UnlearnCanvas: A Stylized Image Dataset to Benchmark Machine Unlearning for Diffusion Models

Yihua Zhang[1], Chongyu Fan[1], Yimeng Zhang[1], Yuguang Yao[1], Jinghan Jia[1], Jiancheng Liu[1], Gaoyuan Zhang[2], Gaowen Liu[3], Ramana Kompella[3], Xiaoming Liu[1], Sijia Liu[1,2]

[1]Michigan State University, [2]IBM Research, [3]Cisco Research

Website

Paper

Code

Machine Unlearning for Diffusion Models

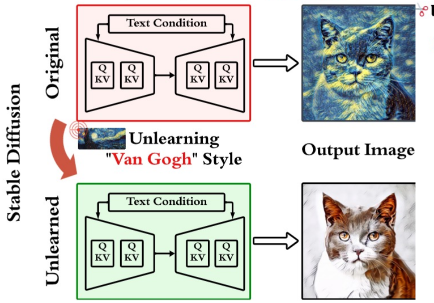

Figure 1: An illustration of MU for DMs.

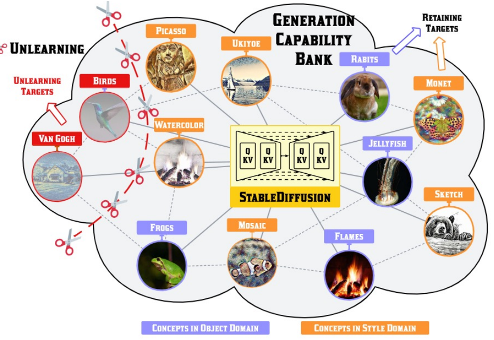

Figure 2: MU with UnlearnCanvas.

Motivation: Challenges in MU evaluation.

- (C1) The absence of a consensus on a diverse unlearning target test repository.
- (C2) The lack of a systematic study on 'retainability' of DMs post-unlearning.
- (C3) The precision challenge in evaluating DM-generated images.

Related Work	Object Unlearning				Style Unlearning			
	Target #	Quant. UE	Quant. RT	Qual.	Target #	Quant. UE	Quant. RT	Qual.
ESD [23]	11	✓	✓	✓	5	✗	✗	✓
CA [16]	4	✓	✗	✓	4	✗	✗	✓
FMN [28]	7	✓	✗	✓	1	✗	✗	✓
UCE [24]	11	✓	✗	✓	7	✓	✗	✓
SA [29]	31	✓	✗	✓	1	✗	✗	✓
SalUn [27]	11	✓	✓	✓	0	✗	✗	✗
SPM [26]	24	✓	✓	✓	5	✗	✗	✓
SEOT [30]	22	✓	✓	✓	2	✗	✗	✓
EDiff [31]	2	✓	✓	✓	0	✗	✗	✗
SHS [32]	1	✓	✗	✓	0	✗	✗	✗

Table 1. Overview of MU evaluations for DMs.

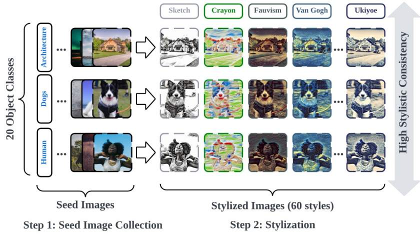

Figure 3. An overview of UnlearnCanvas.

Our Proposal: UnlearnCanvas Dataset

- (A1) Style-object dual supervision enables a rich unlearning target bank.
- (A2) Enabling both 'in-domain' and 'cross-domain' retainability analyses.
- (A3) High stylistic consistency ensures precise style definitions and enables accurate quantitative evaluations.

Evaluation Pipeline with UnlearnCanvas

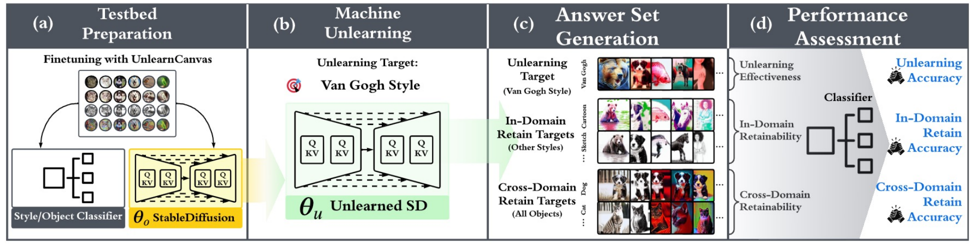

Figure 4. An illustration of the evaluation pipeline proposed in this work using UNLEARNCANVAS when unlearning a specific target concept 'Van Gogh Style'. Unlearning performances are quantitatively assessed (marked in blue) to accurately reflect the unlearning performance portrait. The unlearning target of the pipeline could traverse all the styles and objects to achieve a comprehensive evaluation.

