ICCV OCT 19-23, 2025 HONOLULU HAWAII

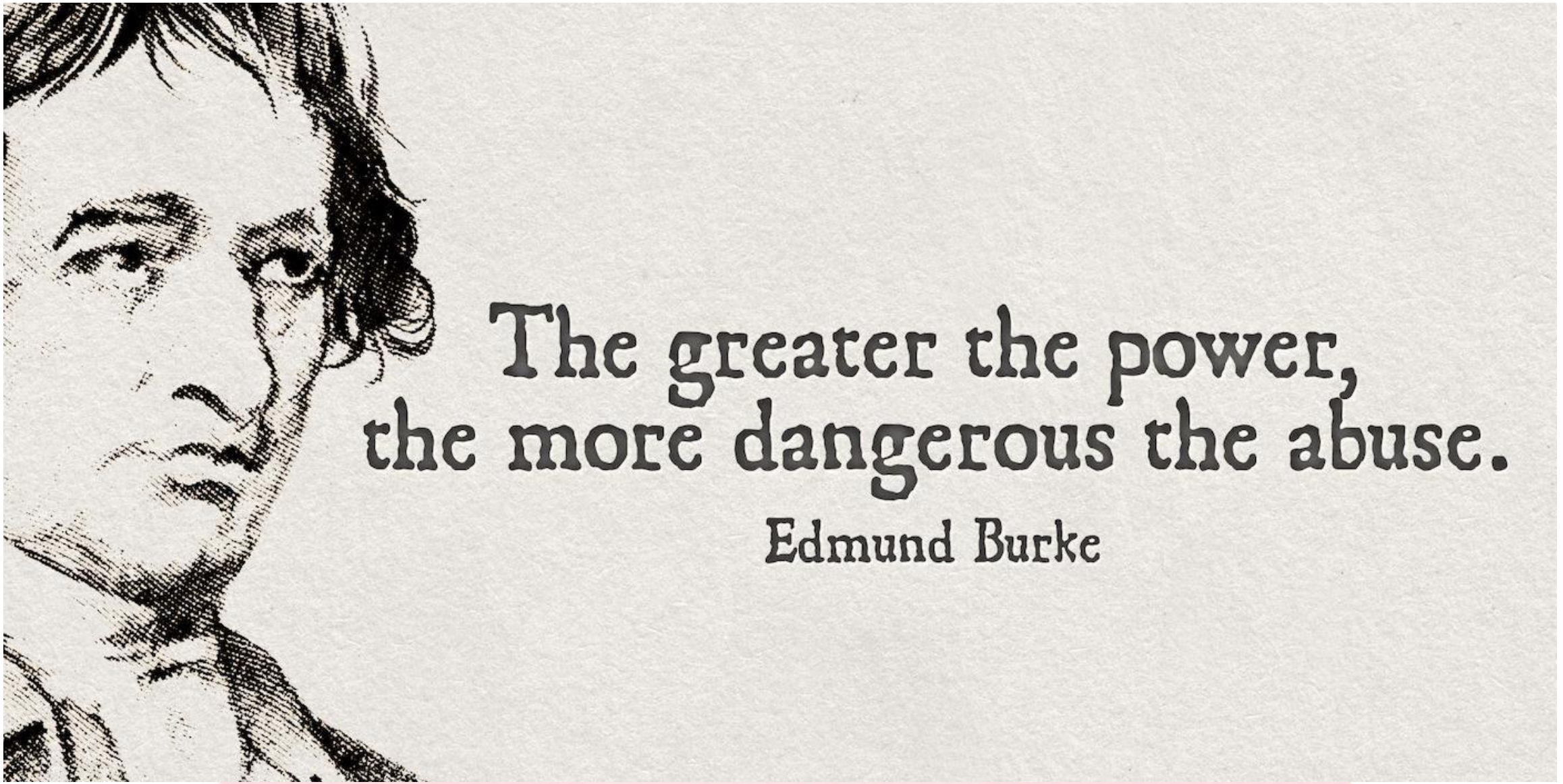
International Conference on Computer Vision, ICCV 2025

