Medical Manifestation-Aware De-Identification

Yuan Tian[1], **Shuo Wang**[2], **Guangtao Zhai**[2*]

[1] Shanghai AI Laboratory
[2] Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University
tianyuan168326@outlook.com, zhaiguangtao@sjtu.edu.cn

Abstract

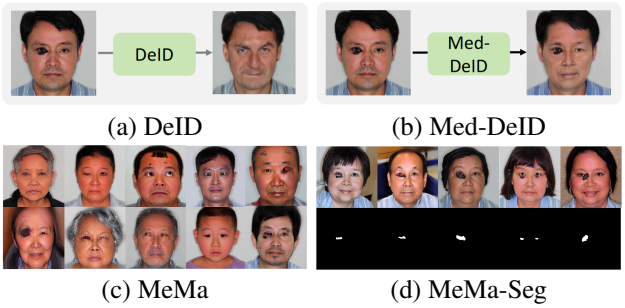
Face de-identification (DeID) has been widely studied for common scenes, but remains under-researched for medical scenes, mostly due to the lack of large-scale patient face datasets. In this paper, we release MeMa, consisting of over 40,000 photo-realistic patient faces. MeMa is re-generated from massive real patient photos. By carefully modulating the generation and data-filtering procedures, MeMa avoids breaching real patient privacy, while ensuring rich and plausible medical manifestations. We recruit expert clinicians to annotate MeMa with both coarse- and fine-grained labels, building the first medical-scene DeID benchmark. Additionally, we propose a baseline approach for this new medical-aware DeID task, by integrating data-driven medical semantic priors into the DeID procedure. Despite its conciseness and simplicity, our approach substantially outperforms previous ones.

Dataset and Code —
https://github.com/tianyuan168326/MeMa-Pytorch

Introduction

The public sharing of large-scale image datasets has facilitated the rapid progress in Artificial Intelligence (AI). However, this also poses great privacy concerns, especially for facial images, which are widely used for identity authentication. To address this issue, many de-identification (DeID) algorithms (Cao et al. 2021; Maximov, Elezi, and Leal-Taixé 2020; Gu et al. 2020; Li et al. 2023; Cai et al. 2024) have been continuously proposed for protecting the facial identity, achieving promising results on common-scene facial datasets (Karras, Laine, and Aila 2019; Karras et al. 2017).

However, rare researches are conducted for the medical scenes, although patient privacy leakage is a big concern in the medical AI era (Price and Cohen 2019). Research on medical-aware DeID (Med-DeID) mainly faces two obstacles. *First*, there are few medical-scene facial datasets available, due to the difficulty in accessing patients compared to healthy individuals. Moreover, it is often not acceptable to package real patient faces as datasets and make them publicly downloadable. *Second*, the current DeID approaches may not be appropriate for protecting medical facial images, due to not particularly preserving the disease manifestations of the origin image. This leads to the lost of diagnosis-necessary disease signs, deteriorating the medical utilities.



(a) DeID (b) Med-DeID

(c) MeMa (d) MeMa-Seg

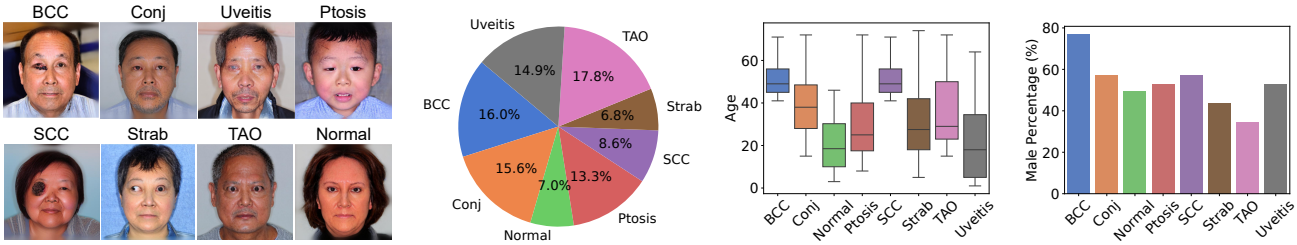
Figure 1: (a) Common DeID approaches, focus on removing identity. (b) Our medical-aware DeID (Med-DeID), also considers preserving the diagnosis-necessary medical information. (c) Our MeMa, a large-scale patient face dataset. (d) Our MeMa-Seg, the tumor segmentation subset of MeMa.

In this paper, we release a Medical Manifestation-rich patient face dataset, termed **MeMa**, containing over 40,000 photo-realistic virtual patient images. To construct MeMa, we obtained permission from the hospital's medical ethics committee to photograph patients. Then, these patient photos are annotated by expert physicians, before being used to train a specialized generative model. By carefully modulating the sampling procedure of the generative model and filtering the generated data, we created a diverse, high-quality, and real-world-like patient face dataset.

Furthermore, we propose a baseline medical semantics-preserved DeID approach, termed MedSem-DeID, to eliminate the patient identity from the facial image, at the premise of preserving the medical utility. Concretely, we first condense the rich medical priors within the MeMa into a medical semantics encoder, and then adopt it to (1) enhance the medical knowledge of the features within the DeID pipeline, and (2) minimize the medical-aware distortion of the de-identified images. Despite its simplicity, our approach easily outperforms previous DeID approaches for medical scenes, thanks to the rich medical manifestation knowledge embodied in the MeMa dataset. Our main contributions are:

- We release, to the best of our knowledge, the first large-

*Corresponding Author.



(a) Disease Example (b) Patient distribution (c) Age distribution (d) Gender distribution

Figure 2: Examples and the distribution characteristics of the proposed MeMa dataset.

scale patient face dataset of rich medical manifestations, **MeMa**, which is expected to facilitate research in the field of medical-scene privacy protection.

- We propose a baseline approach for this novel medical DeID problem, which particularly preserves the disease signs during the DeID procedure, by making full use of the rich medical priors within MeMa.
- We build the first medical-scene DeID benchmark, by comprehensively evaluating the proposed baseline and other recent DeID approaches on MeMa. Our approach is consistently superior in various aspects.

Related Work

Facial Datasets. Amounts of large-scale face datasets (Karras, Laine, and Aila 2019; Karras et al. 2017) have been proposed, but they primarily feature healthy individuals, limiting their use for medical-scene DeID. In contrast, we introduce a large-scale patient face dataset with rich medical manifestations. Our dataset also includes rich medical annotations, facilitating face DeID field in medical scenes.

