Semantics versus Identity: A Divide-and-Conquer Approach towards Adjustable Medical Image De-Identification

Anonymous ICCV submission

Paper ID 822

Abstract

Medical imaging has significantly advanced computer-aided diagnosis, yet its re-identification (ReID) risks raise critical privacy concerns, calling for de-identification (DeID) techniques. Unfortunately, existing DeID methods neither particularly preserve medical semantics, nor are flexibly adjustable towards different privacy levels. To address these issues, we propose a divide-and-conquer framework comprising two steps: (1) Identity-Blocking, which blocks varying proportions of identity-related regions, to achieve different privacy levels; and (2) Medical-Semantics-Compensation, which leverages pre-trained Medical Foundation Models (MFMs) to extract medical semantic features to compensate the blocked regions. Moreover, recognizing that features from MFMs may still contain residual identity information, we introduce a Minimum Description Length principle-based feature decoupling strategy, to effectively decouple and discard such identity components. Extensive evaluations against existing approaches across seven datasets and three downstream tasks, demonstrates our state-of-the-art performance.

1. Introduction

In the era of digital medicine, large-scale medical images, such as X-rays and fundus photographs [6], are routinely processed by AI-based diagnostic models [29, 84, 94, 107, 109] to aid clinical decision-making. However, the increasing availability of these images raises significant concerns regarding patient privacy [18, 56, 78, 86].

Although explicit personal details such as patient name can be easily removed from medical image headers [1, 70, 82] or burned-in texts [90, 111], re-identification (ReID) remains feasible for the intrinsic bio-identifiers, such as anatomical markers visible in chest X-rays [36, 73]. This enables sensitive information breaches [9, 51, 87], compromising patient privacy (see Figure 1(a)).

Several studies have attempted to defend against ReID attacks. For instance, some approaches [12, 31, 33, 45] focus on removing facial features to obfuscate identity. However, such methods cannot be applied to other body parts like the chest, where identity information is deeply interwoven with diagnostic semantics. Standard image filtering techniques, such as blurring [91], pixelation [42], and masking [96], indiscriminately degrade critical diagnostic details, thereby impairing downstream medical task performance. Moreover, under high privacy settings, the severe degradation of image quality further deteriorates task performance. Differential privacy methods [20, 28, 57, 100] mitigate identity information via noise injection, but this operation also perturbs diagnostic features. Identity adversarial learning methods [54, 74] train generators by jointly maximizing identity discrepancy between the generated and original images, while minimizing distortion of medical information. Nevertheless, given the inherent entanglement

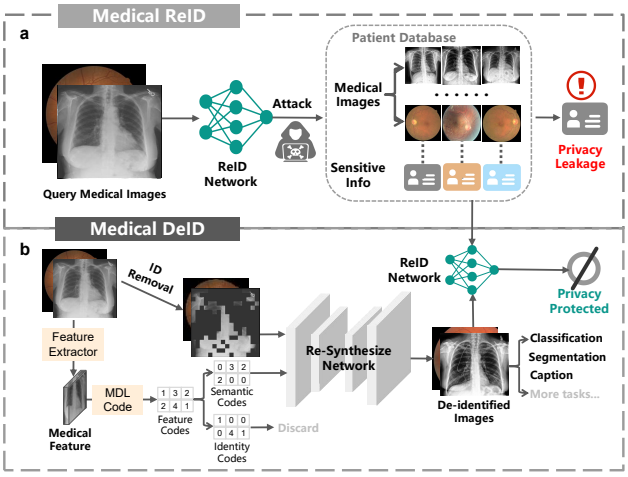

Figure 1. (a) Given the query medical image, the ReID model can retrieve sensitive patient information from a leaked database. (b) Our DeID framework, removing identity and then compensating medical semantics, ensures adjustable identity protection, while preserving downstream task utility. Besides, a Minimum Description Length (MDL) principle-based code space is introduced, to decouple and discard the identity information in medical features.

between identity and diagnostic features, these methods fail to preserve diagnostic semantics at high privacy levels adequately. Recently, diagnostic annotation-conditioned generative models [16, 27, 44, 88, 97] have yielded promising results, yet they remain limited to task-specific semantics and cannot offer adjustable privacy levels. In summary, *no existing method preserves task-generalizable semantics, while supporting a wide range of adjustable privacy levels*.

To address these issues, we introduce a novel divide-and-conquer framework DCM-DeID, which decouples identity removal from semantic preservation, to achieve semantic-rich yet adjustable de-identification. Our approach includes three steps, i.e., *ID-Blocking*, which masks identity-related regions to achieve adjustable privacy levels; *Medical Semantics Extraction*, which leverages pre-trained medical foundation models (MFMs) [71, 105] to extract semantic-rich medical features; *Image Re-Synthesis*, which employs a diffusion model [43, 83] to synthesize de-identified images, given the above ID-masked image and the medical features. Moreover, considering that the features from MFMs may also contain some identity information, we introduce a novel minimum description length [34]-based feature decoupling strategy, which excludes identity-associated information from the vanilla MFM features in a minimum-codelength latent space. This effectively prevents the reintroduction of identity information during the image re-synthesis step. Our contributions are:

- We reveal that existing medical DeID methods fall short in preserving task-generalizable semantics, and do not adjust seamlessly across privacy levels. We build the first benchmark for this problem, by reproducing previous approaches fairly on seven datasets.
- We propose the DCM-DeID framework, which performs identity removal and medical semantics preservation in separate steps, enabling both adjustable privacy protection and medical task utility.
- We introduce a Minimum Description Length-based decoupling strategy, which decouples identity cues from medical features in a compact code space, further improving the privacy protection capability.
- Our framework demonstrates state-of-the-art performance. Extensive Analysis is performed to verify its inner designs.

2. Related Works

Image Privacy Protection. Early methods applied low-level filters to obscure image details, including downsampling [21], blurring [91], and pixelation [42]. Later, encryption in alternate domains such as JPEG bitstreams [79, 89] and DCT coefficients [102, 103] was explored, though these often introduced severe distortions that hindered downstream tasks. Homomorphic encryption [98, 112] addresses inference on encrypted images, but suffers from high computational cost [75] and limited compatibility with advanced models like Vision Transformers [25]. Additionally, approaches for face images [12, 35, 66] leverage facial priors from StyleGAN [52] or face recognition networks [23, 106], which may not readily generalize to other domains, such as the medical-domain images in our work.

Medical Image De-Identification. Early methods (e.g., FreeSurfer [31], PyDeface [33], SynthStrip [45]) focus on removing facial features in brain MRI. For common medical images, early approaches use pixel-domain filters (like blurring [91] and pixelation [42]) or frequency-domain techniques [32], but these hand-crafted solutions also severely degrade the image details, leading to substantially degraded results. Differential Privacy methods [20, 28, 57, 100] inject noise into the training data, which compromises inference-time utility. More recent generative models [16, 27, 44, 88, 97] synthesize images conditioned on disease labels or lesion masks. However, they tend to lack task generalizability and struggle to balance privacy-utility trade-offs, which are addressed by our approach.

Feature Decoupling. Early variational auto-encoder (VAE)-based works [14, 41, 55, 85] decouple representations, by constraining the variables in latent space independent. Generative adversarial network (GAN)-based methods [15, 59] are unsupervised, leaving factors unaligned with explicit semantic or identity information. For face images, there are methods [24, 49, 52, 53, 62] targeting identity separation. However, these methods rely on strong facial priors that may not generalize to medical images. In contrast, our approach effectively decouples identity in medical images, within a minimum-codelength space.