Moderating the Generalization of Score-based Generative Model

Wan Jiang[1], He Wang[2] , Xin Zhang[3], Dan Guo[1] , Zhaoxin Fan[4], Yunfeng Diao[1], Richang Hong[1]

[1]Hefei University of Technology [2]University College Londo
[3]San Diego State University [4]Beihang University

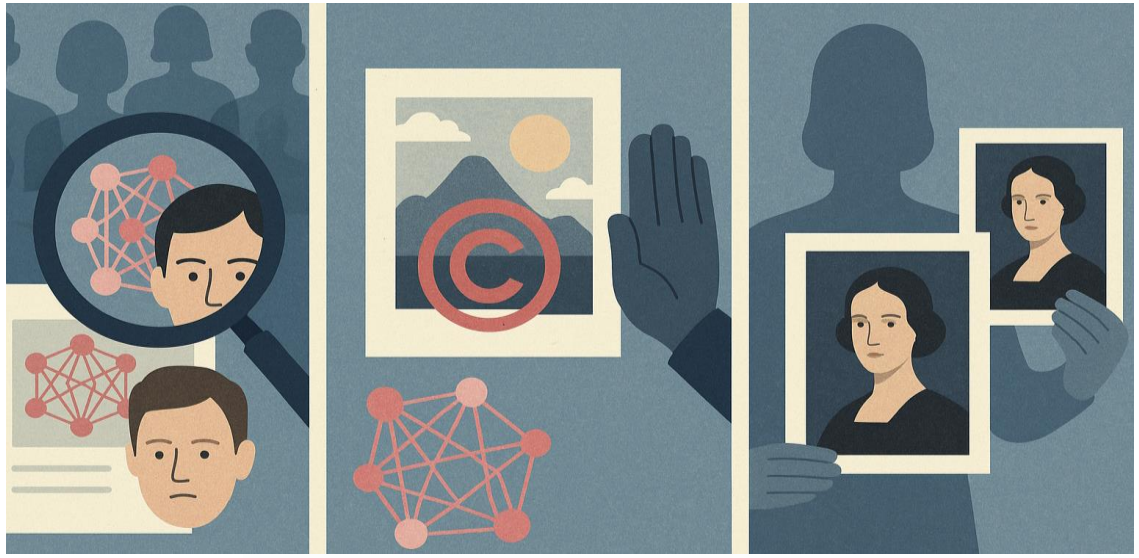
MSGM: Background &Motivation



MSGM: Background &Motivation

Unintended generalization

Generative models can produce unseen yet realistic data, causing **privacy leaks, copyright infringement, and style imitation** as their generalization power grows.



Can we make generative models forget undesired data?

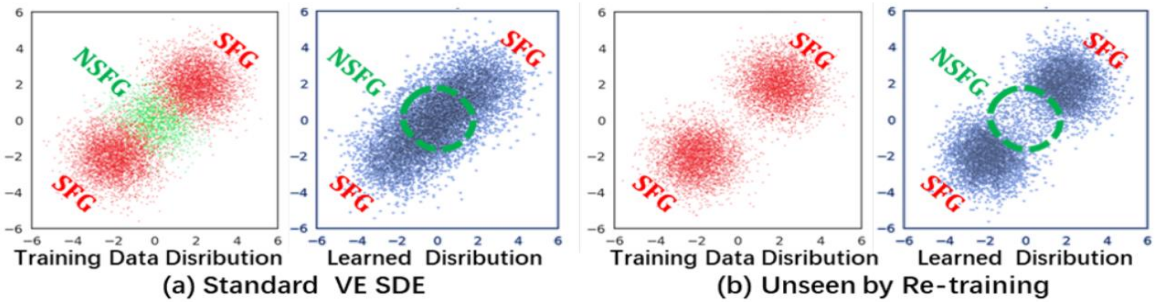
Limitations of existing methods:

Existing MU methods are designed for DDPMs or conditional generation, and due to differences in objectives and architectures, they cannot be directly applied to SGMs or more general unconditional generation.