Face De-identification. Early De-ID methods (Jourabloo, Yin, and Liu 2015) used the K-same algorithm. Recent approaches (Hukkelås, Mester, and Lindseth 2019; Maximov, Elezi, and Leal-Taixé 2020) leverage generative models to remove facial identity, while often compromising utility. More recent methods (Wen et al. 2023; Cai et al. 2024; Ren, Lee, and Ryoo 2018) aim to preserve more facial attributes and better serve common utilities such as gaze detection and image/video recognition (Kong and Fu 2022; Tian et al. 2022, 2020, 2021, 2019; Yan et al. 2023; Gao et al. 2024; Tan et al. 2024; Che et al. 2021), but not specifically medical signs (Chen et al. 2024a). In contrast, our approach leverages medical manifestation representations learned from real patient photos, preserving medical attributes during DeID. Additionally, it is reversible, similar to (Gu et al. 2020; Cao et al. 2021; Li et al. 2023), enabling reversal for medical audits.

Semantic Representation. Effectively modeling semantic information is crucial for modifying facial images, while maintaining perceptual quality (Min et al. 2024; Yi et al. 2021; Duan et al. 2022; Chen et al. 2024b; Li et al. 2024; Gao et al. 2022, 2021; Yi, Jiang, and Zhou 2024) and preserving medical utility. Previous approaches have leveraged contrastive learning (Tian et al. 2024b, 2023b) and masked image modeling (Tian et al. 2023a; Tian, Lu, and Zhai 2024) for self-supervised learning of image semantics. Recent studies have shown that pre-trained visual foundation models, such as stable diffusion (Rombach et al. 2022), exhibit even stronger semantic representations (Zhang et al. 2024; Tian, Lu, and Zhai 2025; Hedlin et al. 2024). In this work, we present the first adaptation of diffusion model-extracted semantics to the medical DeID problem.

Medical-scene Face Privacy Protection. Progress on this problem has been slow, often relying on simple methods like blurring or replacing faces with 3D masks (Yang et al. 2022), which discard critical disease signs. The progress gap is attributed to the lack of large-scale medical-scene facial datasets. Our work aims to address this gap.

Approach

We first build a new patient face dataset, termed **MeMa**. It addresses the lack of medical-scene facial datasets. MeMa is synthesized from real patient photos. Its synthetic nature avoids potential ethical problems. Expert physicians recognize its validity. Further, we propose a baseline model for the medical-aware facial DeID (Med-DeID) problem.

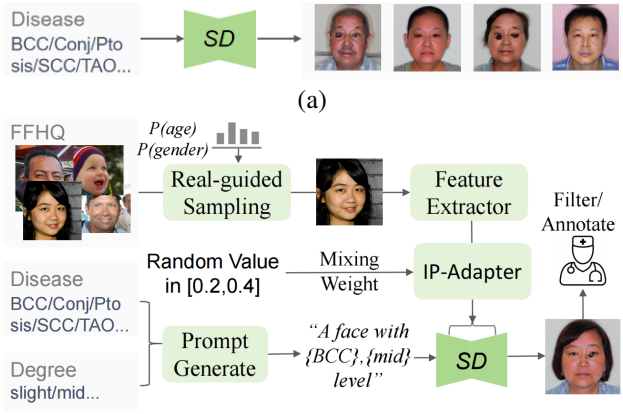


Figure 3: MeMa building pipeline. (a) Training patient face generation model on real patient data. (b) Rich-condition patient face sampling. $P(age)$ and $P(gender)$ denote the age and gender distributions, which are statistically derived from the real patients. 'SD' denotes the stable diffusion model.

MeMa Dataset

The overview of MeMa is shown in Fig. 2, which consists of 42,307 synthetic patient face images. MeMa closely mimics real patients in both visual appearance and statistics. Patient age and gender are estimated using the DeepFace framework (Serengil 2020; Serengil and Ozpinar 2021). We describe the main steps for assembling MeMa as follows.

Disease Categories: We take the eye clinic as an exemplar scene, since most eye diseases show typical external facial manifestations. In our study, we included patients with seven eye diseases. These are Basal Cell Carcinoma (BCC), Conjunctivitis (Conj), Uveitis, Ptosis, Squamous Cell Carcinoma (SCC), Strabismus (Strab), and Thyroid Associated Ophthalmopathy (TAO). We also included clinically Normal cases. Examples are shown in Fig. 2(a). The detailed manifestations of the above diseases can be found in the MSD medical manual (Merck & Co. 2024).

Real Patient Data Collection: We collected 39,323 photos of 12,467 real patients. They attended the Eye Clinic at Shanghai Ninth People's Hospital(SNPH) between January 2020 and June 2023. The photo-taking procedure was approved by the hospital's ethics committee. The patients' diagnosis results were collected from their medical records.

Generating MeMa from Real Data: As shown in Fig. 3, we first train a medical-aware generative model with the collected patient data. Then, we sample the virtual patients from the model by using proper conditions, aiming to generate *safe* and *diverse* samples. Finally, we recruit expert physicians to filter the images of bad medical quality, then annotate the filtered images. The steps are detailed as follows.

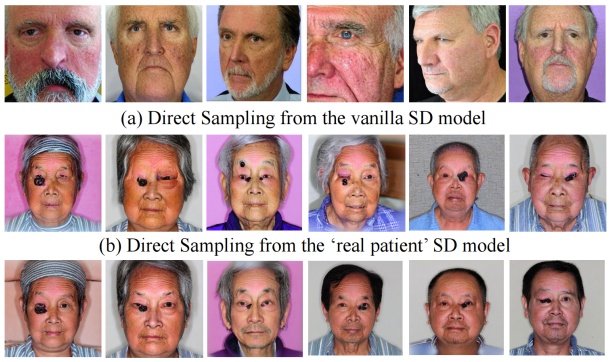
Step1: Medical-aware Generative Model Training: We first translate the disease type into the prompt caption 'A face, eye with {disease name}'. With the paired data of the real patient photographs and the disease type caption, we fine-tune the diffusion model (Rombach et al. 2022), producing the patient face generation model. As compared in Fig. 4 (a) and (b), after fine-tuning the SD model on our real patient dataset, the generated image shows typical medical signs and manifestations, while the vanilla SD model can not effectively generate images with reasonable medical manifestations, due to its limited medical knowledge.

Step2: Rich-Condition Patient Face Synthesis: Directly sampling from the real-patient generation model with the simple prompt 'A face, eye with {disease name}' is not enough, which shows two problems. First, identity leakage: the identity of most sampled patients can be found in the training dataset, potentially leaking the privacy of real patients. Second, mode collapse: the samples tend to be less diverse, with collapsed medical manifestation modes.

To address the *identity leakage* problem, we propose injecting facial attributes from public faces into the generation process. Specifically, we randomly sample face images from the FFHQ dataset and use the IP-Adapter (Ye et al. 2023) to inject these attributes. As shown in Tab. 1, this substantially reduces the average identity leakage percentage from 71.8% to 1.27%, when being evaluated with multiple face recognition models, *i.e.*, SphereFace (Liu et al. 2017), ArcFace (Deng et al. 2019), and CosFace (Wang et al. 2018).

