# Unlearning in Diffusion Models under Data Constraints: A Variation Inference Framework

Subhodip Panda, Varun MS, Shreyans Jain, Sarthak Maharana, Prathosh AP

Transaction on Machine Learning Research (TMLR), 2026

May 19, 2026

# Table of Contents: Variational Diffusion Unlearning

# Introduction

- **Problem Statement**
  1. Can we forget generating certain undesired samples from a pretrained Diffusion Models?
  2. Can we forget generating data in data-constrained settings, such as without access to a full training dataset or retaining dataset?

# Introduction

- **Problem Statement**
  1. Can we forget generating certain undesired samples from a pretrained Diffusion Models?
  2. Can we forget generating data in data-constrained settings, such as without access to a full training dataset or retaining dataset?

- **Objectives**
  1. Propose a machine-unlearning methodology that can prevent the generation of outputs containing undesired features from a pre-trained diffusion model.
  2. Model's generation quality shouldn't degrade much.

# Contributions

- Proposed a variational inference framework.
- Our proposed method is more data-efficient. No access to the full training data corpus.
- Methodology to unlearn different classes from a pre-trained unconditional DDPM and unlearning the generation of particular user-identified features, such as artistic styles, from a pretrained Stable Diffusion Model trained on the LAION-5B dataset.

# Problem Formulation

- A pre-trained DDPM model, denoted as $f_{\theta^*}$, with initial parameters $\theta^* \in \Theta \subseteq \mathbb{R}^d$ trained on $D = \{x_i\}_{i=1}^m$.
- The entire data space can be expressed as the union of $\mathcal{X}_r$ and $\mathcal{X}_f$, where $\mathcal{X} = \mathcal{X}_r \bigcup \mathcal{X}_f$ or $\mathcal{X}_r = \mathcal{X} \backslash \mathcal{X}_f$. We denote the distributions over $\mathcal{X}_f$ and $\mathcal{X}_r$ as $P_{\mathcal{X}_f}$ and $P_{\mathcal{X}_r}$, respectively.
- The objective of the unlearning mechanism is to output a sanitized model $\theta^u$ that does not produce outputs within the domain $\mathcal{X}_f$. This implies that the model should be trained to generate data samples conforming to the distribution $P_{\mathcal{X}_r}$ only.
- Assuming access to the whole training dataset, a computationally expensive approach to achieving this is by retraining the entire model from scratch using a dataset $D_r = \{x_i\}_{i=1}^s \sim P_{\mathcal{X}_r}^s$ or equivalently, $D_r = D \setminus D_f$, where $D_f = \{x_i\}_{i=1}^t \sim P_{\mathcal{X}_f}^t$.
- It is important to note that we do not have access to $D_r$ in our setting.

# Variational Inference Framework

## Bayesian Inference Lens

- Using Bayesian inference, retrained parameters $\theta^r$ are a sample from the retrained model's parameter posterior distribution $P(\theta|D_r)$ i.e., $\theta^r \sim P(\theta|D_r)$. Similarly, $\theta^* \sim P(\theta|D_r, D_f)$.
- Using simple Bayes Theorem, we try to approximate the retrained model's parameter posterior distribution $P(\theta|D_r)$ as follows:

$$P(\theta|D_r) \propto \frac{P(\theta|D_r, D_f)}{P(D_f|\theta)}$$

# Variational Inference Framework

## Bayesian Inference Lens

- Using Bayesian inference, retrained parameters $\theta^r$ are a sample from the retrained model's parameter posterior distribution $P(\theta|D_r)$ i.e., $\theta^r \sim P(\theta|D_r)$. Similarly, $\theta^* \sim P(\theta|D_r, D_f)$.
- Using simple Bayes Theorem, we try to approximate the retrained model's parameter posterior distribution $P(\theta|D_r)$ as follows:

$$P(\theta|D_r) \propto \frac{P(\theta|D_r, D_f)}{P(D_f|\theta)}$$

- We don't know or find it hard to estimate the normalizing constant.
- A variational KL-divergence minimization over a set of probable approximate posterior distributions $\mathcal{Q}$ as follows:

$$Q^*(\theta) = \underset{Q(\theta)\in\mathcal{Q}}{\operatorname{argmin}} D_{KL}\left(Q(\theta)\middle|\middle|Z \cdot \frac{P(\theta|D_f, D_r)}{P(D_f|\theta)}\right) \tag{1}$$

# Proposed Method

**Theorem (Informal)**

*Gaussian mean-field approximation,* $Q(\theta) = \prod_{i=1}^{d} \mathcal{N}(\theta_i, \sigma_i^2)$ *and* $P(\theta|D_r, D_f) = \prod_{i=1}^{d} \mathcal{N}(\mu_i^*, \sigma_i^{*2})$

$$D_{KL}\left(Q(\theta) \,\middle\|\, Z \cdot \frac{P(\theta \mid D_f, D_r)}{P(D_f \mid \theta)}\right) \gtrsim -\sum_{x_0 \in D_f} \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)} \left[\frac{(1-\alpha_t)}{\alpha_t(1-\bar{\alpha}_{t-1})} \|\epsilon_0 - \epsilon_\theta(x_t, t)\|^2\right] + \sum_{i=1}^{d} \left[\frac{(\theta_i - \mu_i^*)^2}{2\sigma_i^{*2}} + \frac{\sigma_i^2}{2\sigma_i^{*2}} + \log \frac{\sigma_i^*}{\sigma_i} - \frac{1}{2}\right].$$