MSGM: Background &Motivation

◆ Shortcomings of the "gold standard" MU

- In 2D Gaussian and high-dimensional experiments, the current "gold standard" MU for SGMs still regenerates data that should be forgotten, showing unintended generalization persists.

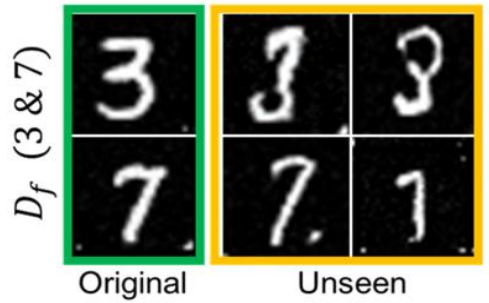


$$p_{\text{data}} = \underbrace{\frac{4}{5}\mathcal{N}((-2,-2),I)}_{\mathcal{D}_g} + \underbrace{\frac{2}{5}\mathcal{N}((0,0),I)}_{\mathcal{D}_f} + \underbrace{\frac{4}{5}\mathcal{N}((2,2),I)}_{\mathcal{D}_g},$$

Table 1: The Negative log-likelihood (NLL) values of different methods with respect to the data from p_{data}.

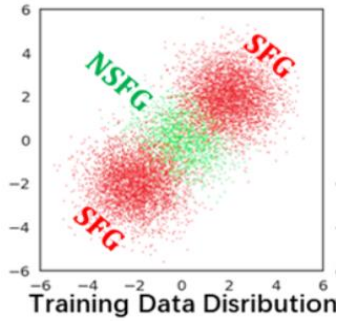
Test	Standard	Unseen	MSGM
$\mathcal{D}_g$	10.91	10.63	10.64
$\mathcal{D}_f$	10.73	11.59	39.01

- This raises doubts about the robustness and reliability of existing MU methods.



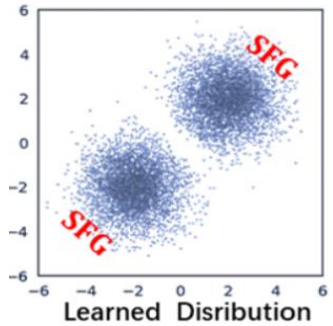
MSGM: Method

- **Key Idea：** Modify **the score function** to steer sampling away from NSFG data while preserving the SFG score to maintain generation quality.



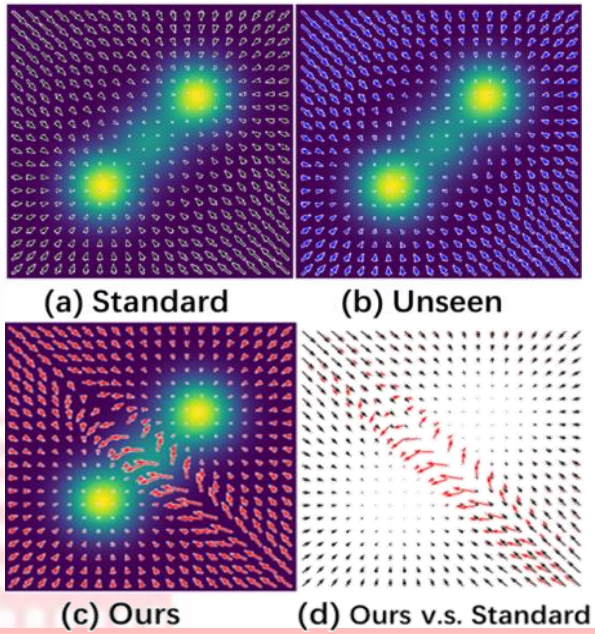
$$\arg\min_{\theta\in\Theta}\{D(p_\theta(\mathbf{x}),p_g(\mathbf{x}))-D(p_\theta(\mathbf{x}),p_f(\mathbf{x}))\}$$