	SphereFace	ArcFace	CosFace	Average
Direct Sample	73.45%	76.72%	65.34%	71.83%
Face Injection	1.62%	0.97%	1.24%	1.27%

Table 1: Effectiveness of injecting public face attribute for reducing the identity leakage percentage.



(a) Direct Sampling from the vanilla SD model

(b) Direct Sampling from the 'real patient' SD model

(c) *Rich-Condition* Sampling from the 'real patient' SD model

Figure 4: Comparison of different image generation strategies. We take the Basal Cell Carcinoma (BCC) disease as an example. 'SD' denotes the Stable Diffusion.

Sampling Strategy	Disease	Age	Gender
Random sampling	0.256	0.143	0.157
Real-guided sampling	**0.003**	**0.002**	**0.002**

Table 2: Wasserstein distance (Rüschendorf 1985) between the generated and the real patient image distributions. Smaller distance indicates more real-world alike generation.

To mitigate the *mode collapse* problem, we first randomly sample the public face injection weight from the range $[0.2, 0.4]$, instead of using a fixed weight. This leads to better feature fusion flexibility and improves output diversity. Second, we enhance the text prompt with severity descriptions, e.g., 'A face, eye with {disease name}, {slight/mid/heavy}-level'. This further improves diversity, even though no disease severity is annotated in the collected patient captions. The reason may be that the base SD model has learned a large dictionary of word semantics and can automatically connect the common 'severity' description words to the image generation process. As compared in Fig. 4 (b) and (c), with rich conditions injected, both the quality and diversity of the generated images are substantially improved.

To ensure the generated dataset's statistical characteristics match those of real patients, we calculate the distributions of real patient disease types, ages, and genders. We control the generated images to follow the above distributions. To control the disease type, we simply modify the disease name of the prompt. To control the age and gender of the generated images, we label FFHQ images using an age and gender estimation model (Serengil 2020), then select images based on this metadata for attribute injection. As shown in Tab. 2, the real distribution-guided sampling strategy produces a dataset with similar statistical characteristics to real patients.

Step3: Filtering and Annotation: After generating the images, we remove those with small identity feature distance to the original real patient set, ensuring the privacy of the real patients will not be leaked. Then, the physicians filter out the images with low medical utility quality. Finally, these physicians label the per-image disease information of the filtered dataset. Moreover, considering that lesion segmentation is another representative medical imaging task. We also ask the physicians to segment the tumor mask of the subset SCC images, producing the MeMa-Seg subset. The annotation procedure is assisted by the SAM model (Kirillov et al. 2023) and then refined by the physicians.

A Baseline Approach for Med-DeID

We propose a baseline approach to incorporate the rich medical manifestation knowledge within MeMa into the DeID procedure, which consists of two sub-modules: medical semantics encoding and medical semantics-preserved DeID.

Medical Semantics Encoding. The Med-DeID task requires preserving as much medical information as possible while obfuscating other identifying details. This necessitates a semantic encoder that recognizes local medical semantics.

Motivated by the strength of diffusion models in extracting fine-grained local semantics (Tian et al. 2024a; Tang et al. 2023), we train another diffusion model on the proposed MeMa dataset to learn the medical semantics. We adopt its first several blocks as the medical encoder Enc_{med}, instead of the whole network, for reducing the computational cost.

It should be mentioned that the roles of the diffusion models in the previous section and here are fundamentally different: the previous one is for high-quality image generation, whereas the one here is for extracting rich medical semantics. Our approach is very flexible, and the semantic encoder can be other choices, as analyzed in the experiment section.

Medical Semantics-Preserved DeID (MedSem-DeID). As illustrated in Fig. 5, our approach leverages the medical encoder Enc_{sem} to inject medical knowledge into the feature extraction procedure, as well as regularize the medical utility of the de-identified image.

Given the original image X, where H and W denote its height and width, an image encoder transforms X into the facial feature $f_{face} \in \mathbb{R}^{512\times\frac{H}{32}\times\frac{W}{32}}$. Meanwhile, we use Enc_{med} to extract the medical feature $f_{med} \in \mathbb{R}^{320\times\frac{H}{16}\times\frac{W}{16}}$. The f_{med} is downscaled and concatenated with f_{face}, passing through three consecutive residual blocks (He et al. 2016), producing f. Then, we employ a group of Transformer blocks (Vaswani et al. 2017), termed ID-Encryptor, to encrypt the ID information within f. Specifically, we flatten the spatial dimension of f, concatenate it with the password vector $P \in \mathbb{R}^{512}$, and feed the concatenated vector into ID-Encryptor, producing the encrypted feature f_{enc}. f_{enc} is passed through an image decoder network to result in the encrypted image X_{enc}. Please refer to the supplementary material for the network architecture details.

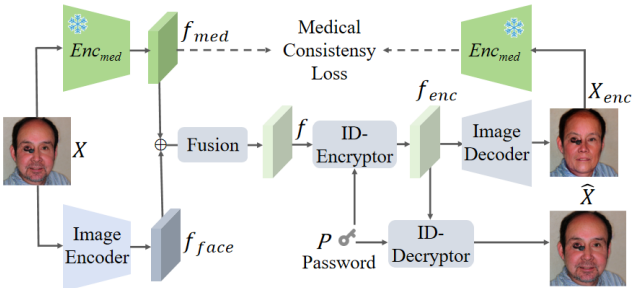


Figure 5: Overview of the proposed baseline model MedSem-DeID. The snow icon indicates the Enc_{med} is frozen during training DeID networks. The image decoder after the ID-decryptor is omitted for briefness. $\oplus$ denotes the channel-wise concatenation operation.

In medical contexts, it is often necessary to rigorously recheck results with expert physicians on the original image. Moreover, the Good Clinical Practice (GCP) guideline (Guideline 2001) mandates that all medical materials involved in the diagnosis process must be traceable. Therefore, we design our method to be reversible, enabling the recovery of the original image from the encrypted features. Given the original password P, f_{enc} can be decrypted back to $\hat{f}$, by another group of Transformer blocks termed ID-Decryptor. Then, $\hat{f}$ is reconstructed as the original image $\hat{X}$ by the image decoder. When an incorrect password is used, f_{enc} is reconstructed into a wrong image X_{wrong}.

