayed with the cosine annealing strategy [48]. The total number of training steps is 300,000. The mini-batch size is 196 for single-image medical imaging, while 24 for multiple-slice medical sequences. We utilize the AdamW optimizer [47] implemented in PyTorch [60] with CUDA support. The values of β_1 and β_2 are set to 0.9 and 0.999, respectively. The weight decay is set to 0.05. The entire training process takes about two days on a machine equipped with four NVIDIA RTX 4090 GPUs.

Medical Foundation Models. For X-ray modality, we adopt the Med-Unic [81, 82]. For CT modality, we adopt the CT-CLIP [24, 25]. For fundus image modality, we adopt the RetFound [24, 115]. For histopathology modality, we adopt the CHIEF [86, 87].

Evaluation Metrics. Following [27], we adopt the cumulative matching characteristics (CMC) [4] at Rank-1 (R1), i.e., CMC-R1, to evaluate the ReID performance.

4.2. Datasets

Training and Internal Validation Sets. We re-organize the public datasets with multiple images per patient, excluding those with less than two images, to ensure each patient has at least one query and target images for re-identification. The re-organized datasets include, (1) 111333 **X-ray** images from *MIMIC-X* [38]. (2) 2460 **lung CT** scans from *CCII* [102]. (3) 211 **abdominal CT** scans from *HCC-TACE* [53]. and (4) 35126 **eye fundus** images *EyePACS* (5) 6068 **eye fundus** images from *ODIR* [1]. (6) 542 **histopathology** images from *LUAD* [13]. The train/validation splitting protocols and dataset details are provided in the supplementary material.

External Validation Sets. We also evaluate our model on six external validation sets, the results of which can reflect the model's generalization capability. (1) To build external **X-ray** set, we sample 6569 images of 1000 patients from *Chest-X* [85]. (2) To build **abdominal CT** set, we sample 239 CT scans of 70 patients from *KIRC* [3]. (3) As another **abdominal CT** set, we sample 194 CT scans of 56 patients from *LIHC* [3]. It is worth mentioning that a little proportional of LIHC contains the MRI images. (4) To build **brain MRI** set, we use all 55 MRI scans of 20 patients from *OASIS2* [51]. (5) To build **eye fundus** image set, we use 700 fundus images of 350 patients from *Mess2* [14]. (6) As another **eye fundus** image set, we use all 521 images of 144 patients from *GRAPE* [34].

4.3. Results

MedReID Benchmark. As shown in Table 2, we evaluate various visual foundation models, visual-language foundation models, Person-ReID model, medical foundation models, and single-modality MedReID models. To fully release their potential, we fine-tune some representative models using our training datasets, ensuring a fair comparison.

For *visual foundation models*, contrastive learning approaches like MoCoV3 and DINOv2 achieve decent performance, with accuracies of 84.79% and 91.52% on the CCII (Lung-CT) dataset, respectively. In contrast, masked learning models such as MAE and MaskFeat perform much worse, achieving only 68.33% and 19.95% on the same dataset. However, after fine-tuning for the MedReID task, MAE† outperforms MoCoV3† on most datasets. These findings align with previous research [26, 54], i.e., contrastive features are more linearly separable when being directly deployed, while MAE-style models excel after adaptation due to their more powerful representations.

For *visual-language foundation models*, CLIP consistently outperforms other methods by a substantial margin, achieving 93.02% accuracy on CCII and 70.00% on OASIS2. In contrast, Align and BLIP perform much worse, with accuracies below 20% on OASIS2. These results highlight that CLIP, trained on approximately one billion image-text pairs, learns highly generalizable visual features. After further tuning, the fine-tuned CLIP† shows another significant improvement, surpassing both MoCoV3† and MAE† models by a large margin. For example, on the Chest-X dataset, CLIP†, MoCoV3†, and MAE† achieve, 73.00%, 64.00%, and 68.60%, respectively.