3. Methodology

In this section, we first describe the medical re-identification (ReID) models used for privacy attacks. Next, we introduce our de-identification model, which divides the task into two stages. First, identity information is removed via region blocking with an adjustable threshold. Second, lost medical semantics are compensated. This approach flexibly adjusts privacy while preserving rich, generalizable medical features for downstream tasks.

3.1. Medical ReID Models

Given a query medical image, ReID models aim to retrieve all images belonging to the same individual, from a medical record database. Concretely, the model first extracts identity (ID) embedding from the query image, and then compare it with that of each image within the database. Then, the image with the closest Euclidean distance is adopted as the re-identified image.

We build two medical ReID models, i.e., ViT [25] and VisionMamba [110]-based ones, which are separately adopted in the training and the evaluation stages. These

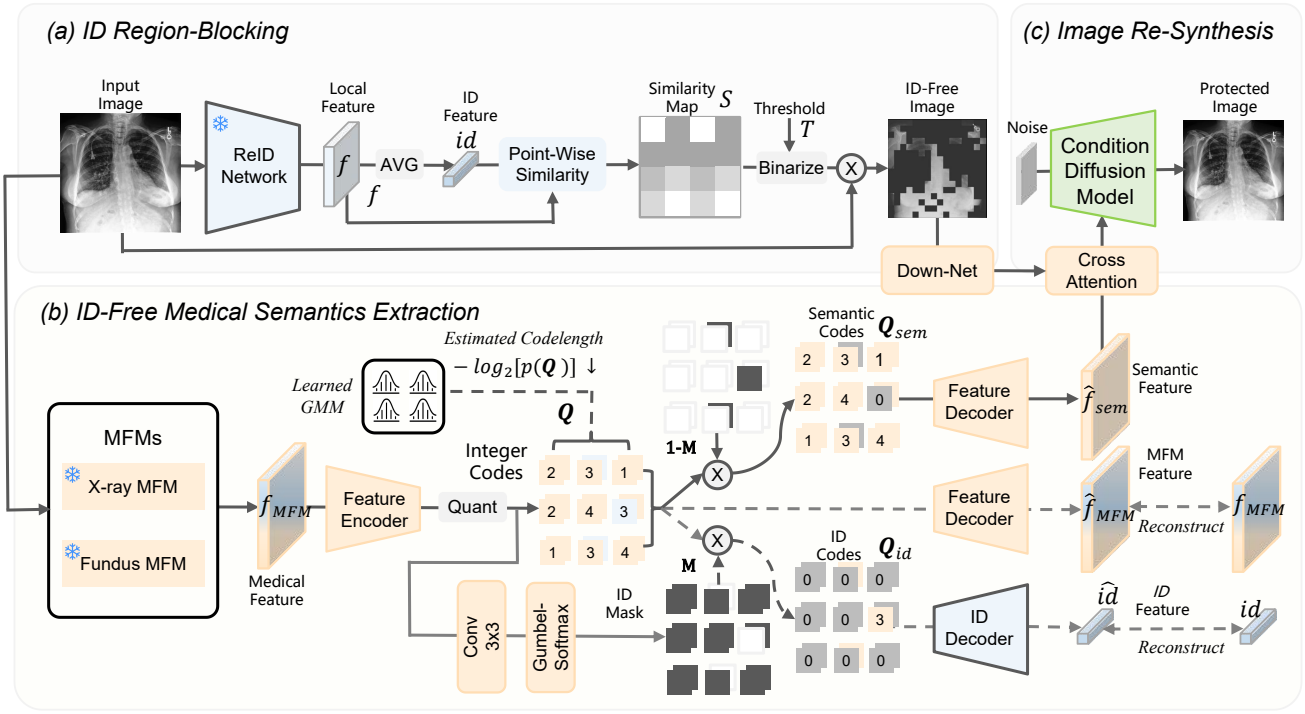

Figure 2. Overview of the proposed divide-and-conquer framework, DCM-DeID. (a) *ID-Blocking*: A pre-trained ReID network produces the identity-similarity map, which is binarized by different thresholds to adjust privacy level. (b) *ID-Free Medical Semantics Extraction*: Medical foundation models (MFMs) extract features that are encoded into a code space under the minimum-codelength regularization. A learned mask partitions the codes into identity- and medical semantics-related ones, where only the latter one is preserved. (c) *Image Re-Synthesis*: A diffusion model re-synthesizes images that are privacy-preserving and semantics-rich, generalizing to various downstream tasks. We illustrate with X-ray images, but the framework is also applicable to other modalities such as fundus images. ❄ denotes frozen models, while gray dashed lines indicate components used solely for learning identity-semantic decoupling. The channel number of codes Q is arbitrary; two channels are shown for conciseness. $\otimes$ denotes the element-wise multiplication.

ReID models are optimized with a combination of classification loss and triplet loss [40], following the previous object ReID work [39].

3.2. A Divide-and-Conquer Approach

To defend against attacks on medical ReID models, we propose DCM-DeID, a divide-and-conquer approach for medical image de-identification. DCM-DeID operates in three stages: *ID Blocking*, which removes identity-related image regions; *ID-Free Medical Semantics Extraction*, which extracts rich medical information without reintroducing identity information; and *Image Re-Synthesis*, which generates the final de-identified medical image.

ID-Blocking. Given an input image $X \in \mathbb{R}^{3\times H\times W}$, where H and W denote the image spatial scales, we use a ViT-based ReID model to extract local features $f \in \mathbb{R}^{768\times h\times w}$, where $h = H/16$ and $w = W/16$. Spatial average pooling is applied to f to obtain an identity embedding $id \in \mathbb{R}^{768}$. For each spatial position in f, the cosine similarity with id is computed, resulting in a similarity map $S \in \mathbb{R}^{h\times w}$. To generate the ID-blocked image, the similarity map S is binarized by a threshold T. Then, the ID-blocked image is computed as: $X^{noID} = X \odot \text{Upsample}(S > T)$, where Upsample denotes nearest-neighbor interpolation to match the resolution of S to X.

ID-Free Medical Semantics Extraction. Although X^{noID} effectively removes identity information, it inevitably distorts medical cues such as lung shadows. To amend this, we employ pre-trained medical foundation models (MFMs), e.g., MGCA [93] for X-ray images, to extract rich medical feature f_{MFM} from X. Since f_{MFM} contains both semantic cues and local details that may encode identity, we introduce a feature decoupling strategy (Section 3.3) to decouple and remove the identity information, yielding the identity-free semantic feature $\hat{f}_{sem}$.

Image Re-Synthesis. Given X^{noID} and $\hat{f}_{sem}$, a dual-conditioned diffusion model synthesizes the de-identified image that inherits the rich semantics within MFMs, while also protecting privacy. Since the synthesized image is highly realistic, it can be directly deployed to downstream medical AI applications, without further adaptation. The model details are elaborated in Section 3.4.

3.3. Medical Semantics Decoupling

Medical features extracted by the MFM encode both diagnostic semantics (e.g., lesion morphology) and identity-

related cues (e.g., rib patterns in chest X-rays). For effective privacy-preserving, it is imperative to decouple these two types of information, and discard the identity cues. We achieve this by learning a minimum-length code space, and separating the two parts in this space.