$$\min_\theta L_{MSGM}=\min_\theta \mathbb{E}_{t\sim\mathcal{U}(0,T)}\left(\alpha L_g+(1-\alpha)L_f\right)$$



$$p_{\text{data}}=\underbrace{\frac{4}{5}\mathcal{N}((-2,-2),I)}_{\mathcal{D}_g}+\underbrace{\frac{2}{5}\mathcal{N}((0,0),I)}_{\mathcal{D}_f}+\underbrace{\frac{4}{5}\mathcal{N}((2,2),I)}_{\mathcal{D}_g},$$

$$L_g=\lambda(t)\{\mathbb{E}_{\mathbf{x}(0)}\mathbb{E}_{\mathbf{x}(t)}[\|s_\theta^{\mathbf{u}}(\mathbf{x}^g(t),t)-\nabla_{\mathbf{x}^g(t)}\log p_{0t}(\mathbf{x}^g(t)\mid\mathbf{x}^g(0))\|_2^2]\},\ \mathbf{x}^g\in\mathcal{D}_g.$$



Orthogonal-MSGM: Orthogonal complement steering for distributions that are moderately separable.

$$L_f=\lambda(t)\{\mathbb{E}_{\mathbf{x}(0)}\mathbb{E}_{\mathbf{x}(t)}[\|s_\theta^{\mathbf{u}}(\mathbf{x}^f(t),t)\cdot\nabla_{\mathbf{x}^f(t)}\log p_{0t}(\mathbf{x}^f(t)\mid\mathbf{x}^f(0))\|_2^2]\},\ \mathbf{x}^f\in\mathcal{D}_f.$$

Obtuse-MSGM: Negative correlation steering for highly overlapping distributions.

$$L_f=\lambda(t)\{\mathbb{E}_{\mathbf{x}(0)}\mathbb{E}_{\mathbf{x}(t)}[s_\theta^{\mathbf{u}}(\mathbf{x}^f(t),t)\cdot\nabla_{\mathbf{x}^f(t)}\log p_{0t}(\mathbf{x}^f(t)\mid\mathbf{x}^f(0))]\},\ \mathbf{x}^f\in\mathcal{D}_f.$$

MSGM: Experiments

◆ Class-wise/Feature-wise Ungeneration

Table 1. Quantitative results of unlearning undesirable features or classes.

Dataset	Model	Feature/Class	Unlearning Ratio (%) (↓)					Test	Negative Log-Likelihood ($\mathcal{D}_g$ (↓) and $\mathcal{D}_f$ (↑))				
			Standard	Ort	Obt	Unseen	EraseDiff		Standard	Ort	Obt	Unseen	EraseDiff
MNIST	VESDE	3	11.0	**0.4**	1.5	1.8	1.4	$\mathcal{D}_g$	2.82	3.92	3.70	3.07	3.30
		7	15.8	**0.8**	3.6	2.3	1.1						
		3 and 7	26.8	**1.2**	5.1	4.1	2.5	$\mathcal{D}_f$	2.78	**13.23**	12.08	3.01	3.74
CIFAR-10	VPSDE	automobile	11.2	1.9	**0.9**	3.4	9.7	$\mathcal{D}_g$	3.12	3.22	3.28	3.09	3.09
		dog	13.4	10.0	11.5	10.8	**8.2**						
		automobile and dog	24.6	**11.9**	12.4	14.2	17.9	$\mathcal{D}_f$	3.20	**5.94**	4.37	3.21	4.10
STL-10	VPSDE	airplane	12.1	**2.4**	3.6	3.8	2.6	$\mathcal{D}_g$	2.90	2.90	2.92	2.90	-
								$\mathcal{D}_f$	2.19	8.94	**9.25**	2.32	-
CelebA	VPSDE	bangs	19.6	3.5	**0.7**	6.7	1.2	-	-	-	-	-	-	-

MSGM: Experiments

◆ Unlearning DDPM and Fine-tune

Table 2. Fine-tune quantitative results.