Person ReID method TransReID has generally produced

Method	Dataset										
	MIMIC-X X-Ray	Chest-X X-Ray	CCII Lung-CT	HCC-TACE Ab-CT	KIRC Ab-CT	LIHC Ab-CT	OASIS2 Br-MRI	Mess2 Fundus	ODIR Fundus	GRAPE Fundus	LUAD Histo
Visual Foundation Models											
ImageNet-Sup [16]	34.10	39.90	84.04	50.00	47.14	26.78	47.99	47.14	32.70	44.30	29.13
MoCoV3 [10]	45.10	46.50	84.79	45.24	46.43	30.36	50.00	56.86	42.26	59.93	47.24
DINOv2 [54]	36.40	37.60	91.52	50.00	42.86	28.57	46.00	36.00	23.72	41.37	42.52
BEITv2 [61]	35.10	35.30	89.53	52.38	25.00	30.36	70.00	52.00	37.89	54.40	45.67
CAE [11]	36.20	32.40	71.32	45.24	28.57	21.43	40.00	41.43	28.34	50.16	47.24
MAE [26]	23.80	23.10	68.33	35.71	32.14	23.21	15.99	47.14	23.15	33.55	30.71
MaskFeat [89]	9.20	11.60	19.95	28.57	17.86	14.29	10.00	20.00	8.32	16.94	14.17
MoCoV3†	84.20	64.00	92.52	71.43	46.43	33.93	56.00	70.99	65.90	67.43	51.97
MAE†	88.20	68.60	93.27	76.19	57.14	41.07	60.00	72.57	61.12	60.91	45.67
Visual-Language Foundation Models											
Align [35]	0.40	0.90	43.39	4.76	17.86	12.50	0.00	13.71	3.38	7.82	5.51
BLIP [46]	3.10	4.80	79.05	21.42	25.00	14.29	10.00	33.43	10.54	17.59	24.41
CLIP [63]	33.10	31.60	93.02	45.24	35.71	28.57	68.00	41.14	30.15	50.81	46.46
CLIP†	92.30	73.00	93.52	69.05	57.14	51.79	68.00	73.71	66.06	60.52	40.94
Object ReID Model											
TransReID [27]	29.30	33.90	88.78	33.33	39.29	26.79	69.99	42.29	30.89	36.81	30.71
TransReID†	86.80	68.60	93.52	80.95	47.14	39.29	64.00	74.00	65.52	60.36	54.33
Medical Foundation Models											
BioMedClip [103]	25.20	24.00	82.04	40.48	32.14	26.79	32.00	23.14	19.44	27.68	33.07
RetFound [115]	12.10	15.00	61.85	35.71	39.29	16.07	15.99	53.71	28.83	35.50	25.98
CT-CLIP [25]	3.80	5.30	87.03	9.52	33.14	13.51	5.99	33.14	17.79	16.61	16.54
Med-Unic [81]	48.70	44.90	77.06	33.33	32.14	25.00	23.99	27.71	21.75	35.83	28.35
BioMedClip†	20.10	19.00	83.04	52.38	25.00	26.79	36.00	28.57	18.62	27.69	42.52
RetFound†	54.80	42.80	92.27	66.67	28.57	35.71	50.00	74.14	66.70	61.10	37.80
CT-CLIP†	19.70	19.70	94.04	47.62	21.43	28.57	37.99	29.71	19.93	29.97	42.52
Med-Unic†	92.90	74.30	69.08	57.14	39.29	35.71	41.99	24.57	16.39	22.15	25.20
Modality-specialized MedReID Models											
Packhäuser *et al*. [56]	92.42	88.21	68.63	45.24	35.7	32.11	36.02	23.74	15.18	23.77	29.19
Ganz *et al*. [21]	11.40	11.90	53.62	33.33	39.29	33.93	28.00	27.43	22.65	25.08	56.76
All-in-One MedReID Models											
Ours	**96.89**	**91.49**	**95.01**	**88.09**	**82.68**	**76.82**	**85.00**	**85.71**	**71.34**	**71.00**	**68.75**

Table 2. Comparison of different approaches on medical image re-identification in terms of CMC-R1. † indicates the model is further tuned on the medical datasets same as ours. MIMIC and ChestX indicate the MIMIC-CXR and ChestX-Ray14 datasets. 'Ab-' and 'Br-' denotes the 'Abdominal' and 'Brian'. All models adopt the ViT-Base [16] architecture with a similar parameter number, for a fair comparison. The best and the second best results are marked with **gray bold** and gray, respectively.

suboptimal results when applied to medical images, largely attributed to the substantial domain gap between person images and medical images. After fine-tuning, TransReID† improves somewhat, but still lags far behind CLIP†.

For *medical foundation models*, BioMedClip performs much inferior to CLIP, due to the smaller training dataset PMC-15M. Specialized models like Med-UniC achieve decent performance in their training modality, such as 48.70% accuracy on X-ray images, but perform poorly on other modalities like fundus and CT. This is similar to CT-CLIP and RetFound. After fine-tuning, CT-CLIP†, RetFound†, and Med-UniC† show a further performance boost on the modalities consistent with their pre-training dataset, demonstrating that their pre-trained medical priors are beneficial for the ReID task, but perform unsatisfactorily on other modalities. For example, RetFound† achieves 74.14% on Mess2 (fundus), outperforming the strong CLIP†, but only 42.80% on Chest-X (X-ray).

Single-modality MedReID methods [21, 56] fail to generalize to the modalities out of the training scope. For instance, the X-ray ReID model [56] attains 92.42% accuracy on MIMIC-X (X-ray) but only 15.18% on ODIR (fundus). In contrast, we outperform them by a large margin, due to learning and combining identity cues from several diverse-modality training sources. Additionally, we surpass fine-tuned medical foundation models, such as RetFound† and Med-UniC†, by inheriting and adapting their medical priors to the MedReID problem. Our approach also surpasses various visual foundation models, achieving state-of-the-art performance across all modalities and datasets.

We benchmark eight *large visual-language* model on the medical ReID task. The results are detailed in the supplementary material. Our approach also demonstrates obvious superiority, achieving 98.80% accuracy on Chest-X, while QWen-VL-Max and GPT-4o only achieves 76.80% and 62.50%, respectively.