Theoretical Motivation. From an information-theoretic perspective, the Minimum Description Length (MDL) principle [4, 34] states that the best representation for a given set of data is the one that minimizes the total codelength needed to describe the data, where each group of features tends to capture the independent or low-correlation information parts. In our context, let $\boldsymbol{Q}$ be the latent representation of the MFM feature f_{MFM} and let $H(\boldsymbol{Q})$ denote its expected codelength. The MDL principle objective can be seen as balancing a reconstruction loss and a compression term, i.e., the so-called rate-distortion loss (RD loss) [5]:

$$\mathcal{L}_{\text{code-all}} = \min_{\mathcal{E},\mathcal{D}} \underbrace{\|f_{MFM} - \hat{f}_{MFM}\|_2}_{\text{Feature Reconstruction}} + \underbrace{\beta H(\boldsymbol{Q})}_{\text{Codelength}}, \quad (1)$$

where $\boldsymbol{Q} = \mathcal{E}(f_{MFM})$, $\hat{f}_{MFM} = \mathcal{D}(\boldsymbol{Q})$, and β denotes balancing weight. $\mathcal{E}$ and $\mathcal{D}$ represent a pair of feature encoder and decoder networks.

Discrete Code-based Codelength Estimation. Directly calculating the $H(\boldsymbol{Q})$ for the continuous variable $\boldsymbol{Q}$ is non-trivial [65]. Fortunately, the neural data compression community [2, 3, 67, 69] have verified that the codelength of *integer* latent variables can be quite precisely estimated with a *learnable entropy model*. Therefore, we append the quantization operation at the tail of the encoder $\mathcal{E}$, to make elements within $\boldsymbol{Q}$ discrete values, and estimate its codelength.

Concretely, $\mathcal{E}$ comprises three residual blocks [37] with 256 channels, followed by a convolutional layer to reduce dimensionality and a rounding operation that outputs a 32-channel integer code $\boldsymbol{Q}$. The decoder network $\mathcal{D}$ is symmetric to $\mathcal{E}$, except it omits the rounding operation. During training, the straight-through estimator [68] is employed to backpropagate gradients through the rounding step.

Following [2], the expected codelength of encoding $\boldsymbol{Q}$ is calculated as the log-likelihood, i.e., $H(\boldsymbol{Q}) = -\log_2 p(\boldsymbol{Q})$, where the probability $p(\boldsymbol{Q})$ is modeled using a Gaussian Mixture Model (GMM) [81] with K components:

$$p(\boldsymbol{Q}) = \sum_{k=1}^{K} w^k \cdot \mathcal{N}\left(\boldsymbol{Q}; \mu^k, e^{\sigma^k}\right), \quad (2)$$

where $\{\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}\}$ are the learnable mixture weights, means, and log variance scalers of the GMM components, respectively, which are shared across spatial positions, not unshared along the channel axis [2]. Following [17], K is set to three. For each integer element $q \in \boldsymbol{Q}$, the probability is computed over the quantization bin [19, 69]:

$$p(q) = \mathcal{F}(q + 0.5) - \mathcal{F}(q - 0.5), \quad (3)$$

where $\mathcal{F}(x) = \sum_{k=1}^{K} w^k \, \Phi\left(x; \mu^k, e^{\sigma^k}\right)$ is the cumulative distribution function (CDF) of the Gaussian Mixture Model (GMM), $\Phi\left(x; \mu, e^{\sigma}\right) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sqrt{2}\, e^{\sigma}}\right)\right]$. We not that the CDF can be efficiently calculated by the modern deep learning framework such as PyTorch [77].

Learning of Identity-Associated Code Mask. A single convolution layer predicts a binary mask $\mathbf{M}$ from $\boldsymbol{Q}$, with the same dimensions as $\boldsymbol{Q}$. The Gumbel-Softmax algorithm [48] is applied to enable gradient propagation through the binary mask. The identity-associated codes are then obtained by element-wise masking, $\boldsymbol{Q}_{\text{id}} = \boldsymbol{Q} \odot \mathbf{M}$. A lightweight convolutional network, composed of three residual blocks followed by average pooling, predicts the identity embedding $\hat{id}$ from $\boldsymbol{Q}_{\text{id}}$. Then, the RD loss for reconstructing identity can be given by:

$$\mathcal{L}_{\text{code-id}} = \|\hat{id} - id\|_2 + \beta H(\boldsymbol{Q}_{\text{id}}), \quad (4)$$

where $H(\tilde{\boldsymbol{Q}}_{\text{id}})$ is calculated similarly to $H(\boldsymbol{Q})$, sharing the same GMM parameters and balancing weight β as in Equation 1, since they operate in the same latent space.

Reconstruction of Medical Semantics. By suppressing identity-related codes via the inverse mask $(1-\mathbf{M})$, we obtain the semantics-part codes $\boldsymbol{Q}_{sem} = (1-\mathbf{M}) \otimes \boldsymbol{Q}$. Finally, the final ID-free medical semantic feature is reconstructed as: $\hat{f}_{sem} = \mathcal{D}(\boldsymbol{Q}_{sem})$, which preserves critical diagnostic semantics, excluding the identity information.

3.4. Image Re-Synthesis Model

Given the ID-masked image X^{noID} and the ID-free medical semantic feature $\hat{f}_{sem}$, we employ a diffusion model to synthesize de-identified medical images. First, we utilize a Down-Net to project the high-resolution X^{noID} into the low-resolution feature $f^{noID} \in \mathbb{R}^{512 \times \frac{H}{32} \times \frac{W}{32}}$. The Down-Net consists of the VAE encoder from Stable Diffusion [83], followed by two convolution layers of kernel size 5 and stride size 2. Next, we adopt a bi-directional cross-attention mechanism [13] to fuse f^{noID} and $\hat{f}_{sem}$, producing a fused feature $f_{fuse} \in \mathbb{R}^{512 \times \frac{H}{32} \times \frac{W}{32}}$, which is further processed through a series of convolutional layers. This produces a set of features with dimensions matching those of the UNet's intermediate feature maps within the diffusion model. These features are added to the UNet layers, guiding the diffusion process toward two objectives: maintaining the privacy level of X^{noID}, while preserving the medical semantics in $\hat{f}_{sem}$.

3.5. Learning Strategy

The whole framework is end-to-end optimized, with the following objective,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{code-all}} + \mathcal{L}_{\text{code-id}} + \mathcal{L}_{\text{diffuse}}, \quad (5)$$

where $\mathcal{L}_{\text{diffuse}}$ denotes the diffusion loss [43]. We do not introduce the balancing weight, since we found directly adding the loss terms already achieves satisfactory results.