Learning Objectives. The learning objective of the proposed MedSem-DeID is formulated as follows, $\mathcal{L} = \mathcal{L}_{deid} + \mathcal{L}_{rev-id} + \mathcal{L}_{wrong} + \lambda_{med}\mathcal{L}_{med} + \lambda_{rev}\mathcal{L}_{rev} + \mathcal{L}_{GAN}$. $\mathcal{L}_{deid} = \cos(\phi(X), \phi(X_{\text{enc}}))$ enforces the identity of the encrypted image apart from the original image, where ϕ denotes the pre-trained identity recognition network ArcFace (Deng et al. 2019), cos denotes the cosine similarity. $\mathcal{L}_{rev-id} = -\cos(\phi(X), \phi(\hat{X}))$ enforces the identity of reversibly decrypted image is the same as the original image. $\mathcal{L}_{wrong} = \cos(\phi(X), \phi(X_{enc}^{wrong}))$ enforces the identity of the image decrypted by the wrong password far away from the original image. $\mathcal{L}_{med} = \ell_2(f_{med}, Enc_{med}(X_{\text{enc}}))$ facilitate the encrypted image is similar to the original image in terms of medical semantics. $\mathcal{L}_{rev} = \ell_1(X, \hat{X})$ regularizes the appearance of the recovered image by right password is similar to the original one. ℓ_1 and ℓ_2 denote the mean absolute error (MSE) and the mean squared error (MSE) functions, respectively. The $\mathcal{L}_{\text{GAN}}$ is the adversarial generative network (GAN) loss, enforcing the photo-realism of all images. λ_{med} and λ_{rec} denote the balancing weights.

Experiments

Datasets. *MeMa:* the proposed MeMa dataset consists of 42,307 images in total, which is split into a training set (34,000 images), a hyper-parameter selection set (3,729 images), and a validation set (4,578 images). All images are labeled with the disease category. *MeMa-Seg:* for the BCC (basal cell carcinoma) disease type, we randomly select 600 images from the training set and 150 images from the validation set of MeMa, annotating the tumor masks for these images. This results in the MeMa-Seg dataset, which can be used to evaluate the fine-grained medical performance

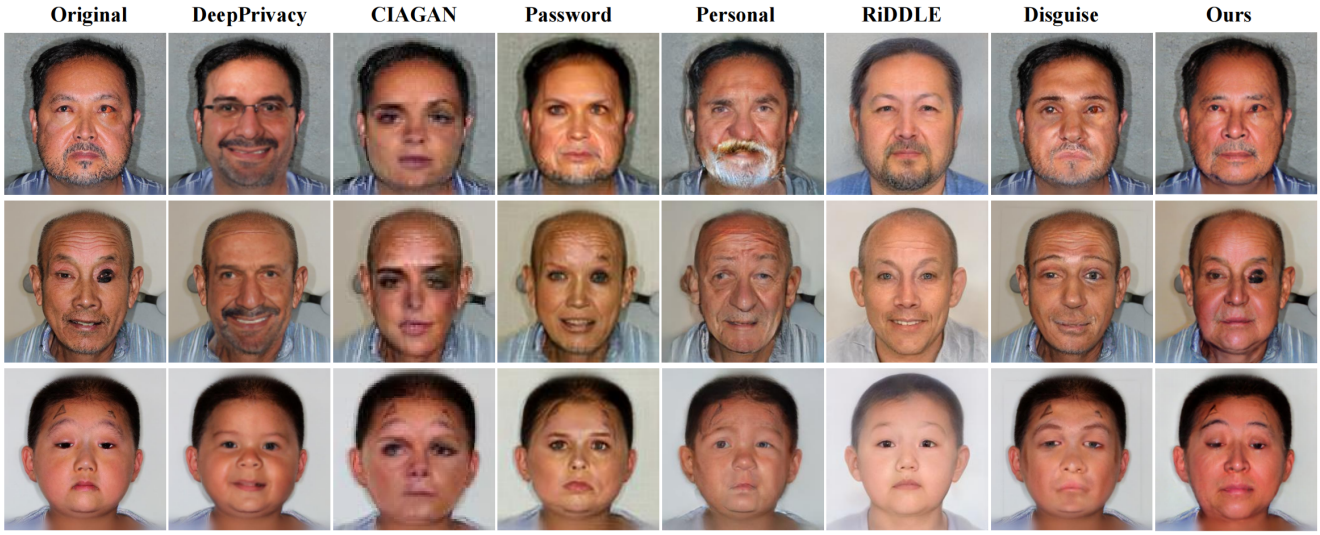


Figure 6: Qualitative results of different methods on the MeMA validation set.

Method	Rev	Utility Prior	Classification (%) ↑									Segmentation ↑	
			All	BCC	Conj	Normal	Ptosis	SCC	Strab	TAO	Uveitis	Dice	Jaccard
DeepPrivacy	✗	Landmark	40.80	91.20	4.52	80.38	69.10	0.95	3.66	59.23	17.32	0.0041	0.0021
CIAGAN	✗	Landmark	41.36	30.96	3.85	96.93	49.36	13.52	7.33	31.81	97.11	0.1280	0.0757
Disguise	✗	Landmark+Gaze	76.01	57.19	95.32	98.63	90.68	58.29	37.00	90.69	80.31	0.2751	0.1984
Password	✓	Unet	54.21	43.99	95.99	87.37	47.53	52.38	5.93	79.26	21.22	0.6336	0.5192
Personal	✓	Face Attributes	30.17	67.17	9.69	61.95	31.99	4.19	2.09	33.39	30.90	0.0136	0.0073
RiDDLE	✓	StyleGAN	30.01	29.27	0.00	74.40	58.32	0.00	0.00	4.04	74.02	0.0031	0.0017
Ours	✓	Med-Knowledge	**86.70**	**96.45**	**99.83**	**98.98**	**92.32**	**87.05**	**63.70**	**97.72**	57.56	**0.6775**	**0.5453**

Table 3: Comparison of various DeID methods on medical tasks. Classification and Segmentation tasks are evaluated on validation sets of MeMa and MeMa-Seg, respectively. 'Rev' denotes if the method is reversible.

of different DeID approaches. *Real-ECXHCSU:* we also collaborate with Eye Center of Xiangya Hospital of Central South University (ECXHCSU), enrolling 129 patients to conduct a real-world clinical trial. This aims to validate whether our algorithm, trained on the synthetic MeMa dataset, remains effective for real-world patients. Moreover, ECXHCSU is geographically distant from SNPH used to develop the MeMa dataset. This aims to further emphasize the generalization capability of our approach.