# Proposed Method

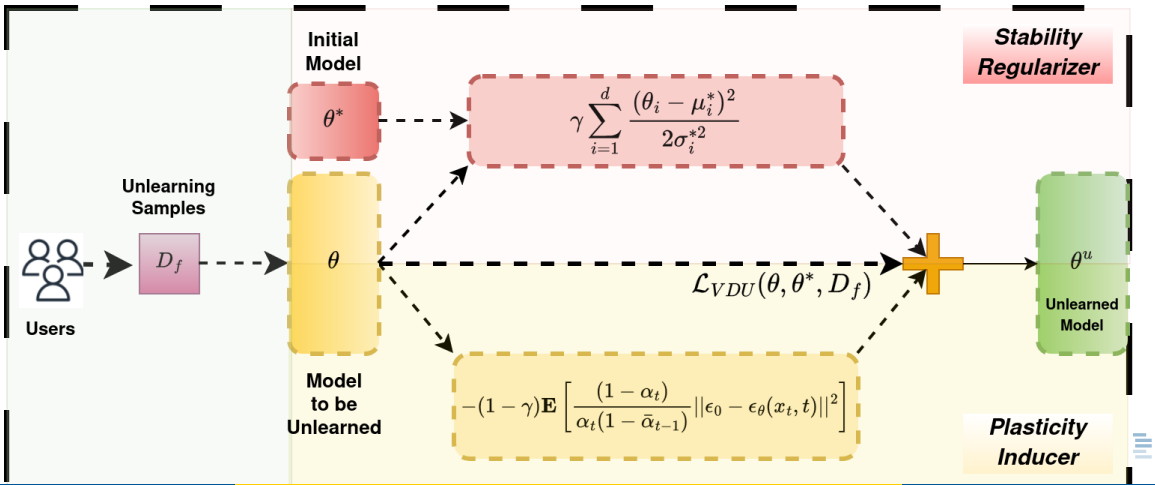
**Theorem (Informal)**

*Gaussian mean-field approximation,* $Q(\theta) = \prod_{i=1}^{d} \mathcal{N}(\theta_i, \sigma_i^2)$ *and* $P(\theta|D_r, D_f) = \prod_{i=1}^{d} \mathcal{N}(\mu_i^*, \sigma_i^{*2})$

$$D_{KL}\left(Q(\theta) \,\middle\|\, Z \cdot \frac{P(\theta \mid D_f, D_r)}{P(D_f \mid \theta)}\right) \gtrsim -\sum_{x_0 \in D_f} \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)} \left[\frac{(1-\alpha_t)}{\alpha_t(1-\bar{\alpha}_{t-1})} \|\epsilon_0 - \epsilon_\theta(x_t, t)\|^2\right] + \sum_{i=1}^{d} \left[\frac{(\theta_i - \mu_i^*)^2}{2\sigma_i^{*2}} + \frac{\sigma_i^2}{2\sigma_i^{*2}} + \log\frac{\sigma_i^*}{\sigma_i} - \frac{1}{2}\right].$$

- We use the above lower-bound as a fine-tune loss to modify the model

$$\mathcal{L}_{VDU}(\theta, \theta^*, D_f) = -(1-\gamma) \mathop{\mathbb{E}}_{\substack{x_0 \sim D_f \\ \epsilon_0, t}} \left[\frac{(1-\alpha_t)}{\alpha_t(1-\bar{\alpha}_{t-1})} ||\epsilon_0 - \epsilon_\theta(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_0, t)||^2\right] + \gamma \sum_{i=1}^{d} \frac{(\theta_i - \mu_i^*)^2}{2\sigma_i^{*2}} \tag{2}$$



# Experimental results

| Datasets | Unlearned Classes | Fine-tuning | | SA | | SalUn | | ESD | | VDU (Ours) | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | PUL (%) | u-FID | PUL (%) | u-FID | PUL (%) | u-FID | PUL (%) | u-FID | PUL (%) | u-FID |
| **MNIST** | Digit-1 | 90.07 | 5.90 | 15.01 | 301.26 | 45.28 | 68.75 | 56.66 | 66.15 | **75.06** | **14.43** |
| | Digit-8 | 94.29 | 6.15 | 48.95 | 161.09 | 32.53 | 50.18 | 32.68 | 55.67 | **68.96** | **38.20** |
| **CIFAR-10** | Automobile | 88.01 | 9.02 | 2.25 | 92.89 | 12.01 | 66.47 | 8.97 | 73.48 | **60.87** | **30.86** |
| | Ship | 92.86 | 10.74 | **85.02** | 249.89 | 55.13 | 75.48 | 58.27 | 77.12 | 71.63 | **24.46** |
| **tIN** | Fish | 82.56 | 13.75 | **67.7** | 32.00 | 39.33 | 42.70 | 41.67 | 45.19 | 58.75 | **29.92** |
| | Butcher Shop | 83.37 | 16.82 | 6.7 | 24.00 | 2.87 | 33.47 | 8.76 | 39.48 | **66.67** | **24.17** |

*(a) Original "0"* *(b) Unlearned "0"* *(c) Original "bottle"* *(c) Unlearned "bottle"*

Unlearning Concept = "Van_Gogh"

Prompt = "a portrait of van gogh"

Original | Unlearned epoch=1 | Unlearned epoch=2 | Unlearned epoch=3 | Unlearned epoch=4 | Unlearned epoch=5

# Conclusion and Future works

- Parametric view is restricted more generalized function space variational inference can be used.
- Further Gaussian assumption in parameter space can also be relaxed.
- This framework can be extended to Language Models.