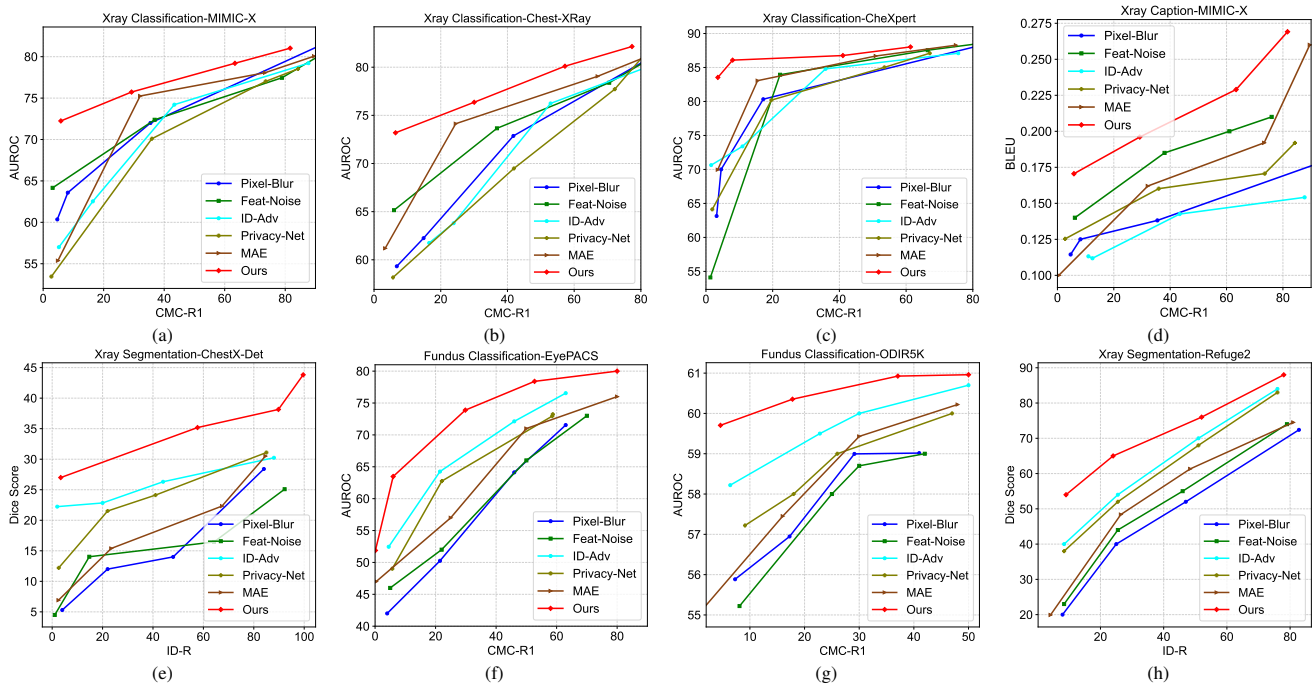

Figure 3. Identity-performance trade-off curves of various medical privacy protection methods.

Attack SR	Method	X-ray Classify AUROC (%)			X-ray Caption BLEU	X-ray Seg Dice (%)	Fundus Classify AUROC (%)		Fundus Seg Dice (%)
		MIMIC-X	Chest-XRay	CheXpert	MIMIC-X	ChestX-Det	EyePACS	ODIR5K	Refuge2
10%	Pixel-Blur [91]	64.15	60.48	74.45	0.1259	7.53	44.83	56.17	22.35
	Feat-Noise [104]	65.84	66.25	66.49	0.1453	10.74	47.76	55.53	24.85
	ID-Adv [74]	59.36	61.75	73.11	0.1131	22.23	56.28	58.50	41.23
	Privacy-Net [54]	57.09	59.51	71.47	0.1329	15.75	52.63	57.29	39.23
	MAE [38]	59.11	65.26	77.05	0.1186	9.94	50.91	56.50	27.33
	Ours	**72.86**	**73.66**	**86.12**	**0.1750**	**27.98**	**65.23**	**59.96**	**54.66**
20%	Pixel-Blur [91]	67.23	64.27	80.67	0.1307	11.25	49.56	57.42	34.11
	Feat-Noise [104]	68.27	69.01	80.78	0.1595	14.28	51.29	57.17	37.20
	ID-Adv [74]	64.08	62.71	77.53	0.1193	22.83	63.24	59.28	49.47
	Privacy-Net [54]	62.11	62.63	80.25	0.1434	20.55	61.01	58.25	47.47
	MAE [38]	66.43	71.38	83.52	0.1383	13.99	54.97	58.01	40.13
	Ours	**74.35**	**75.01**	**86.32**	**0.1859**	**29.49**	**69.59**	**60.41**	**62.04**
	Original	82.13	84.82	87.24	0.3218	52.89	81.46	61.53	90.08

Table 1. Performance comparison of medical image privacy protection methods, under different attack Success-Rates (SR). For measuring SR, we adopt CMC-R1 metric for MIMIC-X, Chest-Xray, CheXpert, EyePACS, and ODIR5K, using ID-R metric for ChestX-Det and Refuge2. Original denotes the performance on original images, which is the performance upper-bound of privacy-removal images.

4. Experiments

4.1. Implementation Details

For *Med-ReID* models, we adopt the AdamW optimizer [63] during training, with a learning rate of 1e-5 scheduled by cosine decaying strategy and a weight decay of 1e-2. The training process consists of 300,000 steps. The batch size is 256. We apply random cropping and blurring as image augmentation strategies, and the input image resolution to networks is 256×256. The ViT-based models are initialized with CLIP-pretrained weights [80], while the VisionMamba-based models are initialized with ImageNet-pretrained weights [22]. Training a single ReID model takes about 24 hours with four NVIDIA RTX 4090 GPUs.

For *DCM-DeID* model, the UNet within the diffusion model follows the same architecture as the Stable Diffusion [83], also performing the diffusion procedure in the latent space. The feature channels within UNet are reduced

to [128, 256, 512, 1024], for the four stages of both the down-pathway and up-pathway, to reduce computational cost. The identity-similarity map threshold T is defined as the r-th quantile of the similarity map S. r is selected from [0.95, 0.7, 0.4, 0.2] to cover wide privacy levels. We adopt MGCA-ResNet [93] and RetFound-ViT [109] MFMs for X-ray and fundus images, respectively. During training, we apply random flipping and random cropping 256 × 256 patches for data augmentation. The codelength loss term weight β is set to 0.5. At test time, we resize the shorter side of the images to 256 and then center-crop the middle 256×256 region. The learning rate is set to 1e-4 and is gradually decayed with the cosine annealing strategy [64]. The total number of training steps is 800,000. The mini-batch size is 64. We utilize the AdamW optimizer [63] implemented in PyTorch [77] with CUDA support. The momentum parameters are set as $\beta_1 = 0.9$ and $\beta_2 = 0.99$, and the gradient norm is clipped to a maximum value of 1. The entire training process takes about three days on a machine equipped with eight NVIDIA RTX 4090 GPUs.

4.2. Datasets

We evaluate our approach on two medical image modalities: chest X-rays and eye fundus photographs, with seven public datasets. For the *chest X-ray* modality, we split the MIMIC-X dataset [50] into training, validation, and test sets using an 8:1:1 ratio. For the Chest-Xray and CheXpert datasets, we randomly select 10% patients as the test set. We also adopt the ChestX-Det dataset [60] to evaluate the X-ray segmentation task. For the *eye fundus* modality, we divide the EyePACS dataset [26] into training, validation, and test sets with an 8:1:1 ratio, and we use the Refuge2 dataset [30] to evaluate the fundus segmentation task. ODIR5K [8] is also adopted for evaluating the fundus classification task of systemic diseases such as hypertension. Note that only MIMIC-X and EyePACS are used during training; all other datasets, which differ in environment, demographics, and imaging devices, are never seen during training, to evaluate the domain generalizability of our approach.

4.3. Reproduced Privacy Protection Methods

We implement several privacy protection methods, comparing them with our approach in a fair setting.

Pixel-Blur [91]. This method applies a Gaussian blur to the input image. We experiment with standard deviations of $\{1, 5, 10, 20\}$ to vary the level of de-identification.

Feat-Noise [104]. We train an autoencoder [104] and inject Gaussian noise into its latent features. The noise level is selected from $\{0.1, 0.8, 0.85, 0.9, 0.95\}$.