Finally, we study the cross-modality capability of our model. We evaluate models on a licensed private dataset of 1814 respiratory patients with paired Chest X-ray and CT images. Our all-in-one model learns to associate patient ID across modalities, achieving 87.28% accuracy (Tab. R1), outperforming single-modality-only models. This suggests

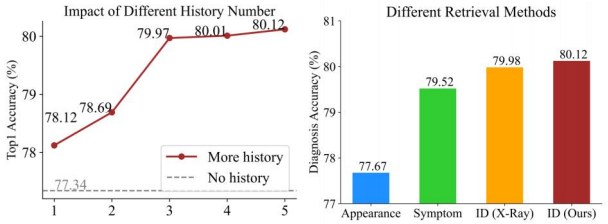

Figure 3. Impact of the historical image number on diagnosis outcome. We use the proposed MaMI to collect the historical image, as the auxiliary information, to aid the diagnosis.

the all-in-one paradigm benefits cross-modality ReID. The fine-tuning further improves the result to 94.38%.

MaMI(X-ray only)	MaMI(CT only)	MaMI(Ours)	MaMI(Tuned)
76.42%	78.19%	87.28%	94.38%

Table 3. Cross-modality ReID: Using X-ray images to retrieve matching CT images of the same patient, on the test set of the respiratory dataset. MaMI (Ours) refers to our all-in-one model trained without cross-modality image pairs. MaMI (Tuned) denotes fine-tuning on the dataset's cross-modality image pair.

Application I: Longitudinal Personalized Healthcare. Further, we consider a realistic scenario where patients' past medical images are not under good management. Given the current image, we use MaMI to retrieve relevant historical images and combine them with the current image for diagnosis. Notably, only the images themselves are utilized, without any historical labels. To integrate features from multiple historical images, we employ a simple MLP.

As shown in Figure 3 *left*, the enhancement through historical image retrieval boosts diagnostic accuracy, due to more longitudinal observations. Specifically, when retrieving five historical images, the accuracy increases from 77.34% to 80.12%, a gain of 2.78%. This demonstrates that MaMI can effectively enhance clinical utility by retrieving relevant historical data from unstructured archives. We further compare different image retrieve approaches, as shown in Figure 3 *right*. Our approach consistently outperforms the appearance(DINOv2 [54])-based, symptom(Med-Unic [81])-based, and X-ray-specialized ReID (Packhäuser *et al*. [56]) methods.

	MaMI	MAE†	CLIP†	Med-Unic†	[56]
Original	91.49%	68.60%	73.00%	74.30%	88.21%
Protected	21.23%	14.52%	11.86%	13.94%	15.68%

Table 4. MedReID on the protected Chest-X dataset. The medical visual cues are detected by our MaMI model, while the privacy removal images can resist attacks from other ReID models. † indicates the model is further tuned on the medical datasets same as ours, for a fair comparison.

Application II: Privacy Protection. We adopt a simple U-Net [65] to predict identity-related visual cues and remove them from the original images. The training objective is to minimize the identity similarity distance between the identity-removed image and the original image, while maximizing their medical feature similarity. Details are in the supplementary material. The identity-removal U-Net is trained on the MIMIC-X dataset and evaluated on the Chest-X dataset. As shown in Table 4, the protected images resist re-identification attacks from various ReID models. We also train disease classification models on both the original and privacy-protected datasets. The accuracies are 81.24% and 80.67%, respectively, indicating that the privacy-protected images preserve the data utility well.

Model	ComPA	MFMs	Internal Validation		*External Validation*	
			MIMIC-X	HCC-TACE	Chest-X	GRAPE
M_{base}	✗	✗	92.29	69.04	86.21	66.51
M_{compa}	✓	✗	96.60	80.95	89.35	69.45
M_{ours}	✓	✓	**96.89**	**88.09**	**91.49**	**71.00**

Table 5. Ablation study on the two core designs, including the continuous-modality parameter adapter (ComPA) and the medical prior learning from a bag of medical foundation models (MFMs).

4.4. Model Analysis

Framework-level Ablation Study. As shown in Table 5, the baseline model M_{base}, which naively fine-tunes the CLIP model on our multi-modality training dataset, results in the poorest performance. Introducing the modality-adaptive component ComPA, the resulting M_{compa} achieves substantial gains on various modalities, i.e., 4.31% and 11.91% gains on MIMIC-X (X-ray) and HCC-TACE (CT), due to handling inter-modality heterogeneity.

Further alignment with Medical Foundation Models (MFMs) to enrich the model's medical prior, resulting in M_{ours}, yields additional performance gains, especially in data-scarce situations. On the HCC-TACE dataset, which contains only 127 training samples, performance increases from 80.95% to 88.09%. This demonstrates that MFMs mitigate the data scarcity issue common in medical imaging. On external datasets (Chest-X and GRAVE), M_{ours} surpasses M_{compa} by 2.14% and 1.55%, respectively. The good results on external validation datasets highlight the generalizability of features derived from MFMs.

In summary, both ComPA and MFM alignment are crucial. The ComPA improves overall performance on various modalities, while MFM alignment mitigates the data-scarcity problem and enhances generalization capability.