Implementation Details. For training the *patient face generator model*, we fine-tune Stable Diffusion v1-5 (Rombach et al. 2022) using the low-rank adaptation (LoRA) (Hu et al. 2021) technique, with the real patient data. The rank number is set to 64. We use the Adam optimizer (Kingma 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The learning rate starts at 1×10^{-4} and follows a cosine decay schedule. The batch size is 32, and the model is trained for ten epochs. It takes about five days to train the model on a machine equipped with two Nvidia A6000 GPUs. For training the *medical semantics encoder*, we use the same training strategy as above, except that the training data comes only from the MeMa training set. For training the *MedSem-DeID model*, we use the Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.99$. The initial learning rate is 2×10^{-4} and is halved after 150,000 iterations. The total iteration number is 300,000. The batch size is 16. Training takes approximately two days on a machine equipped with four Nvidia 4090 GPUs.

Benchmark Methods. For DeepPrivacy (Hukkelås, Mester, and Lindseth 2019), Password (Gu et al. 2020), CIAGAN (Maximov, Elezi, and Leal-Taixé 2020), and RiDDLE (Li et al. 2023), we adopt their officially released codes and models. For Disguise (Cai et al. 2024) and Personal (Cao et al. 2021), we request the materials from the authors.

Evaluation Protocol and Metrics. *Medical utility:* for the disease classification task, we fine-tune the DiNov2 model (Oquab et al. 2023) on the MeMa training set. We evaluate its Top1 accuracy on the MeMa validation set processed by various DeID approaches. For the tumor segmentation task, we use the nnU-Net (Isensee et al. 2021) to evaluate different methods on MeMa-Seg, adopting the Dice score (Kamnitsas et al. 2017) and Jaccard index (Fletcher, Islam et al. 2018) as metrics. *Real-word clinical utility:* we recruit three physicians to manually diagnose the images in Real-ECXHCSU, that are de-identified by various DeID approaches. Each image is diagnosed by all three physicians, and the final diagnosis is determined by a majority voting

strategy. We use Cohen's Kappa (k) (Banerjee et al. 1999) to measure the diagnosis consistency between the original and the de-identified images. k is a common metric for evaluating clinical trial outcomes in the medical field. *Identity protection:* following recent works (Cao et al. 2021; Wen et al. 2023), we use Euclidean distance between the identity features of de-identified and original faces, denoted as 'ID-Dis', to quantitatively evaluate the effectiveness of identity protection. Identity features are extracted by FaceNet (Schroff, Kalenichenko, and Philbin 2015) trained on CASIA (Yi et al. 2014), FaceNet trained on VGGFace2 (Cao et al. 2018), and SphereFace (Liu et al. 2017), which are not used in the training procedure. *Other utilities:* following previous methods (Li et al. 2023; Cai et al. 2024), we adopt the Dlib (King 2009) and L2CS-Net (Abdelrahman et al. 2023) to evaluate the landmark detection and gaze estimation performances. *Reversibility:* we compare our method against the previous reversible methods, in terms of ID similarity, medical results, and visual quality of the reconstructed original image.

Results

Medical Utility. As shown in Tab. 3, Our method achieves the best overall classification accuracy, outperforming the second-best method, Disguise, by more than 10%. For SCC disease, our approach outperforms DeepPrivacy, CIAGAN, Disguise, Password, Personal, and RiDDLE by 86.10%, 73.53%, 27.76%, 34.67%, 82.86%, and 87.05%, respectively. On the more challenging tumor segmentation task, our approach also performs best, achieving the highest Dice (0.6775) and Jaccard (0.5453) scores.

Moreover, we train Password and Disguise models on our MeMa dataset, improving their classification accuracy to 54.67% and 77.12%, respectively, but still lagging behind our 86.70%. This indicates that MeMa can enhance the efficacy of various DeID methods in medical contexts, and its full potential can be realized through specialized medical-scene methods like our MedSem-DeID.

In Tab. 3 (3rd column), we summarize the priors employed by different methods. Landmark priors (DeepPrivacy and CIAGAN) and high-level common priors (face attributes/StyleGAN adopted by Personal/RiDDLE) perform poorly in medical applications, *i.e.*, less than 50% accuracy and 0.2 segmentation Dice score. The gaze prior (Disguise) is effective for coarse-grained classification (76.01%) but fails in fine-grained segmentation task (0.2751). Password, using a U-Net to preserve high-frequency signals, excels in low-level segmentation (0.6336) but fails in high-level classification task (54.21%). This also introduces visual artifacts (Fig. 6, 4rd column). In contrast, our method, leveraging the medical semantics within MeMa, achieves superior performance in both classification (86.70%) and segmentation (0.6775) without handcrafted designs such as landmark.

Real-World Clinical Utility. We conduct a clinical trial on the Real-ECXHCSU cohort. As shown in Tab. 4, our method largely outperforms the recent DeID methods (Disguise and Password) across all five disease categories, achieving near-perfect consistency in the clinical outcomes, *i.e.*, k>0.81. Moreover, our approach is more flexible and effective than a recent hand-crafted DeID approach that is delicately designed for eye diseases, namely, digital mask (DM) (Yang et al. 2022). On complex diseases, such as BCC and Eyelid Nevus (EyelidN), DM does not work (k = 0.0566/0.0988) while our approach achieves satisfactory results (k = 0.8245/0.8346). Moreover, this real-world evaluation introduces additional disease categories not seen during training, *i.e.*, Entropion and EyelidN, highlighting the robustness and generalizability of our method. The diagnosis accuracy is provided in the supplementary material.

Method	Cohen's Kappa (k) ↑				
	BCC	TAO	Ptosis	Entropion	EyelidN
DM	0.0566	0.8159	0.8276	0.1879	0.0988
Disguise	0.7534	0.5824	0.7134	0.2467	0.1387
Password	0.4657	0.2758	0.4289	0.1329	0.0459
RiDDLE	0.1201	0.0751	0.0826	0.0937	0.0811
Ours	**0.8245**	**0.8278**	**0.8302**	**0.8256**	**0.8346**

Table 4: Comparison of different DeID methods in terms of the diagnosis outcomes, on the real-world cohort Real-ECXHCSU. k>0.81 indicates perfect clinical consistency.

Method	Rev	ID-Dis ↑		
		$\text{FaceNet}_{\text{VGGFace2}}$	$\text{FaceNet}_{\text{CASIA}}$	Sphere
DeepPrivacy	✗	1.1548	1.1831	1.1818
CIAGAN	✗	1.2843	1.2566	1.2881
Disguise	✗	1.3976	1.3607	1.3128
Password	✓	1.3380	1.3139	1.2629
Personal	✓	1.2819	1.2944	1.2351
RiDDLE	✓	**1.4278**	**1.3694**	*1.3583*
Ours	✓	*1.4007*	*1.3609*	**1.3601**

Table 5: Comparison of different methods on the MeMa validation set. The higher the ID-Dis, the better de-identified. **Bold** and *italic* indicates the best and the second-best result.

Method	DeepPrivacy	Password	Disguise	RiDDLE	Ours
Rate(%) ↓	5.76	7.76	2.89	2.13	**1.76**

Table 6: Face matching rate on Real-ECXHCSU.