Dataset	Model	Feature/Class	Unlearning Ratio (%) (↓)				
			Stand	Ort	Obt	Unseen	EraseDiff
CIFAR-10	VPSDE	automobile	11.2	2.7	**0.6**	3.4	9.4
		dog	13.4	**8.7**	8.9	10.8	5.2
		automobile and dog	24.6	11.4	**9.5**	14.2	14.6
	DDPM	automobile	13.1	3.3	**1.6**	2.7	3.0
		dog	13.9	5.4	**3.6**	4.5	4.4
		automobile and dog	27.0	8.7	**5.2**	7.2	7.4
CelebA	VPSDE	bangs	19.6	2.6	**0.1**	6.7	1.9

Table 3. Fine-tuned NLL values for CIFAR-10 on VP SDE.

Test	Standard	Unseen	Unlearning	Unseen	EraseDiff
$\mathcal{D}_g$	2.89	3.06	4.36	2.92	3.06
$\mathcal{D}_f$	2.91	10.36	14.96	2.95	4.38

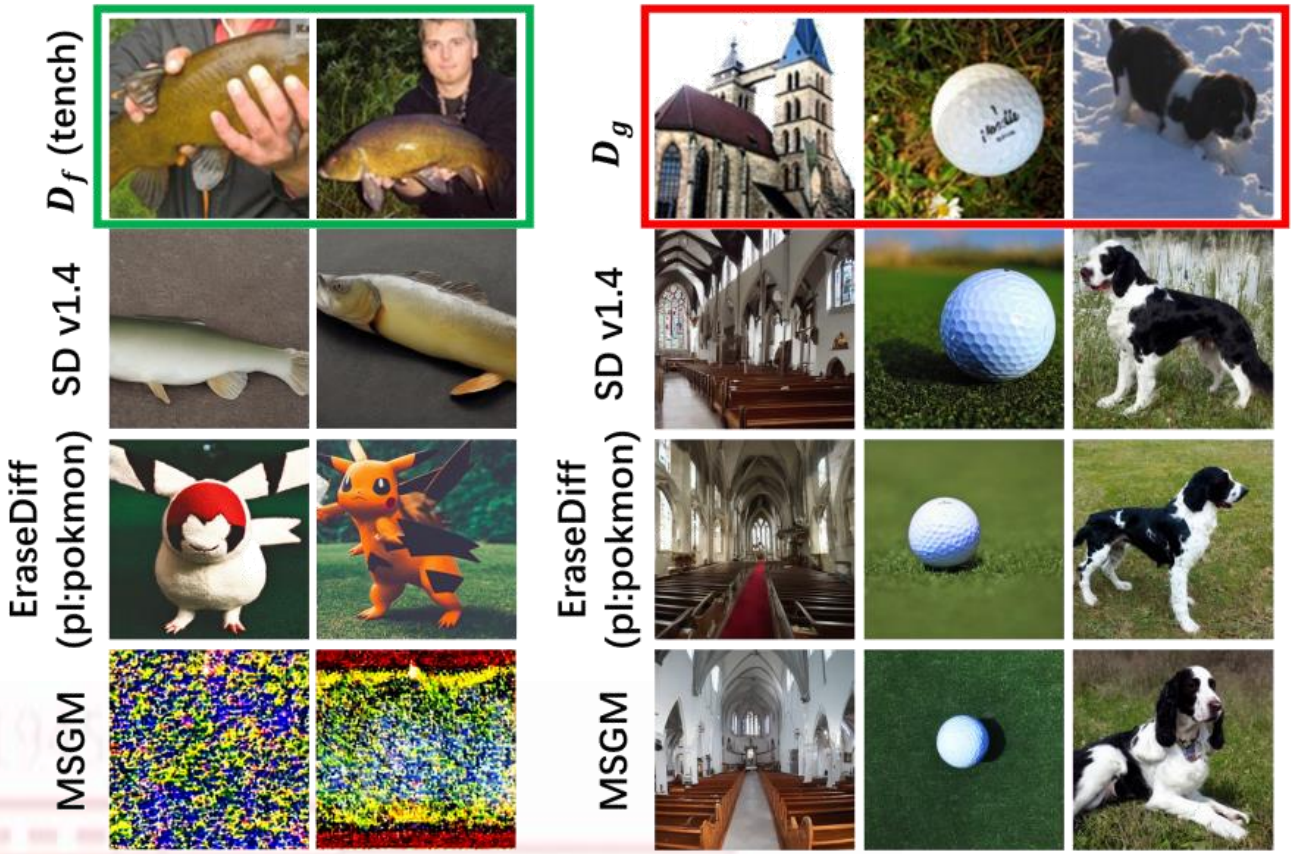
MSGM: Experiments

◆ Unlearning T2I Generation on High-resolution Datesets

Table 4. Unlearning T2I models on Imagenette.

	SD v1.4	ESD	EraseDiff	MSGM
FID of $\mathcal{D}_g(\downarrow)$	4.89	3.09	3.09	3.08
CR($\downarrow$)	0.74	0.00	0.00	0.00

Comparison of EraseDiff using semantically similar pseudo-label ‘cyprinoid’ (for ‘tench’) versus our pseudo-label-free MSGM approach.



Visualization of diverse unlearning methods applied to fine-tune SD v1.4 on the Imagenette dataset. The left green box displays NSFG images sampled from forgetting datasets. ‘pl’ indicates the pseudo-label used during training.

MSGM: Experiments

◆ Application to Downstream Tasks

Table 5. The inpainting comparison results.