Ablation Study on ComPA. We further investigate if all designs within ComPA are necessary. As shown in Table 6, without considering any modality specificity, the baseline model $M_{mod\text{-}no}$ achieves 92.36% and 86.42% on MIMIC-X and Chest-X datasets, respectively.

With the discrete modality labels, such as X-ray as 0, Lung CT as 1, Abdominal CT as 2, etc, as the input condition, the produced M_{mod1} substantially improves upon $M_{mod\text{-}no}$ by 1.45% on Chest-X, proving that modality information is critical for a unified MedReID model. Further

Model	Adaptive	Modality	Codebook	MIMIC-X	Chest-X
$M_{mod\text{-}no}$	✗	-	-	92.36%	86.42%
M_{mod1}	✗	Discrete	-	94.67%	87.87%
M_{mod2}	✓	Continuous	✗	96.78%	90.12%
M_{ours}	✓	Continuous	✓	**96.89%**	**91.49%**

Table 6. Ablation study on ComPA. All models are incorporated with the $\mathcal{L}_{med-align}$ loss item for rigorous ablation.

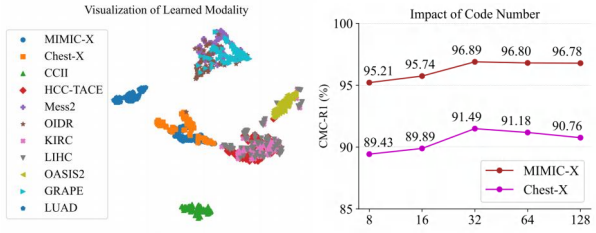

Figure 4. *Left:* t-SNE map of the learned continuous modality. *Right:* Impact of the code number of the codebook within ComPA.

Rank	8	16	32
Chest-X	85.56%	**91.49%**	91.50%

Group	32	64	128
Chest-X	91.45%	**91.49%**	90.52%

Table 7. Impact of (*left*) rank number of the generated parameters, and (*right*) group number of the parameter-generation layers.

introducing instance-adaptive continuous modality design, M_{mod2} surpasses M_{mod1} by another 2.11% on MIMIC-X, indicating that the continuous design better captures data nuances. The introduction of codebook design leads to further improvements, particularly on the external validation set Chest-X (+1.37%). This suggests that the codebook enhances the model's out-of-domain generalizability.

Next, we visualize the learned instance-adaptive modality features by t-SNE[79]. Figure 4 *left* shows a clear separation between different modalities (MIMIC-X and Mess2), while the same modality datasets (Mess2 and OIDR) cluster closely. We observe that LIHC contains some outliers, as a small proportion of LIHC cases are abdominal MRI scans instead of CT scans. Notably, our model autonomously groups OASIS2 MRI images, despite not training with the brain-MRI data, underscoring the high robustness of our modality representation. Then, we train different variant models by tuning the codebook size. As shown in Figure 4 *right*, a small code size such as 8 severely reduces performance on all datasets (96.89% $\rightarrow$ 95.21% for MIMIC-X and 91.49% $\rightarrow$ 89.43% for Chest-X), while a large codebook size such as 128 mainly degrades the model generalizability, i.e., 91.49% $\rightarrow$ 90.76% on external Chest-X.

We further investigate the impact of other ComPA hyperparameters. As shown in Table 7, a small rank constrains model representation, while ranks larger than 16 lead to performance saturation and increased computational cost. For group number, performance is stable at 32 and 64 but degrades at 128 due to much-reduced parameters of FFN- and Att-PNet. For λ, our model achieves very similar performance for 0.1 and 0.01 (91.32% *v.s.* 91.49% on Chest-X), but inferior performance 89.82% for 0.001, due to the too loose medical prior regularization.

Model	Feature	Inter-Image	Relation Operator	Chest-X
$M_{med\text{-}no}$	-	-	-	88.54%
M_{med1}	Global	-	-	88.87%
M_{med2}	Local	✗	-	89.10%
M_{med3}	Selected	✗	-	90.02%
M_{med4}	Selected	✓	MLP	91.22%
M_{ours}	Selected	✓	Subtraction	**91.49%**

Table 8. Strategies of learning medical priors from MFMs.

Learning Strategy of Medical Priors. As shown in Table 8, compared to the baseline model $M_{med\text{-}no}$ (no medical priors), introducing global medical priors (M_{med1}) yields minimal gains, as global features fail to capture subtle identity information. Naive local priors (M_{med2}) marginally surpass M_{med1} by 0.23%. After the modality-specific feature selection operation, feature semantics is significantly improved, reflected by a substantial gain of 0.92%.

Replacing single-image feature alignment with inter-image feature relation alignment, where the relation feature is obtained by concatenating the features from different images and feeding them into a three-layer MLP, further boosts performance by 1.20% in M_{med4}. Finally, substituting the MLP with a subtraction operation in M_{ours} enforces the model's focus on subtle image differences, achieving a final performance of 91.49%. This proves that modeling the inter-image relationship is crucial for the ReID problem, regardless of the specific relationship operator. Both the MLP and our subtraction operation achieve good results.