De-Identification Performance. As shown in Tab. 5, our method achieves the best DeID performance of ID-Dis value 1.3601, when evaluated with the SphereFace face recognition model. With the $\text{FaceNet}_{\text{VGGFace2}}$ and $\text{FaceNet}_{\text{CASIA}}$ models, our approach outperforms all methods except RiDDLE. RiDDLE maps person images into the very low-dimensional StyleGAN (Karras, Laine, and Aila 2019) latent space and selects a sample with the maximum identity distance from this space. While this over-dimension-reduction operation benefits identity protection, it sacrifices much original face information, resulting in poor downstream utilities, as evidenced in Tab. 3 and Tab. 4.

Furthermore, we evaluate our approach on the LFW dataset (Huang et al. 2008). With the SphereFace facial recognition network, our approach achieves a face verification accuracy close to random guessing (50%).

Moreover, we simulate a real-world identity authentica-

Method	Rev	Landmark Error ↓				Gaze Error ↓	
		All	Eye	Mouth	Nose	Pitch	Yaw
DeepPrivacy	✗	193.91	5.86	155.02	33.02	7.72	7.06
CIAGAN	✗	313.36	23.87	215.11	74.37	13.00	7.78
Disguise	✗	94.09	5.78	67.29	21.01	*4.85*	*5.71*
Password	✓	**65.90**	*4.38*	**42.52**	*18.98*	5.34	9.97
Personal	✓	109.46	5.87	71.20	32.38	7.24	7.45
RiDDLE	✓	136.79	5.73	91.28	39.77	7.01	7.93
Ours	✓	*87.62*	**2.94**	*66.76*	**17.92**	**3.96**	**4.97**

Table 7: Comparison of different DeID methods, in terms of common utilities, on the MeMa validation set. Landmark error is calculated as the averaged pixel distance between the original and the de-identified image. **Bold** and *italic* indicates the best and the second-best performance.

tion system. We use ID-card photos of Real-ECXHCSU patients as the identity database. We then match the de-identified clinical photos within the ID photo database. Note that the ID photo may be a long time away from the clinical photo. As shown in Tab. 6, our approach achieves the lowest successful face matching rate of 1.76%, compared to other approaches such as Disguise (2.89%) and RiDDLE (2.13%).

Qualitative Results. As shown in Fig. 6, our approach uniquely preserves both coarse- and fine-grained medical cues, such as drooping eyelids and conjunctival redness. In contrast, DeepPrivacy masks and replaces the original face, Password retains color but distorts shapes. Disguise, Personal, and RiDDLE sacrifice medical cues for privacy. Besides, our approach demonstrates good visual quality.

Other Downstream Utilities. As shown in Tab. 7, our approach shows competitive or best results on facial landmark detection and gaze detection tasks. For example, our approach achieves an eye landmark detection error of 2.94, much lower than the second-best approach, Password, which achieves 4.38. This is due to our particularly preserved eye-related medical semantics. For gaze detection, although Disguise explicitly introduces the gaze detection loss, it still obtains a larger gaze error of 4.85 *v.s.* 3.96, proving the power of our semantics learned on MeMa.

Reversibility. As shown in Tab. 8, compared with other reversible methods, our method performs better in terms of identity recovery, image fidelity, and perceptual quality, achieving ID-Dis, Peak Signal-to-Noise Ratio (PSNR), and LPIPS (Zhang et al. 2018) values of 0.5642, 27.02dB, and 0.2098, respectively. Moreover, our approach exhibits the best disease classification accuracy of 89.34%, while the second-best Personal only obtains 73.08%.

We provide the qualitative results in Fig. 7. Only our approach precisely preserves the clinical diagnosis necessary sign, *i.e.*, the discolored left eye iris. RiDDLE generates high-quality facial textures, while obsoleting medical details. Password and Personal can not generate sharp details.

Model Analysis

Ablation Study for MedSem-DeID Model. Recalling that MedSem-DeID enhances the medical knowledge of the DeID pipeline in both the feature extraction procedure (f_{med}) and the loss function ($\mathcal{L}_{med}$), we verify the effectiveness of both strategies. As shown in Tab. 9, when trained on the common-scene facial dataset FFHQ without using any medical prior, the resulting model (M1) achieves an ID-Dis score of 1.3689 and a disease classification accuracy of 42.13%. When the training dataset is replaced with MeMa, the medical accuracy of the resulting model (M2) improves to 46.82% without compromising the DeID performance. After further introducing medical priors, no matter the f_{med} or the $\mathcal{L}_{med}$, the resulting models M3 and M4 show an obvious improvement in classification accuracy, *i.e.*, 69.94% and 71.35%, while slightly compromising the DeID results. When combining both strategies, the final model achieves a strong medical performance of 86.70%.

Different Medical Semantic Encoders. Our method is flexible, not relying on the typical implementation of the medical encoder. To verify this, we trained two other semantic encoders on MeMa. We fine-tuned a pre-trained ViT model (Sharir, Noy, and Zelnik-Manor 2021) using supervised and self-supervised learning strategies, specifically the

Method	ID-Dis↓	PSNR↑	LPIPS↓	Med-Class↑
Password	0.6382	26.29dB	0.2752	67.99%
Personal	0.5723	25.92dB	0.2240	73.08%
RiDDLE	0.8593	14.00dB	0.3732	47.46%
Ours	**0.5642**	**27.02dB**	**0.2098**	**89.34%**

Table 8: Comparison of the recovered image by various reversible approaches on MaMa validation set. Lower ID-Dis indicates the recovered identity is more similar to the original. Lower LPIPS indicates better perceptual quality.

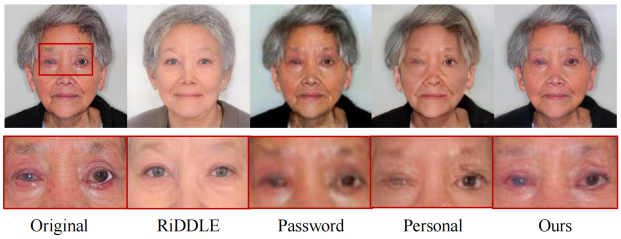


Figure 7: Qualitative results of the recovered image of different reversible DeID approaches. The ocular region is zoomed-in for a more clear comparison.

Model	Dataset	f_{med}	$\mathcal{L}_{med}$	ID-Dis↑	Med-Class↑
M1	FFHQ	✗	✗	1.3609	42.13%
M2	MeMa	✗	✗	1.3609	46.82%
M3	MeMa	✓	✗	1.3602	69.94%
M4	MeMa	✗	✓	1.3601	71.35%
Ours	MeMa	✓	✓	1.3601	**86.70%**

Table 9: Framework ablation Study. Both the MeMa dataset and the utilization of medical priors are useful. ID-Dis is calculated with the SphereFace network. Med-Class denotes the disease classification accuracy.