Experiment Results

Benchmarking Current DM Unlearning Methods for Style and Object Unlearning

Method	Effectiveness: Style Unlearning UA (↑)	Style Unlearning IRA (↑)	Style Unlearning CRA (↑)	Object Unlearning UA (↑)	Object Unlearning IRA (↑)	Object Unlearning CRA (↑)	FID (↓)	Efficiency: Time (s) (↓)	Memory (GB) (↓)	Storage (GB) (↓)
ESD [23]	98.58%	80.97%	93.96%	92.15%	55.78%	44.23%	65.55	6163	17.8	4.3
FMN [28]	88.48%	56.77%	46.60%	45.64%	90.63%	73.46%	131.37	350	17.9	4.2
UCE [24]	98.40%	60.22%	47.71%	94.31%	39.35%	34.67%	182.01	434	5.1	1.7
CA [25]	60.82%	96.01%	92.70%	46.67%	90.11%	81.97%	54.21	734	10.1	4.2
SalUn [27]	86.26%	90.39%	95.08%	86.91%	96.35%	99.59%	61.05	667	30.8	4.0
SEOT [30]	56.90%	94.68%	84.31%	23.25%	95.57%	82.71%	62.38	95	7.34	0.0
SPM [26]	60.94%	92.39%	84.33%	71.25%	90.79%	81.65%	59.79	29700	6.9	0.0
EDiff [31]	92.42%	73.91%	98.93%	86.67%	94.03%	48.48%	81.42	1567	27.8	4.0
SHS [32]	95.84%	80.42%	43.27%	80.73%	81.15%	67.99%	119.34	1223	31.2	4.0

Table 2. Performance overview of DM unlearning methods on UNLEARNCANVAS: Metrics (UA, IRA, CRA, FID) are averaged over style and object cases. Arrows (↑/↓) indicate desired value direction. Best results are green; underperforming ones are red, highlighting areas for improvement.

- Retainability Matters in Performance Assessment and sole reliance on Unlearning Accuracy (UA) is insufficient
- Retaining cross-domain concepts (CRA) is harder than within-domain concepts (IRA)
- No single method performs consistently across all domains

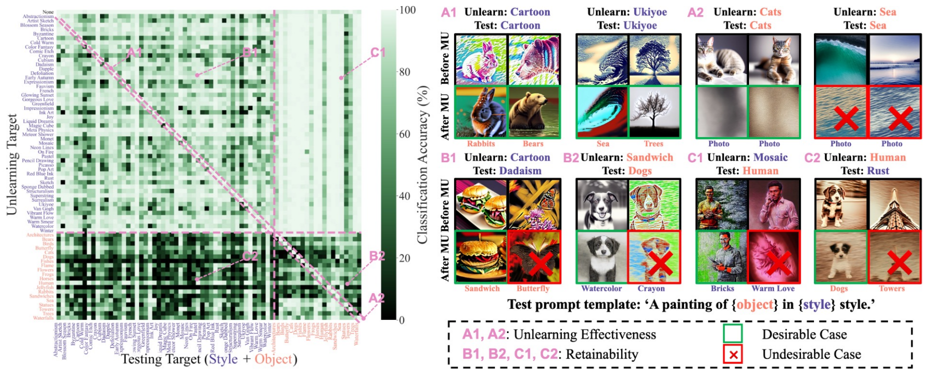

Figure 5. Left: Heatmap of ESD's unlearning accuracy (UA) and retainability (IRA, CRA) on UnlearnCanvas. The x-axis lists tested concepts, y-axis shows unlearning targets, with styles (blue) and objects (orange). Regions A, B, and C correspond to UA, IRA, and CRA for style ('1') and object ('2') unlearning. Lighter colors indicate better performance; the first row shows pre-unlearning reference. Right: Example images before and after unlearning.

Benchmarking Current DM Unlearning Methods in More Challenging Scenarios

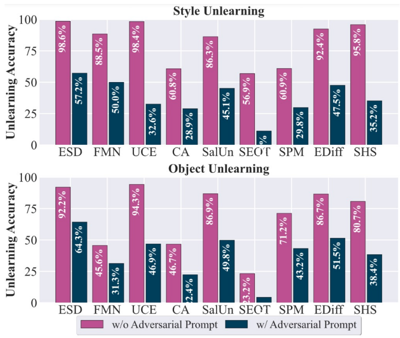

Figure 6. Adversarial prompts expose significant vulnerabilities in MU methods, with UA dropping below 60%, emphasizing the need for worst-case evaluations.

Method	UA	SC	OC	UP
ESD [23]	91.42%	4.88%	14.72%	84.38%
FMN [28]	45.37%	68.73%	62.74%	83.25%
UCE [24]	75.97%	4.53%	5.72%	35.42%
CA [25]	47.92%	10.08%	56.35%	81.54%
SalUn [27]	42.21%	62.45%	70.93%	87.28%
SEOT [30]	29.32%	45.31%	53.64%	85.45%
SPM [26]	45.72%	41.34%	36.32%	67.82%
EDiff [31]	71.33%	35.23%	26.32%	51.52%
SHS [32]	55.32%	14.34%	24.32%	83.95%

Table 3. Unlearning style-object combinations is significantly harder, with UA dropping over 20% and retainability falling below 20%, highlighting challenges in defining precise unlearning scopes.

Contact: zhan1908@msu.edu