Model Complexity. With batch size 128, the inference time of our model is 151.56 ms on a machine with an NVIDIA 4090 GPU, compared to 141.21 ms for the vanilla ViT-Base. The ComPA module only additionally consumes 10ms, as it primarily consists of several simple MLPs to compute modality-specific parameters. Given its brought substantial result gains, this minor increase in latency is justified. Further, the MFM alignment procedure incurs no inference cost, as it only regularizes the training procedure.

5. Conclusion

In this paper, we have introduced a comprehensive benchmark and a unified model for a novel MedReID problem, covering a wide range of medical modalities. We have proposed a modality-adaptive architecture to enable a single model to handle diverse medical modalities at runtime. Additionally, we integrate medical priors into our model by exploiting the pre-trained medical foundation models. Our model substantially outperforms all previous approaches.

Acknowledgement This work was supported by Shanghai Artificial Intelligence Laboratory, National Natural Science Foundation of China (Grant No.72293585, No.72293580, No.62225112), the Fundamental Research Funds for the Central Universities, National Key R&D Program of China 2021YFE0206700, Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), and STCSM 22DZ2229005.

References

[1] Peking university international competition on ocular disease intelligent recognition (odir-2019). https://odir2019.grandchallenge.org/. Accessed: 2022-02-10. 5

[2] Hugo JWL Aerts. The potential of radiomic-based phenotyping in precision medicine: a review. *JAMA oncology*, 2 (12):1636–1642, 2016. 1

[3] O Akin, P Elnajjar, M Heller, R Jarosz, BJ Erickson, S Kirk, Y Lee, MW Linehan, R Gautam, R Vikram, et al. The cancer genome atlas kidney renal clear cell carcinoma collection (tcga-kirc)(version 3). the cancer imaging archive, 2016. 5

[4] Ruud M Bolle, Jonathan H Connell, Sharath Pankanti, Nalini K Ratha, and Andrew W Senior. The relation between the roc curve and the cmc. In *Fourth IEEE workshop on automatic identification advanced technologies (AutoID'05)*, pages 15–20. IEEE, 2005. 5

[5] Jerrold T Bushberg and John M Boone. *The essential physics of medical imaging*. Lippincott Williams & Wilkins, 2011. 4

[6] Lei Cai, Jingyang Gao, and Di Zhao. A review of the application of deep learning in medical image classification and segmentation. *Annals of translational medicine*, 8(11), 2020. 1, 2

[7] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022. 2

[8] Haiwen Chen, Zhiyuan Qu, Yuan Tian, Ning Jiang, Yuan Qin, Jie Gao, Ruoyan Zhang, Yanning Ma, Zuolin Jin, and Guangtao Zhai. A cross-temporal multimodal fusion system based on deep learning for orthodontic monitoring. *Computers in Biology and Medicine*, 180:109025, 2024. 2

[9] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15050–15061, 2023. 2

[10] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021. 4, 6

[11] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, 132(1):208–223, 2024. 6

[12] Zijian Chen, Wei Sun, Yuan Tian, Jun Jia, Zicheng Zhang, Wang Jiarui, Ru Huang, Xiongkuo Min, Guangtao Zhai, and Wenjun Zhang. Gaia: Rethinking action quality assessment for ai-generated videos. *Advances in Neural Information Processing Systems*, 37:40111–40144, 2024. 2

[13] National Cancer Institute Clinical Proteomic Tumor Analysis Consortium et al. The clinical proteomic tumor analysis consortium lung adenocarcinoma collection (cptac-luad)(version 12). the cancer imaging archive website. 5

[14] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, John-Richard Ordóñez-Varela, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, pages 231–234, 2014. 3, 5

[15] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 2

[16] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 6

[17] Jiansheng Fang, Huazhu Fu, and Jiang Liu. Deep triplet hashing network for case-based medical image retrieval. *Medical image analysis*, 69:101981, 2021. 2

[18] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 2

[19] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. Unsupervised pre-training for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14750–14759, 2021. 2

[20] Keisuke Fukuta, Toshiaki Nakagawa, Yoshinori Hayashi, Yuji Hatanaka, Takeshi Hara, and Hiroshi Fujita. Personal identification based on blood vessels of retinal fundus images. In *Medical Imaging 2008: Image Processing*, pages 630–638. SPIE, 2008. 2

[21] Jonathan Ganz, Jonas Ammeling, Samir Jabari, Katharina Breininger, and Marc Aubreville. Re-identification from histopathology images. *Medical Image Analysis*, 99: 103335, 2025. 2, 6

[22] Aggeliki Giakoumaki, Sotiris Pavlopoulos, and Dimitris Koutsouris. Secure and efficient health data management through multiple watermarking on medical images. *Medical and Biological Engineering and Computing*, 44:619–631, 2006. 1

[23] Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021. 2

[24] Ibrahim Ethem Hamamci. CT-CLIP: Contrastive learning of image and text representations for computed tomography. https://github.com/ibrahimethemhamamci/CT-CLIP, 2024. 5