ID-Adv [74]. A UNet is trained to generate a de-identified image Y from the original image X, optimizing the loss $\mathcal{L} = \lambda \cos(id_X, id_Y) + \|med_X - med_Y\|_2 + \mathcal{L}_{reg}$, where id_X and id_Y are identity features extracted by a ViT-based ReID model, and med_X and med_Y are medical features obtained from MFMs same as our approach. $\mathcal{L}_{reg}$ is a GAN regularization loss ensuring visual plausibility, $\cos(\cdot, \cdot)$ denotes cosine similarity, and $\|\cdot\|_2$ the ℓ_2 norm. The trade-off weight λ is chosen from $\{0.1, 0.5, 1, 2\}$.

Privacy-Net [54]. This method updates the identity model and the de-identification network adversarially, enhancing de-identification performance. The original Privacy-Net focuses solely on segmentation tasks, supervised by segmentation masks. To enable task-agnostic de-identification, we train it using the same objective as ID-Adv. Since the identity model is adversarially updated and are stronger, we use smaller λ values compared to ID-Adv, i.e., $\{0.05, 0.25, 0.5, 1\}$.

MAE [38]. Following [96], we transfer the concept of masked auto-encoders (MAE) [38] to the adjustable privacy protection problem, by masking a random proportion of patches to obscure identity information. It adopts the same diffusion model as our approach to generate the masked regions. This model can also serve as a degenerated version of our model, where both semantic compensation and identity-region similarity designs are removed.

4.4. Downstream Task Models

For the *identity recognition*, we adopt the VisionMamba-based ReID model, which differs from the ViT-based model employed during the training of privacy protection methods, ensuring the method generalization capability across different ReID models. For the *X-ray classification*, we use the ViT model pre-trained with Med-UniC [92]. For *X-ray captioning*, we employ the visual-language model CXR-LLaVA-v2 [58], which is specifically designed for X-ray images. For *X-ray segmentation*, we adopt CGRSeg [72]. For *fundus classification*, we use the ViT model pre-trained with KeepFit [99]. Finally, for *fundus segmentation*, given the limited dataset size, we employ nnUNet [47].

4.5. Evaluation Metrics

For privacy evaluation, we adopt the cumulative matching characteristics (CMC) [10] at Rank-1, i.e., CMC-R1, on datasets with patient ID information available (i.e., MIMIC-X, Chest-Xray, CheXpert, EyePacs, and ODIR5K). For datasets without patient ID information (i.e., CheX-det and REFUGE2), we adopt the recognition rate, i.e., ID-R, which determines whether the distance between the ID feature of the original and de-identified image exceeds a predefined threshold. The thresholds are set to 1.1 and 1.35 for the X-ray and fundus modalities, respectively, based on statistics from the validation sets of MIMIC-X and EyePACS. For the disease diagnosis task, we employ the area under the receiver operating characteristic curve (AUROC) metric [11]; for the image captioning task, we use the bilingual evaluation understudy (BLEU) metric [76]; and for the image

segmentation task, we adopt the Dice score metric [7].

4.6. Results

X-ray Classification. As shown in Table 1, our method substantially outperforms other approaches, achieving AUROCs of 72.86%, 73.66%, and 86.12% on MIMIC-X, Chest-XRay, and Chexpert, respectively, under CMC-R1=10%. Notably, although our model is trained on MIMIC-X, it generalizes well to the other two datasets.

Among the other compared approaches, Feat-Noise obtains the second-best performance, i.e., AUROC of 65.84% at CMC-R1=10% on MIMIX-X, by condensing image pixels into a compact latent feature space. In contrast, methods that jointly optimize a trade-off between de-identification and medical preservation, i.e., ID-Adv and PrivacyNet, yield unsatisfactory performances. As shown in Figure 3 (a), under the ID-R=5% setting, ID-Adv and PrivacyNet attain AUROCs of only 56.32% and 54.21% on MIMIC-X, respectively, which are much lower than the simple pixel blurring baseline (61.62%). This indicates that directly optimizing the two conflicting objectives is suboptimal. In contrast, our approach decouples the objectives into two separate steps, identity removal and medical semantic compensation, achieving consistently superior performance.

As for MAE, which employs the same diffusion model as ours, it achieves competitive results at a high attacking rate, with an AUROC of 76.12% @CMC-R1=40% on MIMIC-X, outperforming all other approaches except ours. However, at a low attacking rate CMC-R1=10%, it falls behind our method by over 13% AUROC. This highlights that our superior performance is not solely due to the generative power of the diffusion model, but rather stems from the effectiveness of our core idea of semantic compensation.

X-ray Caption. As shown in Table 1, our method attains a BLEU score of 0.1750, remarkably surpassing Pixel-Blur (0.1259), Feat-Noise (0.1453), and MAE (0.1186), at CMC-R1=10%. This proves that our approach can comprehensively preserve the clinic-required information, beyond only the classification label.

X-ray Segmentation. Furthermore, we evaluate the methods on a fine-grained task: segmentation. As shown in Table 1, at ID-R1=10%, our method achieves a Dice score of 27.98%, outperforming Pixel-Blur (7.53%), Feat-Noise (10.74%), Privacy-Net (15.75%), ID-Adv (22.23%), and MAE(9.94%). This proves that our semantic compensation scheme not only preserves the global semantics for classification, but also effectively retains the local semantics for segmentation. Pixel-Blur and Feat-Noise perform poorly, since they severely corrupts the image details. In contrast, ID-Adv and Privacy-Net, which incorporate a medical feature-matching loss, achieve slightly decent performance, but still lag far behind our approach. For instance, under ID-R=80%, our method outperforms Privacy-Net by approximately 8%, as shown in Figure 3(e).

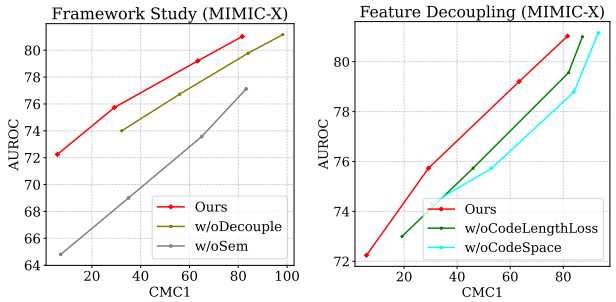

Figure 4. (*Left*) Ablation on the framework design. (*Right*) Ablation study on the feature decoupling strategy.

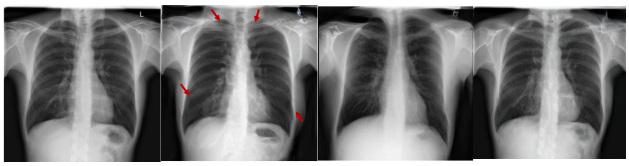

Figure 5. Qualitative comparison of different variant models. The models are described in Figure 4 caption. Red arrow denotes the modified identity-related features. Best to view by zooming-in.

Fundus Classification. Beyond X-ray images, our method also proves effective on fundus data. For instance, on EyePACS and ORID5K, our approach outperforms the second-best competitor ID-Adv by about 9% and 1%, respectively. These results confirm that our approach generalizes well across different imaging modalities.

Fundus Segmentation. Our method achieves a Dice score of 54.66% on REFUGE2 at ID-R1=10% , largely surpassing MAE (27.33%), Privacy-Net (39.23%), and ID-Adv (41.23%). This further validates that our method also effectively preserves fine-grained semantic cues of eye fundus.

