A Notation Table

Notations	Description
x, $\mathcal{X}$	input image, input space
y, $\mathcal{Y}$	disease label, label space
a, $\mathcal{A}$	sensitive attribute label, sensitive label space
z, $\mathcal{Z}$	embedding, embedding Space
$\phi : \mathcal{X} \rightarrow \mathcal{Z}$	freezed FM encoder
$f : \mathcal{Z} \rightarrow \mathcal{Y}$	disease classifier
$g : \mathcal{Z} \rightarrow \mathcal{A}$	sensitive attribute classifier
ϵ	universal edition
λ	regularization coefficient
T	disease target

Table 2: Notation Table

B Visualization

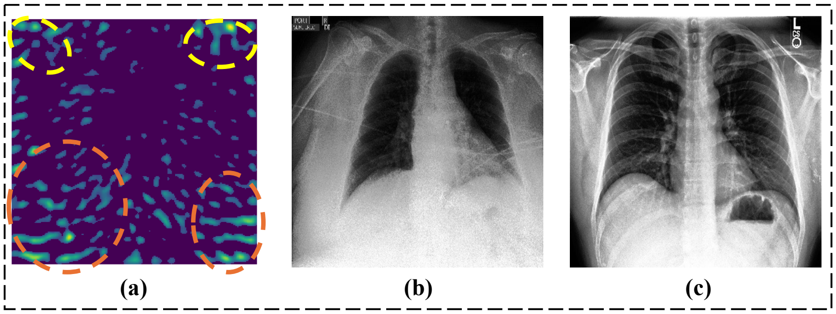

Fig. 3: Visualization of chest X-rays and DNE noise patterns (with Gaussian smoothing applied) to interpret gender-discriminative image regions. (a) The normalized UDE noise map, with larger noise highlighted by brighter color, reveals gender-discriminative features. The large noise circled in orange corresponds to the breast. The large noise circled in yellow reflects artifacts on X-ray, such as text notations. (b) A female chest X-ray. (c) A male chest X-ray.

C Data Distribution

Table 3: Training and Testing Set Distribution for *Pneumonia, Edema, and Pleural Effusion.*

Diseases	Pneumonia				Edema				Pleural Effusion			
	Negative		Positive		Negative		Positive		Negative		Positive	
	M	F	M	F	M	F	M	F	M	F	M	F
# Train Sample	1500	150	150	1500	5000	500	500	5000	5000	500	500	5000
# Test Sample	100	100	100	100	200	200	200	200	200	200	200	200

D Greedy Zero-order Optimization

Algorithm 1 Greedy Zero-order (GeZO) Optimization

Setup: Local Iterations R, step size s, step size decay k, Edit from last global epoch $\epsilon_{\texttt{epoch}}$, sample times C, best loss $\mathcal{L}_{\texttt{best}}$, best direction $d_{\texttt{best}}$, momentum: μ, batch size B.

```
procedure UpdateEditionPerEpoch($\epsilon_{\texttt{epoch}}$):
    $\epsilon^1 \leftarrow \epsilon_{\texttt{epoch}}$, $s^1 \leftarrow 0.01$, $v^1 \leftarrow 0$, $\mu \leftarrow 0.9$, $\mathcal{L}_{\texttt{best}} \leftarrow \infty$, $k \leftarrow 0.95$
    for $r = 1 \rightarrow R$ do
        $d_{\texttt{best}}, \mathcal{L}_{\texttt{best}} \leftarrow$GreedyGradient$(B, \epsilon^r, s^r)$    ▷ find the $d_{\texttt{best}}$ achieve $\mathcal{L}_{\texttt{best}}$
        if $d_{\texttt{best}} \neq$ None then
            $v^{r+1} \leftarrow \mu \cdot v^r + d_{\texttt{best}}$
            $\epsilon^{r+1} \leftarrow \epsilon^r + v^{r+1}$, $s^{r+1} \leftarrow s^r$    ▷ Update $\epsilon$ with momentum
        else
            $s^{r+1} \leftarrow k \cdot s^r$    ▷ Reduce step size if no improvement
    return $\epsilon^R$    ▷ Return the updated $\epsilon_{\texttt{epoch + 1}}$
procedure GreedyGradient($B, \epsilon^r, s^r$)
    $d_{\texttt{best}} \leftarrow$ None
    for $c = 1 \rightarrow C$ do    ▷ Sample $C$ times
        $\delta \leftarrow$ RandomPerturbation() $\cdot\, s_r$    ▷ Generate scaled perturbation
        for $d \in \{-1, 1\}$ do
            $\epsilon' \leftarrow \epsilon^r + d \cdot \delta$    ▷ Apply perturbation in both directions
            $\mathcal{L}_\epsilon = -\left[\frac{1}{B}\sum_{i=1}^{B} \mathcal{L}_{\texttt{CE}}(a_i, h(\phi(x_i + \epsilon')))\right] + \lambda||\epsilon'||_2$    ▷ One batch
            if $\mathcal{L}_\epsilon < \mathcal{L}_{\texttt{best}}$ then
                $\mathcal{L}_{\texttt{best}} \leftarrow \mathcal{L}_\epsilon$, $d_{\texttt{best}} \leftarrow d \cdot \delta$    ▷ Update best loss and direction
    return $d_{\texttt{best}}, \mathcal{L}_{\texttt{best}}$.
```