[25] Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Bastian Wittmann, Enis Simsar, Mehmet Simsar, et al. A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. *arXiv preprint arXiv:2403.17834*, 2024. 2, 5, 6

[26] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 5, 6

[27] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022, 2021. 2, 5, 6

[28] Weizhen He, Yiheng Deng, Shixiang Tang, Qihao Chen, Qingsong Xie, Yizhou Wang, Lei Bai, Feng Zhu, Rui Zhao, Wanli Ouyang, et al. Instruct-reid: A multi-purpose person re-identification task with instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17521–17531, 2024. 2

[29] Andreas Heinrich. Automatic personal identification using a single ct image. *European Radiology*, pages 1–12, 2024. 2

[30] Mattias P Heinrich and Lasse Hansen. Implicit neural compression for privacy preserving medical image sharing. In *Medical Imaging with Deep Learning*. 1

[31] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 2

[32] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 4

[33] Hezhen Hu, Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Lu Yuan, Dong Chen, and Houqiang Li. Personmae: Person re-identification pre-training with masked autoencoders. *IEEE Transactions on Multimedia*, 2024. 2

[34] Xiaoling Huang, Xiangyin Kong, Ziyan Shen, Jing Ouyang, Yunxiang Li, Kai Jin, and Juan Ye. Grape: A multi-modal dataset of longitudinal follow-up visual field and fundus images for glaucoma management. *Scientific Data*, 10(1):520, 2023. 5

[35] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 6

[36] Bingliang Jiao, Lingqiao Liu, Liying Gao, Ruiqi Wu, Guosheng Lin, Peng Wang, and Yanning Zhang. Toward re-identifying any animal. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[37] Cheng Jin, Heng Yu, Jia Ke, Peirong Ding, Yongju Yi, Xiaofeng Jiang, Xin Duan, Jinghua Tang, Daniel T Chang, Xiaojian Wu, et al. Predicting treatment response from longitudinal images using multi-task deep learning. *Nature communications*, 12(1):1851, 2021. 1

[38] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019. 2, 3, 5

[39] Farnaz Khun Jush, Tuan Truong, Steffen Vogler, and Matthias Lenga. Medical image retrieval using pretrained embeddings. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024. 2

[40] Sultan Daud Khan and Habib Ullah. A survey of advances in vision-based vehicle re-identification. *Computer Vision and Image Understanding*, 182:50–63, 2019. 2

[41] Bach Ngoc Kim, Jose Dolz, Pierre-Marc Jodoin, and Christian Desrosiers. Privacy-net: an adversarial approach for identity-obfuscated segmentation of medical images. *IEEE Transactions on Medical Imaging*, 40(7):1737–1749, 2021. 1

[42] Minsu Kim, Seungryong Kim, Jungin Park, Seongheon Park, and Kwanghoon Sohn. Partmix: Regularization strategy to learn part discovery for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18621–18632, 2023. 2

[43] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 2

[44] Philippe Lambin, Ralph TH Leijenaar, Timo M Deist, Jurgen Peerlings, Evelyn EC De Jong, Janita Van Timmeren, Sebastian Sanduleanu, Ruben THM Larue, Aniek JG Even, Arthur Jochems, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nature reviews Clinical oncology*, 14(12):749–762, 2017. 1

[45] Han Li, Long Chen, Hu Han, and S Kevin Zhou. Satr: Slice attention with transformer for universal lesion detection. In *International conference on medical image computing and computer-assisted intervention*, pages 163–174. Springer, 2022. 4

[46] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 6

[47] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[48] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[49] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 2

[50] Hao Luo, Pichao Wang, Yi Xu, Feng Ding, Yanxin Zhou, Fan Wang, Hao Li, and Rong Jin. Self-supervised pre-training for transformer-based person re-identification. *arXiv preprint arXiv:2111.12084*, 2021. 2

[51] Daniel S Marcus, Anthony F Fotenos, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of

imaging studies: longitudinal mri data in nondemented and demented older adults. *Journal of cognitive neuroscience*, 22(12):2677–2684, 2010. 5

[52] Dietrich Meyer-Ebrecht. Picture archiving and communication systems (pacs) for medical application. *International journal of bio-medical computing*, 35(2):91–124, 1994. 1

[53] A Moawad, D Fuentes, A Morshid, A Khalaf, M Elmohr, A Abusaif, JD Hazle, AO Kaseb, M Hassan, A Mahvash, et al. Multimodality annotated hcc cases with and without advanced imaging segmentation [data set]. *The Cancer Imaging Archive*, 2021. 5

[54] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*. 5, 6, 7

[55] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[56] Kai Packhäuser, Sebastian Gündel, Nicolas Münster, Christopher Syben, Vincent Christlein, and Andreas Maier. Deep learning-based patient re-identification is able to exploit the biometric nature of medical chest x-ray data. *Scientific Reports*, 12(1):14851, 2022. 2, 6, 7