4.7. Model Analysis

Framework-Level Ablation Study. As shown in Figure 4 (*Left*), by removing the semantic branch, the AUROC of the resulted model 'w/oSem' dramatically drops by over 8%, at CMC-R1=5%. On the other hand, without the identity-semantics decoupling mechanism, the resulting model 'w/oDecouple' leads to about 15% CMC-R1 increase, for achieving the similar AUROC performance, since substantial identity cues are leaked from the vanilla medical features of MFMs. We further illustrate the protected images from different models. As shown in Figure 5, our results effectively modify identity-related features, such as the shape and location of the clavicle and chest contour. The 'w/oSem' model also removes these regions but significantly alters medical manifestations. In contrast, the 'w/oDecouple' model preserves medical features but fails to sufficiently suppress identity-related features, such as clavicle shape, due to residual identity information in the features from MFMs. These results confirm that both medical

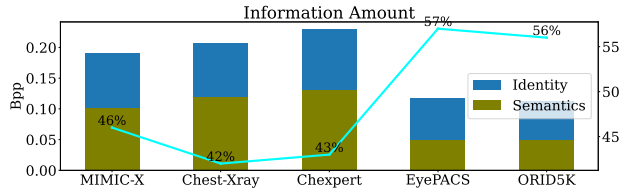

Figure 6. Comparison of semantic and identity information in terms of Bits-per-Pixel (bpp) [61], calculated as the feature codelength divided by the original image size.

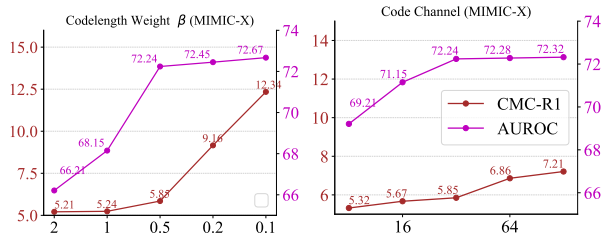

Figure 7. (*Left*) Impact of the rate-distortion weight β. (*Right*:) Impact of the code dimension. All experiments are evaluated by masking 95% identity-related regions, for a fair comparison.

semantics and identity-semantics decoupling are essential for our advanced medical DeID approach.

Ablation Study on the Decoupling Strategy. As shown in Figure 4 (*Right*), omitting the codelength loss terms ('w/oCodeLengthLoss') fails to effectively remove identity information from MFM features, since the loose space cannot effectively decouple the identity and the semantics information. Moreover, removing the discrete code bottleneck ('w/oCodeSpace') further exacerbates identity leakage, leading to further inferior performance.

Furthermore, we quantitatively compare the overall and identity-related information in MFM features, as shown in Figure 6. First, we notice that a significant portion is identity-related, i.e., around 44% and 55% for X-ray and fundus images. Second, the average information amount of the X-ray dataset Chest-Xray is 0.23bpp, much higher than 0.11bpp achieved by the fundus dataset EyePACS. This is aligned with the medical knowledge prior, that X-rays capture multiple organs and tissues, containing much complex information, than the fundus image that only focuses on eyes. This proves that the learned codelength effectively describes the medical data characteristics.

Finally, we analyze the impact of the codelength loss weight β and the latent code channel number. As shown in Figure 7 (*Left*), reducing β from 0.5 to 0.1 significantly increases CMC-R1 from 5.85% to 12.34%, as a loosely constrained code space fails to effectively decouple identity information. Conversely, increasing β from 0.5 to 2 has little effect on CMC-R1 but reduces AUROC performance by approximately 6%, as an overly strong constraint impairs semantic feature reconstruction. The number of code channels also influences performance, by tuning the information capacity of the latent code, as shown in Figure 7 (*Right*). However, since β directly regulates the code-length term, the impact of the channel number is limited.

Discussion with Label-Conditioned Diffusion Models. These methods [46, 95, 108] employ task-specific labels (e.g., disease labels or text reports) to synthesize images, which are limited to label-associated tasks. In contrast, our approach is task-agnostic and applicable to diverse tasks. Moreover, after fine-tuning our approach towards a single task, i.e., replacing the MFM with a supervised classification network, our method achieves 81.92% AUROC at CMC-R1 = 0.30%, surpassing the label-conditioned model, i.e., 80.79% AUROC at CMC-R1 = 0.29%. This confirms that our minimum-codelength representation also benefits the single-task setting, compared to the methods directly using the task labels guiding the diffusion procedure.

Model Complexity. All methods and our model comprise about 380M parameters, for a fair comparison. Our inference time is 540 ms on an NVIDIA 4090 GPU, which is similar to MAE (526ms), but slower than Privacy-Net (120 ms), ID-Adv (122 ms), and Feat-Noise (124 ms), due to the multiple inference steps of diffusion procedure. Nonetheless, given the significant performance gains and that the medical imaging procedure itself is time-consuming, the running time is acceptable and does not hinder clinical workflows. In the future, we will integrate the single-step diffusion technique [101] to accelerate the process.

5. Conclusion, Future Works, and Other

Conclusion. We have presented DCM-DeID, a divide-and-conquer framework for medical image de-identification. By leveraging pre-trained Medical Foundation Models and a minimum codelength-based feature decoupling strategy, our method effectively remove identity cues, while preserving medical task utility. Extensive evaluations demonstrate the superiority of our approach. **Future Works.** Although our study extensively examines the medical privacy protection problem on large-scale public datasets with patient identity annotations, these datasets consist solely of single-slice images. In the future, we will extend our approach to multi-slice images, such as those produced by Magnetic Resonance Imaging (MRI). **Broader Impacts.** Our DeID technique is designed for medical AI applications, aiding the human. We emphasize that all rigorous clinical decisions must be made by human physicians using the original medical images. Furthermore, it is critical to enforce strict ethical guidelines, working in synergy with technological approaches to achieve medical privacy protection.

References

[1] Kadek YE Aryanto, André Broekema, Matthijs Oudkerk, and Peter MA van Ooijen. Implementation of an anonymisation tool for clinical trials using a clinical trial processor integrated with an existing trial patient data information system. *European radiology*, 22:144–151, 2012. 1

[2] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*, 2016. 4

[3] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*, 2018. 4

[4] Andrew Barron, Jorma Rissanen, and Bin Yu. The minimum description length principle in coding and modeling. *IEEE transactions on information theory*, 44(6):2743–2760, 1998. 4

[5] Toby Berger. Rate-distortion theory. *Wiley Encyclopedia of Telecommunications*, 2003. 4

[6] Rui Bernardes, Pedro Serranho, and Conceição Lobo. Digital ocular fundus imaging: a review. *Ophthalmologica*, 226 (4):161–181, 2011. 1

[7] Jeroen Bertels, Tom Eelbode, Maxim Berman, Dirk Vandermeulen, Frederik Maes, Raf Bisschops, and Matthew B Blaschko. Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 92–100. Springer, 2019. 7

[8] Amit Bhati, Neha Gour, Pritee Khanna, and Aparajita Ojha. Discriminative kernel convolution network for multi-label ophthalmic disease detection on imbalanced fundus image dataset. *Computers in Biology and Medicine*, 153:106519, 2023. 6

[9] Vincent Bindschaedler, Paul Grubbs, David Cash, Thomas Ristenpart, and Vitaly Shmatikov. The tao of inference in privacy-protected databases. *Cryptology ePrint Archive*, 2017. 1