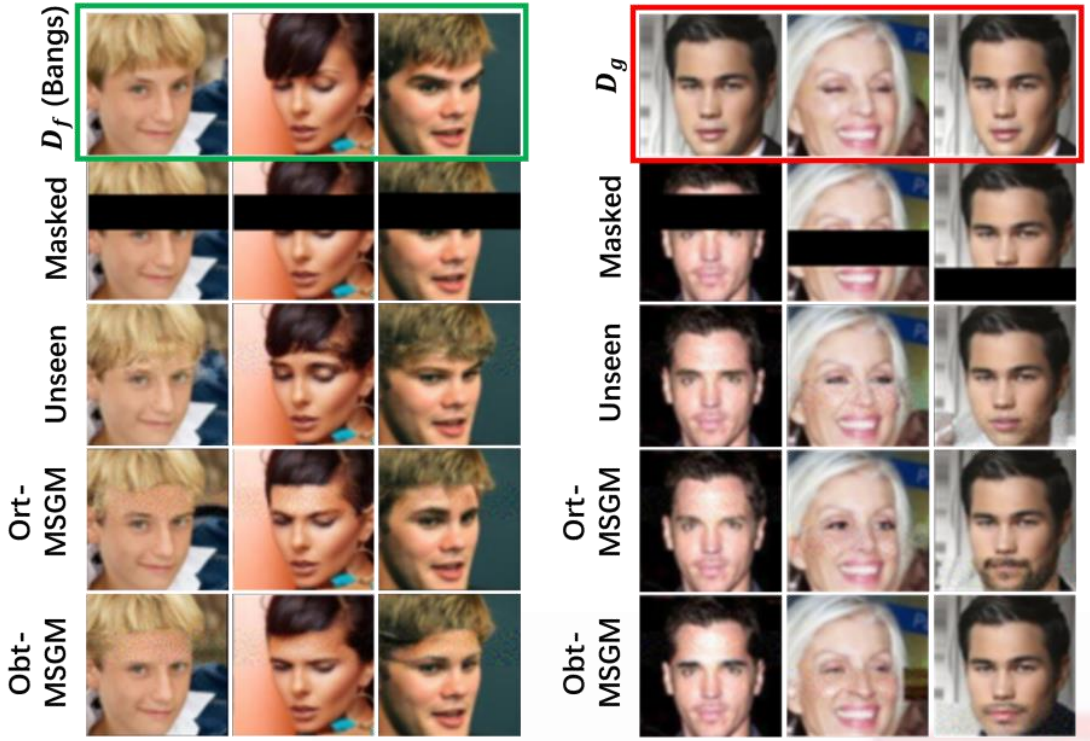
Dataset		ACC (%) ($\mathcal{D}_g(\uparrow)$ and $\mathcal{D}_f(\downarrow)$)					FID of $\mathcal{D}_g(\downarrow)$				CLIP of $\mathcal{D}_g(\downarrow)$				PSNR of $\mathcal{D}_g(\uparrow)$				SSIM of $\mathcal{D}_g(\uparrow)$			
		Clean	Stand	Ort	Obt	Unseen	Stand	Ort	Obt	Unseen	Stand	Ort	Obt	Unseen	Stand	Ort	Obt	Unseen	Stand	Ort	Obt	Unseen
CIFAR-10	$\mathcal{D}_g$	95.4	72.5	75.5	74.7	75.8	13.11	15.96	13.46	13.64	6.80	6.80	6.77	6.72	31.09	31.09	31.01	31.03	0.56	0.55	0.54	0.54
	$\mathcal{D}_f$	95.5	75.0	57.2	**49.6**	59.7																
STL-10	$\mathcal{D}_g$	96.3	83.4	83.6	83.1	84.5	28.48	29.95	28.55	28.56	8.50	8.51	8.50	8.50	31.18	31.17	31.17	31.18	0.59	0.58	0.57	0.59
	$\mathcal{D}_f$	96.3	84.1	59.5	**50.3**	54.9																
CelebA	$\mathcal{D}_g$	98.3	95.5	99.0	99.5	98.0	29.42	30.31	29.43	30.42	8.96	8.96	8.97	8.94	34.54	34.52	34.50	34.54	0.83	0.82	0.81	0.82
	$\mathcal{D}_f$	98.3	53.0	1.0	**0.5**	2.0																

Table 6. The comparison results of reconstruction.