[57] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35:26462–26477, 2022. 4

[58] Andreas S Panayides, Amir Amini, Nenad D Filipovic, Ashish Sharma, Sotirios A Tsaftaris, Alistair Young, David Foran, Nhan Do, Spyretta Golemati, Tahsin Kurc, et al. Ai in medical imaging informatics: current challenges and future directions. *IEEE journal of biomedical and health informatics*, 24(7):1837–1857, 2020. 1

[59] Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association, 2015. 2

[60] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[61] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 6

[62] W Nicholson Price and I Glenn Cohen. Privacy in the age of medical big data. *Nature medicine*, 25(1):37–43, 2019. 1

[63] 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*, pages 8748–8763. PMLR, 2021. 3, 6

[64] Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1025–1034, 2021. 2

[65] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 7

[66] Danli Shi, Weiyi Zhang, Jiancheng Yang, Siyu Huang, Xiaolan Chen, Mayinuer Yusufu, Kai Jin, Shan Lin, Shunming Liu, Qing Zhang, et al. Eyeclip: A visual-language foundation model for multi-modal ophthalmic image analysis. *arXiv preprint arXiv:2409.06644*, 2024. 2

[67] Anushikha Singh, Malay Kishore Dutta, and Dilip Kumar Sharma. Unique identification code for medical fundus images using blood vessel pattern for tele-ophthalmology applications. *computer methods and programs in biomedicine*, 135:61–75, 2016. 2

[68] Chia-Chi Teng, Jonathan Mitchell, Christopher Walker, Alex Swan, Cesar Davila, David Howard, and Travis Needham. A medical image archive solution in the cloud. In *2010 IEEE International Conference on Software Engineering and Service Sciences*, pages 431–434. IEEE, 2010. 1

[69] Yuan Tian, Xiongkuo Min, Guangtao Zhai, and Zhiyong Gao. Video-based early asd detection via temporal pyramid networks. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 272–277. IEEE, 2019. 2

[70] Yuan Tian, Zhaohui Che, Wenbo Bao, Guangtao Zhai, and Zhiyong Gao. Self-supervised motion representation via scattering local motion cues. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 71–89. Springer, 2020. 2

[71] Yuan Tian, Guo Lu, Xiongkuo Min, Zhaohui Che, Guangtao Zhai, Guodong Guo, and Zhiyong Gao. Self-conditioned probabilistic learning of video rescaling. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4490–4499, 2021. 2

[72] Yuan Tian, Yichao Yan, Guangtao Zhai, Guodong Guo, and Zhiyong Gao. Ean: event adaptive network for enhanced action recognition. *International Journal of Computer Vision*, 130(10):2453–2471, 2022. 2

[73] Yuan Tian, Guo Lu, Guangtao Zhai, and Zhiyong Gao. Non-semantics suppressed mask learning for unsupervised video semantic compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13610–13622, 2023. 2

[74] Yuan Tian, Yichao Yan, Guangtao Zhai, Li Chen, and Zhiyong Gao. Clsa: a contrastive learning framework with selective aggregation for video rescaling. *IEEE Transactions on Image Processing*, 32:1300–1314, 2023. 2

[75] Yuan Tian, Guo Lu, Yichao Yan, Guangtao Zhai, Li Chen, and Zhiyong Gao. A coding framework and benchmark towards low-bitrate video understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[76] Yuan Tian, Guo Lu, and Guangtao Zhai. Free-vsc: Free semantics from visual foundation models for unsupervised video semantic compression. In *European Conference on Computer Vision*, pages 163–183. Springer, 2024.

[77] Yuan Tian, Guo Lu, and Guangtao Zhai. Smc++: Masked learning of unsupervised video semantic compression. *arXiv preprint arXiv:2406.04765*, 2024. 2

[78] Yuan Tian, Shuo Wang, and Guangtao Zhai. Medical manifestation-aware de-identification. *arXiv preprint arXiv:2412.10804*, 2024. 2

[79] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 8

[80] Bram Van Ginneken, Cornelia M Schaefer-Prokop, and Mathias Prokop. Computer-aided diagnosis: how to move from the laboratory to the clinic. *Radiology*, 261(3):719–732, 2011. 1

[81] Zhongwei Wan, Che Liu, Mi Zhang, Jie Fu, Benyou Wang, Sibo Cheng, Lei Ma, César Quilodrán-Casas, and Rossella Arcucci. Med-unic: Unifying cross-lingual medical vision-language pre-training by diminishing bias. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 5, 6, 7

[82] Zhongwei Wan, Che Liu, Mi Zhang, Jie Fu, Benyou Wang, Sibo Cheng, Lei Ma, César Quilodrán-Casas, and Rossella Arcucci. Med-unic: Unifying cross-lingual medical vision-language pre-training by diminishing bias. https://github.com/SUSTechBruce/Med-UniC, 2024. 5

[83] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in Neural Information Processing Systems*, 35:33536–33549, 2022. 2

[84] Guangcong Wang, Jianhuang Lai, Peigen Huang, and Xiaohua Xie. Spatial-temporal person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8933–8940, 2019. 2

[85] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 5

[86] Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, Hongping Tang, Kanran Wang, Yu Li, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, pages 1–9, 2024. 5