[10] 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. 6

[11] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997. 6

[12] Jingyi Cao, Bo Liu, Yunqian Wen, Rong Xie, and Li Song. Personalized and invertible face de-identification by disentangled identity information manipulation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3334–3342, 2021. 1, 2

[13] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021. 4

[14] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018. 2

[15] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016. 2

[16] Yan Chen and Pouyan Esmaeilzadeh. Generative ai in medical practice: in-depth exploration of privacy and security challenges. *Journal of Medical Internet Research*, 26:e53008, 2024. 2

[17] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7939–7948, 2020. 4

[18] I Glenn Cohen and Michelle M Mello. Big data, big tech, and protecting patient privacy. *Jama*, 322(12):1141–1142, 2019. 1

[19] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999. 4

[20] William L Croft, Jörg-Rüdiger Sack, and Wei Shi. Obfuscation of images via differential privacy: From facial images to general images. *Peer-to-Peer Networking and Applications*, 14:1705–1733, 2021. 1, 2

[21] Ji Dai, Behrouz Saghafi, Jonathan Wu, Janusz Konrad, and Prakash Ishwar. Towards privacy-preserving recognition of human activities. In *2015 IEEE international conference on image processing (ICIP)*, pages 4238–4242. IEEE, 2015. 2

[22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[23] 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

[24] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5154–5163, 2020. 2

[25] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2

[26] Emma Dugas, Jared, Jorge, and Will Cukierski. Diabetic retinopathy detection. https://kaggle.com/competitions/diabetic-retinopathy-detection, 2015. Kaggle. 6

[27] August DuMont Schütte, Jürgen Hetzel, Sergios Gatidis, Tobias Hepp, Benedikt Dietz, Stefan Bauer, and Patrick

Schwab. Overcoming barriers to data sharing with medical image generation: a comprehensive evaluation. *NPJ digital medicine*, 4(1):141, 2021. 2

[28] Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006. 1, 2

[29] Weijie Fan, Yi Yang, Jing Qi, Qichuan Zhang, Cuiwei Liao, Li Wen, Shuang Wang, Guangxian Wang, Yu Xia, Qihua Wu, et al. A deep-learning-based framework for identifying and localizing multiple abnormalities and assessing cardiomegaly in chest x-ray. *Nature Communications*, 15(1): 1347, 2024. 1

[30] Huihui Fang, Fei Li, Junde Wu, Huazhu Fu, Xu Sun, Jaemin Son, Shuang Yu, Menglu Zhang, Chenglang Yuan, Cheng Bian, et al. Refuge2 challenge: A treasure trove for multi-dimension analysis and evaluation in glaucoma screening. *arXiv preprint arXiv:2202.08994*, 2022. 6

[31] Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012. 1, 2

[32] Alex Gaudio, Asim Smailagic, Christos Faloutsos, Shreshta Mohan, Elvin Johnson, Yuhao Liu, Pedro Costa, and Aurélio Campilho. Deepfixcx: Explainable privacy-preserving image compression for medical image analysis. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 13(4):e1495, 2023. 2

[33] Satrajit Ghosh, Chris Gorgolewski, et al. pydeface: A tool to remove facial features from mri images, 2012. 1, 2

[34] Peter D Grünwald. *The minimum description length principle*. MIT press, 2007. 2, 4

[35] Xiuye Gu, Weixin Luo, Michael S Ryoo, and Yong Jae Lee. Password-conditioned anonymization and deanonymization with face identity transformers. In *European conference on computer vision*, pages 727–743. Springer, 2020. 2

[36] A Gueziec and P Kazanzides. Anatomy-based registration of ct-scan and intraoperative x-ray images for guiding a surgical robot. *IEEE Transactions on Medical Imaging*, 17(5): 715–728, 1998. 1

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[38] 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

[39] 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. 3

[40] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 3

[41] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017. 2

[42] Steven Hill, Zhimin Zhou, Lawrence Saul, and Hovav Shacham. On the (in) effectiveness of mosaicing and blurring as tools for document redaction. *Proceedings on Privacy Enhancing Technologies*, 2016. 1, 2

[43] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 4

[44] Sungmin Hong, Razvan Marinescu, Adrian V Dalca, Anna K Bonkhoff, Martin Bretzner, Natalia S Rost, and Polina Golland. 3d-stylegan: A style-based generative adversarial network for generative modeling of three-dimensional medical images. In *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections: First Workshop, DGM4MICCAI 2021, and First Workshop, DALI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 1*, pages 24–34. Springer, 2021. 2

[45] Andrew Hoopes, Jocelyn S Mora, Adrian V Dalca, Bruce Fischl, and Malte Hoffmann. Synthstrip: skull-stripping for any brain image. *NeuroImage*, 260:119474, 2022. 1, 2

[46] Peng Huang, Xue Gao, Lihong Huang, Jing Jiao, Xiaokang Li, Yuanyuan Wang, and Yi Guo. Chest-diffusion: a light-weight text-to-image model for report-to-cxr generation. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024. 8

[47] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 6

[48] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 4

[49] Zi-Hang Jiang, Qianyi Wu, Keyu Chen, and Juyong Zhang. Disentangled representation learning for 3d face shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11957–11966, 2019. 2

[50] 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. 6

[51] Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311, 2020. 1

[52] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[53] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition, pages 8110–8119, 2020. 2

[54] 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, 5, 6

[55] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International conference on machine learning*, pages 2649–2658. PMLR, 2018. 2

[56] Cody A Koch and Wayne F Larrabee. Patient privacy, photographs, and publication. *JAMA facial plastic surgery*, 15 (5):335–336, 2013. 1

[57] Abhinav Kumar, Sanjay Kumar Singh, K Lakshmanan, Sonal Saxena, and Sameer Shrivastava. A novel cloud-assisted secure deep feature classification framework for cancer histopathology images. *ACM Transactions on Internet Technology (TOIT)*, 21(2):1–22, 2021. 1, 2

[58] Seowoo Lee, Jiwon Youn, Hyungjin Kim, Mansu Kim, and Soon Ho Yoon. Cxr-llava: a multimodal large language model for interpreting chest x-ray images. *European Radiology*, pages 1–13, 2025. 6

[59] Yuheng Li, Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee. Mixnmatch: Multifactor disentanglement and encoding for conditional image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8039–8048, 2020. 2

[60] Jie Lian, Jingyu Liu, Shu Zhang, Kai Gao, Xiaoqing Liu, Dingwen Zhang, and Yizhou Yu. A structure-aware relation network for thoracic diseases detection and segmentation. *IEEE Transactions on Medical Imaging*, 40(8):2042–2052, 2021. 6

[61] Weisi Lin and Li Dong. Adaptive downsampling to improve image compression at low bit rates. *IEEE Transactions on Image Processing*, 15(9):2513–2521, 2006. 8

[62] Yu Liu, Fangyin Wei, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang. Exploring disentangled feature representation beyond face identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2080–2089, 2018. 2

[63] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5, 6

[64] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[65] Charles Marsh. Introduction to continuous entropy. *Department of Computer Science, Princeton University*, 1034, 2013. 4

[66] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5447–5456, 2020. 2

[67] Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, and Luc Van Gool. Conditional probability models for deep image compression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4394–4402, 2018. 4

[68] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*, 2023. 4

[69] David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. *Advances in neural information processing systems*, 31, 2018. 4

[70] Eriksson Monteiro, Carlos Costa, and José Luís Oliveira. A de-identification pipeline for ultrasound medical images in dicom format. *Journal of Medical Systems*, 41(5):89, 2017. 1