	ViT(Supervised)	ViT(MAE)	Diffusion
Med-Class↑	82.97%	85.26%	**86.70%**
ID-Dis↑	1.3600	1.3601	1.3601

Table 10: Impact of various medical semantic encoders.

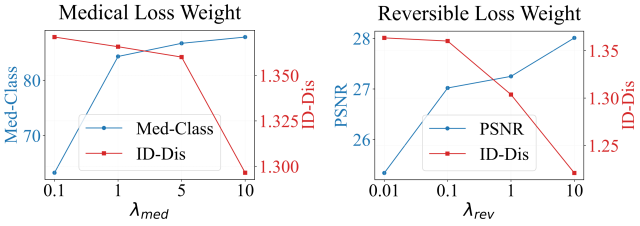


Figure 8: Impact of the loss weights for the medical (*left*) and reversible (*right*) loss terms.

masked auto-encoder (MAE) (He et al. 2022). As shown in Tab. 10, all three variants achieved decent performance. The supervised ViT performed the poorest due to the sparse disease category label for supervision. The diffusion model outperformed ViT(MAE) with 86.38% *vs.* 85.26%. This is likely because the LAION-5B (Schuhmann et al. 2022) dataset used for pre-training the base stable diffusion model is much larger than the pre-training dataset for the base ViT.

Different Loss Weights. As shown in Fig. 8 (*left*), increasing the weight of medical loss (λ_{med}) consistently improves disease classification accuracy, due to the enhanced medical information. However, this also makes the de-identified images more similar to the originals, compromising the DeID performance, *i.e.*, the reduced ID-Dis. We set $\lambda_{med} = 5$ to achieve the best trade-off between medical accuracy and DeID. For the weight controlling reversible reconstruction (λ_{rev}), a similar trade-off between the reconstructed image quality and DeID performance is observed, as shown in Fig. 8 (*right*). We set $\lambda_{rev} = 0.1$ to achieve optimal results.

Conclusion and Limitation

We have released a large-scale patient face dataset, MeMa, to facilitate research on medical privacy protection. Expert physicians validated and annotated MeMa. On this dataset, we established a comprehensive benchmark for medical-scene de-identification, also proposing a new baseline approach that outperforms previous approaches. A limitation is that the current dataset focuses on eye disease-related manifestations. Future work will expand the dataset to include other facial diseases, such as facial paralysis.

Acknowledgments

This work is supported by Strategic Research and Consulting Project of Chinese Academy of Engineering (2024-XBZD-18), National Natural Science Foundation of China (62225112), Shanghai Artificial Intelligence Laboratory, National Natural Science Foundation of China (62101326), National Natural Science Foundation of China (82388101), National Natural Science Foundation of China (72293585), and National Natural Science Foundation of China (72293580). We thank Min Zhou (Doctor of Medicine) and Xuefei Song (Doctor of Medicine) for their invaluable assistance with patient data collection, data annotation, and medical knowledge support.

References

Abdelrahman, A. A.; Hempel, T.; Khalifa, A.; Al-Hamadi, A.; and Dinges, L. 2023. L2cs-net: Fine-grained gaze estimation in unconstrained environments. In *2023 8th International Conference on Frontiers of Signal Processing (ICFSP)*, 98–102. IEEE.

Banerjee, M.; Capozzoli, M.; McSweeney, L.; and Sinha, D. 1999. Beyond kappa: A review of interrater agreement measures. *Canadian journal of statistics*, 27(1): 3–23.

Cai, Z.; Gao, Z.; Planche, B.; Zheng, M.; Chen, T.; Asif, M. S.; and Wu, Z. 2024. Disguise without disruption: Utility-preserving face de-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 918–926.

Cao, J.; Liu, B.; Wen, Y.; Xie, R.; and Song, L. 2021. Personalized and invertible face de-identification by disentangled identity information manipulation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3334–3342.

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 67–74. IEEE.

Che, Z.; Borji, A.; Zhai, G.; Ling, S.; Li, J.; Tian, Y.; Guo, G.; and Le Callet, P. 2021. Adversarial attack against deep saliency models powered by non-redundant priors. *IEEE Transactions on Image Processing*, 30: 1973–1988.

Chen, H.; Qu, Z.; Tian, Y.; Jiang, N.; Qin, Y.; Gao, J.; Zhang, R.; Ma, Y.; Jin, Z.; and Zhai, G. 2024a. A cross-temporal multimodal fusion system based on deep learning for orthodontic monitoring. *Computers in Biology and Medicine*, 180: 109025.

Chen, Z.; Sun, W.; Tian, Y.; Jia, J.; Zhang, Z.; Wang, J.; Huang, R.; Min, X.; Zhai, G.; and Zhang, W. 2024b. GAIA: Rethinking Action Quality Assessment for AI-Generated Videos. *arXiv preprint arXiv:2406.06087*.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4690–4699.

Duan, H.; Shen, W.; Min, X.; Tian, Y.; Jung, J.-H.; Yang, X.; and Zhai, G. 2022. Develop then rival: A human vision-inspired framework for superimposed image decomposition. *IEEE Transactions on Multimedia*, 25: 4267–4281.

Fletcher, S.; Islam, M. Z.; et al. 2018. Comparing sets of patterns with the Jaccard index. *Australasian Journal of Information Systems*, 22.

Gao, C.; Jiang, Y.; Wu, S.; Ma, Y.; Li, L.; and Liu, D. 2024. IMOFC: Identity-Level Metric Optimized Feature Compression for Identification Tasks. *IEEE Transactions on Circuits and Systems for Video Technology*.

Gao, C.; Li, L.; Liu, D.; Chen, Z.; Li, W.; and Wu, F. 2022. Two-step fast mode decision for intra coding of screen content. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8): 5608–5622.

Gao, C.; Liu, D.; Li, L.; and Wu, F. 2021. Towards task-generic image compression: A study of semantics-oriented metrics. *IEEE Transactions on Multimedia*, 25: 721–735.

Gu, X.; Luo, W.; Ryoo, M. S.; and Lee, Y. J. 2020. Password-conditioned anonymization and deanonymization with face identity transformers. In *European conference on computer vision*, 727–743. Springer.

Guideline, I. H. T. 2001. Guideline for good clinical practice. *J Postgrad Med*, 47(3): 199–203.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hedlin, E.; Sharma, G.; Mahajan, S.; Isack, H.; Kar, A.; Tagliasacchi, A.; and Yi, K. M. 2024. Unsupervised semantic correspondence using stable diffusion. *Advances in Neural Information Processing Systems*, 36.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, G. B.; Mattar, M.; Berg, T.; and Learned-Miller, E. 2008. Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments. *Month*.