[87] Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, Hongping Tang, Kanran Wang, Yu Li, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. https://github.com/hms-dbmi/CHIEF, 2024. 5

[88] Andrew Webb. *Introduction to biomedical imaging*. John Wiley & Sons, 2022. 1

[89] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022. 6

[90] Yi Wei, Meiyi Yang, Meng Zhang, Feifei Gao, Ning Zhang, Fubi Hu, Xiao Zhang, Shasha Zhang, Zixing Huang, Lifeng Xu, et al. Focal liver lesion diagnosis with deep learning and multistage ct imaging. *Nature Communications*, 15(1): 7040, 2024. 2

[91] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*, 2023. 2

[92] Xun Wu, Shaohan Huang, and Furu Wei. Mixture of lora experts. In *The Twelfth International Conference on Learning Representations*. 4

[93] Rui Xie, Chen Zhao, Kai Zhang, Zhenyu Zhang, Jun Zhou, Jian Yang, and Ying Tai. Addsr: Accelerating diffusion-based blind super-resolution with adversarial diffusion distillation. *arXiv preprint arXiv:2404.01717*, 2024. 2

[94] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 4

[95] Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. Clip-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing*, 2023. 2

[96] Yuqi Yang, Peng-Tao Jiang, Qibin Hou, Hao Zhang, Jinwei Chen, and Bo Li. Multi-task dense prediction via mixture of low-rank experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27927–27937, 2024. 4

[97] Jingfeng Yao, Xinggang Wang, Yuehao Song, Huangxuan Zhao, Jun Ma, Yajie Chen, Wenyu Liu, and Bo Wang. Eva-x: A foundation model for general chest x-ray analysis with self-supervised learning. *arXiv preprint arXiv:2405.05237*, 2024. 2

[98] Mang Ye, Shuoyi Chen, Chenyue Li, Wei-Shi Zheng, David Crandall, and Bo Du. Transformer for object re-identification: A survey. *arXiv preprint arXiv:2401.06960*, 2024. 2

[99] Fuwang Yi, Mianyi Chen, Wei Sun, Xiongkuo Min, Yuan Tian, and Guangtao Zhai. Attention based network for no-reference ugc video quality assessment. In *2021 IEEE international conference on image processing (ICIP)*, pages 1414–1418. IEEE, 2021. 2

[100] Jiahang Yin, Ancong Wu, and Wei-Shi Zheng. Fine-grained person re-identification. *International journal of computer vision*, 128(6):1654–1672, 2020. 2

[101] Guodong Zeng, Till D Lerch, Florian Schmaranzer, Guoyan Zheng, Jürgen Burger, Kate Gerber, Moritz Tannast, Klaus Siebenrock, and Nicolas Gerber. Semantic consistent unsupervised domain adaptation for cross-modality medical

image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 201–210. Springer, 2021. 2

[102] Kang Zhang, Xiaohong Liu, Jun Shen, Zhihuan Li, Ye Sang, Xingwang Wu, Yunfei Zha, Wenhua Liang, Chengdi Wang, Ke Wang, et al. Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography. *Cell*, 181(6):1423–1433, 2020. 5

[103] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023. 6

[104] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 2

[105] Yifan Zhang, Ying Wei, Qingyao Wu, Peilin Zhao, Shuaicheng Niu, Junzhou Huang, and Mingkui Tan. Collaborative unsupervised domain adaptation for medical image diagnosis. *IEEE Transactions on Image Processing*, 29: 7834–7844, 2020. 2

[106] Chen Zhao, Weiling Cai, Chenyu Dong, and Chengwei Hu. Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8281–8291, 2024. 2

[107] Chen Zhao, Weiling Cai, Chengwei Hu, and Zheng Yuan. Cycle contrastive adversarial learning with structural consistency for unsupervised high-quality image deraining transformer. *Neural Networks*, 178:106428, 2024.

[108] Chen Zhao, Wei-Ling Cai, and Zheng Yuan. Spectral normalization and dual contrastive regularization for image-to-image translation. *The Visual Computer*, pages 1–12, 2024. 2

[109] Sanqiang Zhao, Yongsheng Gao, and Baochang Zhang. Sobel-lbp. In *2008 15th IEEE International Conference on Image Processing*, pages 2144–2147. IEEE, 2008. 2

[110] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *ACM computing surveys (CSUR)*, 35(4):399–458, 2003. 2

[111] Zhedong Zheng, Tao Ruan, Yunchao Wei, Yi Yang, and Tao Mei. Vehiclenet: Learning robust visual representation for vehicle re-identification. *IEEE Transactions on Multimedia*, 23:2683–2693, 2020. 2

[112] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13001–13008, 2020. 5

[113] Qinqin Zhou, Bineng Zhong, Xiangyuan Lan, Gan Sun, Yulun Zhang, Baochang Zhang, and Rongrong Ji. Fine-grained spatial alignment model for person re-identification with focal triplet loss. *IEEE Transactions on Image Processing*, 29:7578–7589, 2020. 2

[114] S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109 (5):820–838, 2021. 1

[115] Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023. 2, 5, 6