Dataset		ACC (%) ($\mathcal{D}_g(\uparrow)$ and $\mathcal{D}_f(\downarrow)$)					FID of $\mathcal{D}_g(\downarrow)$				CLIP of $\mathcal{D}_g(\downarrow)$				PSNR of $\mathcal{D}_g(\uparrow)$				SSIM of $\mathcal{D}_g(\uparrow)$			
		Clean	Stand	Ort	Obt	Unseen	Stand	Ort	Obt	Unseen	Stand	Ort	Obt	Unseen	Stand	Ort	Obt	Unseen	Stand	Ort	Obt	Unseen
CIFAR-10	$\mathcal{D}_g$	95.4	88.1	87.7	87.0	87.9	5.52	5.71	5.57	5.94	6.91	6.90	6.89	6.90	31.91	32.15	32.19	31.82	0.92	0.92	0.93	0.91
	$\mathcal{D}_f$	95.5	74.4	**48.4**	69.6	70.3																

MSGM: Experiments

Application to Downstream Tasks



The comparison of inpainting results on the CelebA dataset. The mask size is 64 × 16. The restored results on Df are displayed on the left. The restored results on Dg are displayed on the right.

The comparison of reconstruction results on the CIFAR10 dataset. The top, middle and bottom columns are the original images, reconstruction images by Unseen, and reconstruction images by Ort respectively.

Thank you for watching

Moderating the Generalization of Score-based Generative Model

Paper Link: https://arxiv.org/pdf/2412.07229

Code Link: https://github.com/yunfengdiao/Moderated-Score-based-Generative-Model