Hukkelås, H.; Mester, R.; and Lindseth, F. 2019. Deepprivacy: A generative adversarial network for face anonymization. In *International symposium on visual computing*, 565–578. Springer.

Isensee, F.; Jaeger, P. F.; Kohl, S. A.; Petersen, J.; and Maier-Hein, K. H. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2): 203–211.

Jourabloo, A.; Yin, X.; and Liu, X. 2015. Attribute preserved face de-identification. In *2015 International conference on biometrics (ICB)*, 278–285. IEEE.

Kamnitsas, K.; Ledig, C.; Newcombe, V. F.; Simpson, J. P.; Kane, A. D.; Menon, D. K.; Rueckert, D.; and Glocker, B. 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical image analysis*, 36: 61–78.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

King, D. E. 2009. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10: 1755–1758.

Kingma, D. 2014. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Kong, Y.; and Fu, Y. 2022. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5): 1366–1401.

Li, C.; Zhang, J.; Zhang, Z.; Wu, H.; Tian, Y.; Sun, W.; Lu, G.; Liu, X.; Min, X.; Lin, W.; et al. 2024. R-Bench: Are your Large Multimodal Model Robust to Real-world Corruptions? *arXiv preprint arXiv:2410.05474*.

Li, D.; Wang, W.; Zhao, K.; Dong, J.; and Tan, T. 2023. RiDDLE: Reversible and Diversified De-Identification With Latent Encryptor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8093–8102.

Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; and Song, L. 2017. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 212–220.

Maximov, M.; Elezi, I.; and Leal-Taixé, L. 2020. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5447–5456.

Merck & Co., R. N. U., Inc. 2024. MSD MANUALS: The Trusted Provider of Medical Information since 1899. https://www.msdmanuals.com/.

Min, X.; Duan, H.; Sun, W.; Zhu, Y.; and Zhai, G. 2024. Perceptual video quality assessment: A survey. *Science China Information Sciences*, 67(11): 211301.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Price, W. N.; and Cohen, I. G. 2019. Privacy in the age of medical big data. *Nature medicine*, 25(1): 37–43.

Ren, Z.; Lee, Y. J.; and Ryoo, M. S. 2018. Learning to anonymize faces for privacy preserving action detection. In *Proceedings of the european conference on computer vision (ECCV)*, 620–636.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Rüschendorf, L. 1985. The Wasserstein distance and approximation theorems. *Probability Theory and Related Fields*, 70(1): 117–129.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Serengil, S. 2020. DeepFace: Face Recognition with Deep Neural Networks. https://github.com/serengil/deepface.

Serengil, S. I.; and Ozpinar, A. 2021. HyperExtended LightFace: A Facial Attribute Analysis Framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, 1–4. IEEE.

Sharir, G.; Noy, A.; and Zelnik-Manor, L. 2021. An image is worth 16x16 words, what is a video worth? *arXiv preprint arXiv:2103.13915*.

Tan, X.; Zhu, Y.; Cheng, Z.; Hu, M.; Zhang, X.; Pei, G.; Yu, C.; Li, Q.; Li, W.; and Wang, J. 2024. Low-cost and portable physiological signal monitor using PhysRate model. *Displays*, 81: 102605.

Tang, L.; Jia, M.; Wang, Q.; Phoo, C. P.; and Hariharan, B. 2023. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36: 1363–1389.

Tian, J.; Aggarwal, L.; Colaco, A.; Kira, Z.; and Gonzalez-Franco, M. 2024a. Diffuse Attend and Segment: Unsupervised Zero-Shot Segmentation using Stable Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3554–3563.

Tian, Y.; Che, Z.; Bao, W.; Zhai, G.; and Gao, Z. 2020. 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*, 71–89. Springer.

Tian, Y.; Lu, G.; Min, X.; Che, Z.; Zhai, G.; Guo, G.; and Gao, Z. 2021. Self-conditioned probabilistic learning of video rescaling. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4490–4499.

Tian, Y.; Lu, G.; Yan, Y.; Zhai, G.; Chen, L.; and Gao, Z. 2024b. A coding framework and benchmark towards low-bitrate video understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Tian, Y.; Lu, G.; and Zhai, G. 2024. SMC++: Masked Learning of Unsupervised Video Semantic Compression. *arXiv preprint arXiv:2406.04765*.

Tian, Y.; Lu, G.; and Zhai, G. 2025. Free-VSC: Free Semantics from Visual Foundation Models for Unsupervised Video Semantic Compression. In *European Conference on Computer Vision*, 163–183. Springer.

Tian, Y.; Lu, G.; Zhai, G.; and Gao, Z. 2023a. Non-semantics suppressed mask learning for unsupervised video semantic compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13610–13622.

Tian, Y.; Min, X.; Zhai, G.; and Gao, Z. 2019. Video-based early asd detection via temporal pyramid networks. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, 272–277. IEEE.

Tian, Y.; Yan, Y.; Zhai, G.; Chen, L.; and Gao, Z. 2023b. Clsa: a contrastive learning framework with selective aggregation for video rescaling. *IEEE Transactions on Image Processing*, 32: 1300–1314.

Tian, Y.; Yan, Y.; Zhai, G.; Guo, G.; and Gao, Z. 2022. Ean: event adaptive network for enhanced action recognition. *International Journal of Computer Vision*, 130(10): 2453–2471.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5265–5274.

Wen, Y.; Liu, B.; Cao, J.; Xie, R.; and Song, L. 2023. Divide and conquer: a two-step method for high quality face de-identification with model explainability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5148–5157.

Yan, Z.; Li, S.; Zhao, R.; Tian, Y.; and Zhao, Y. 2023. DHBE: data-free holistic backdoor erasing in deep neural networks via restricted adversarial distillation. In *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security*, 731–745.

Yang, Y.; Lyu, J.; Wang, R.; Wen, Q.; Zhao, L.; Chen, W.; Bi, S.; Meng, J.; Mao, K.; Xiao, Y.; et al. 2022. A digital mask to safeguard patient privacy. *Nature medicine*, 28(9): 1883–1892.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Yi, D.; Lei, Z.; Liao, S.; and Li, S. Z. 2014. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*.

Yi, F.; Chen, M.; Sun, W.; Min, X.; Tian, Y.; and Zhai, G. 2021. Attention based network for no-reference UGC video quality assessment. In *2021 IEEE international conference on image processing (ICIP)*, 1414–1418. IEEE.

Yi, X.; Jiang, Q.; and Zhou, W. 2024. No-reference quality assessment of underwater image enhancement. *Displays*, 81: 102586.

Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.