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

[72] Zhenliang Ni, Xinghao Chen, Yingjie Zhai, Yehui Tang, and Yunhe Wang. Context-guided spatial feature reconstruction for efficient semantic segmentation. In *European Conference on Computer Vision*, pages 239–255. Springer, 2024. 6

[73] 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. 1

[74] Kai Packhäuser, Sebastian Gündel, Florian Thamm, Felix Denzinger, and Andreas Maier. Deep learning-based anonymization of chest radiographs: a utility-preserving measure for patient privacy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 262–272. Springer, 2023. 1, 5, 6

[75] Pascal Paillier. Public-key cryptosystems based on composite degree residuosity classes. In *International conference on the theory and applications of cryptographic techniques*, pages 223–238. Springer, 1999. 2

[76] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6

[77] 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. 4, 6

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

[79] Moo-Ryong Ra, Ramesh Govindan, and Antonio Ortega. P3: Toward {Privacy-Preserving} photo sharing. In *10th USENIX Symposium on Networked Systems Design and Implementation (NSDI 13)*, pages 515–528, 2013. 2

[80] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learn-

ing transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5

[81] Douglas A Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663):3, 2009. 4

[82] David Rodríguez González, Trevor Carpenter, Jano I van Hemert, and Joanna Wardlaw. An open source toolkit for medical imaging de-identification. *European radiology*, 20: 1896–1904, 2010. 1

[83] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 4, 5

[84] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12):2176–2182, 2021. 1

[85] Huajie Shao, Yifei Yang, Haohong Lin, Longzhong Lin, Yizhuo Chen, Qinmin Yang, and Han Zhao. Rethinking controllable variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19250–19259, 2022. 2

[86] Julie K. Taitsman, Christi Macrina Grimm, and Shantanu Agrawal. Protecting patient privacy and data security. *New England Journal of Medicine*, 368(11):977–979, 2013. 1

[87] Adam Tanner. *Our bodies, our data: how companies make billions selling our medical records*. Beacon Press, 2017. 1

[88] Huan Tian, Tianqing Zhu, and Wanlei Zhou. Fairness and privacy preservation for facial images: Gan-based methods. *Computers & Security*, 122:102902, 2022. 2

[89] Matt Tierney, Ian Spiro, Christoph Bregler, and Lakshminarayanan Subramanian. Cryptagram: Photo privacy for online social media. In *Proceedings of the first ACM conference on Online social networks*, pages 75–88, 2013. 2

[90] Gary Kin-wai Tsui and Tao Chan. Automatic selective removal of embedded patient information from image content of dicom files. *American Journal of Roentgenology*, 198(4): 769–772, 2012. 1

[91] Nishant Vishwamitra, Bart Knijnenburg, Hongxin Hu, Yifang P Kelly Caine, et al. Blur vs. block: Investigating the effectiveness of privacy-enhancing obfuscation for images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 39–47, 2017. 1, 2, 5, 6

[92] 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. 6

[93] 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. 3, 6

[94] Guangyu Wang, Xiaohong Liu, Jun Shen, Chengdi Wang, Zhihuan Li, Linsen Ye, Xingwang Wu, Ting Chen, Kai Wang, Xuan Zhang, et al. A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and covid-19 pneumonia from chest x-ray images. *Nature biomedical engineering*, 5(6):509–521, 2021. 1

[95] Jinzhuo Wang, Kai Wang, Yunfang Yu, Yuxing Lu, Wenchao Xiao, Zhuo Sun, Fei Liu, Zixing Zou, Yuanxu Gao, Lei Yang, et al. Self-improving generative foundation model for synthetic medical image generation and clinical applications. *Nature Medicine*, pages 1–9, 2024. 8

[96] Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Zhu, Jiankang Deng, Xinchao Wang, Hakan Bilen, and Yang You. Facemae: Privacy-preserving face recognition via masked autoencoders. *arXiv preprint arXiv:2205.11090*, 2022. 1, 6

[97] Shudong Wang, Zhiyuan Zhao, Yawu Zhao, Luqi Wang, Yuanyuan Zhang, Jiehuan Wang, Sibo Qiao, and Zhihan Lyu. A semantic conditional diffusion model for enhanced personal privacy preservation in medical images. *IEEE Journal of Biomedical and Health Informatics*, 2024. 2

[98] Weiru Wang, Chi-Man Vong, Yilong Yang, and Pak-Kin Wong. Encrypted image classification based on multilayer extreme learning machine. *Multidimensional Systems and Signal Processing*, 28(3):851–865, 2017. 2

[99] Ruiqi Wu, Chenran Zhang, Jianle Zhang, Yi Zhou, Tao Zhou, and Huazhu Fu. Mm-retinal: Knowledge-enhanced foundational pretraining with fundus image-text expertise. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 722–732. Springer, 2024. 6

[100] Hanyu Xue, Bo Liu, Ming Ding, Tianqing Zhu, Dayong Ye, Li Song, and Wanlei Zhou. Dp-image: Differential privacy for image data in feature space. *arXiv preprint arXiv:2103.07073*, 2021. 1, 2

[101] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6613–6623, 2024. 8

[102] Lin Yuan, Pavel Korshunov, and Touradj Ebrahimi. Privacy-preserving photo sharing based on a secure jpeg. In *2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 185–190. IEEE, 2015. 2

[103] Lin Yuan, Pavel Korshunov, and Touradj Ebrahimi. Secure jpeg scrambling enabling privacy in photo sharing. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2015. 2

[104] Junhai Zhai, Sufang Zhang, Junfen Chen, and Qiang He. Autoencoder and its various variants. In *2018 IEEE international conference on systems, man, and cybernetics (SMC)*, pages 415–419. IEEE, 2018. 5, 6

[105] Shaoting Zhang and Dimitris Metaxas. On the challenges and perspectives of foundation models for medical image analysis. *Medical image analysis*, 91:102996, 2024. 2

[106] Xiao Zhang, Rui Zhao, Yu Qiao, Xiaogang Wang, and Hongsheng Li. Adacos: Adaptively scaling cosine logits for effectively learning deep face representations. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10823–10832, 2019. 2

[107] Hong-Yu Zhou, Xiaoyu Chen, Yinghao Zhang, Ruibang Luo, Liansheng Wang, and Yizhou Yu. Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. *Nature Machine Intelligence*, 4(1):32–40, 2022. 1

[108] Xinrui Zhou, Yuhao Huang, Haoran Dou, Shijing Chen, Ao Chang, Jia Liu, Weiran Long, Jian Zheng, Erjiao Xu, Jie Ren, et al. Ctrl-genaug: Controllable generative augmentation for medical sequence classification. *arXiv preprint arXiv:2409.17091*, 2024. 8

[109] 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. 1, 6

[110] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st International Conference on Machine Learning*, pages 62429–62442. PMLR, 2024. 2

[111] Yingxuan Zhu, PD Singh, Khan Siddiqui, and Michael Gillam. An automatic system to detect and extract texts in medical images for de-identification. In *Medical Imaging 2010: Advanced PACS-based Imaging Informatics and Therapeutic Applications*, page 762803. SPIE, 2010. 1

[112] M Tarek Ibn Ziad, Amr Alanwar, Moustafa Alzantot, and Mani Srivastava. Cryptoimg: Privacy preserving processing over encrypted images. In *2016 IEEE Conference on Communications and Network Security (CNS)*, pages 570–575. IEEE, 2016. 2