Patient-level Machine Unlearning in Latent Diffusion Models: On the Limits of the Privacy-Utility Trade-off

Inês Cardoso[1,2]
Tiago Gonçalves[2]
Luís F. Teixeira[2]
Wilson Silva[1] W.J.DOSSANTOSSILVA@UU.NL
[1] *AI Technology for Life, Department of Information and Computing Sciences, Department of Biology, Utrecht University, Utrecht, The Netherlands*
[2] *INESC TEC, Faculty of Engineering, University of Porto, Porto, Portugal*

Editors: Under Review for MIDL 2026

Abstract

While federated learning frequently uses synthetic images for case-based explanations, diffusion models pose a privacy risk by potentially memorizing and recreating patient-identifiable data. Machine unlearning offers a way to remove specific training data influences, but its effectiveness for patient-level anonymization in generative models is not yet well understood. In this work, we present the first empirical analysis of patient-level unlearning in latent diffusion models, testing three strategies, including our novel KL-Away approach. Our results reveal a critical trade-off: methods that successfully unlearn data degrade diagnostic utility, whereas utility-preserving techniques fail to protect privacy, leaving over 20% of patients re-identifiable. We attribute this to feature entanglement and distributed memorization, suggesting that existing unlearning techniques are currently insufficient for reliable patient anonymization.

Keywords: case-based explanations, federated learning, machine unlearning, privacy.

1. Introduction

Case-based explanation systems justify artificial intelligence (AI) predictions by retrieving visually similar examples, aligning closely with clinical reasoning in radiology (Montenegro and Cardoso, 2025). In federated learning (FL) settings, where data cannot leave local institutions, such approaches require synthetic image catalogs as proxies for real patient data (Campos et al., 2024). Among generative approaches, diffusion models have emerged as a dominant paradigm for high-quality medical image synthesis (Croitoru et al., 2023). However, literature shows that these models memorize training data and reproduce patient-identifiable features in their outputs (Carlini et al., 2023), thus violating clinical privacy requirements and regulations such as the EU General Data Protection Regulation (GDPR), which mandates a Right to Be Forgotten (Dessers and Valcke, 2025).

Machine unlearning (MU) offers a potential solution by selectively removing the influence of specific training samples from a trained model without full retraining (Bourtoule et al., 2021). While MU has been studied for classification tasks and concept erasure in text-to-image models, its application to data-point unlearning in image-to-image diffusion models (particularly at the patient level in medical imaging) remains unexplored. Closest related works include SISS (Alberti et al., 2025), which targets individual entities one at a time, and

the approach proposed by Li et al. (2024), which degrades forget-set outputs to noise rather than preserving image coherence. Neither is designed for multi-patient anonymization.

We present a novel systematic study of patient-level MU in latent diffusion models for synthetic medical image generation. Using a re-identification pipeline, we identify memorized patients and construct *Forget* and *Remain* sets, and introduce KL-Away, a Kullback-Leibler divergence-based unlearning method tailored to latent diffusion models. Through experiments on MIMIC-CXR-JPG (Johnson et al., 2019), we show why existing MU methods fail to reliably anonymize synthetic medical images, revealing a fundamental privacy-utility trade-off driven by feature entanglement and distributed memorization in latent space.

2. Method

Given a latent diffusion model, we aim to remove patient-specific memorization such that synthetic outputs can no longer expose real individuals. Our pipeline has two stages: (1) identifying memorized patients and constructing the *Forget* and *Remain* sets; and (2) applying MU methods to suppress their influence on the model's output.

We generated 4,000 synthetic chest radiographs and applied a patient re-identification pipeline. For this purpose, we trained two ResNet-50 (He et al., 2016) Siamese networks on the real training data: a Patient Retrieval Network (contrastive loss) and a Patient Verification Network (classification loss). For each synthetic image, the retrieval network retrieves the three most similar real training images; the verification network then classifies whether the top match belongs to the same patient. Synthetic images with a verification score $\geq 50\%$ are flagged as non-anonymous.

Real images of patients associated with at least one flagged synthetic image form the *Forget* set D_f; the remaining patients form the *Remain* set D_r. This yields 885 non-anonymous synthetic images linked to 631 patients, leaving 3,115 images in D_r.

We evaluated three unlearning strategies—KL-Away, SISS, and SalUn+KL-Away—operating in the latent space of the Medfusion VAE (Müller-Franzes et al., 2023). We additionally tested **Gradient Ascent (GA)**, but it led to model collapse and is therefore excluded from quantitative comparison. Each method combines a forgetting loss on D_f with the standard diffusion loss on D_r.

KL-Away (proposed) encourages the current model to diverge from a frozen copy of the original model θ_0 on D_f. The loss is:

$$\mathcal{L}_{\text{KL-Away}} = -\frac{1}{2}\|\hat{\epsilon}_\theta(x_f, t, c_f) - \hat{\epsilon}_{\theta_0}(x_f, t, c_f)\|^2. \tag{1}$$

Under Gaussian diffusion assumptions, maximizing KL divergence between the current model and a frozen reference reduces to this squared difference. Unlike GA, KL-Away diverges from the model's own predictions, providing a more targeted forgetting signal.

SISS (Alberti et al., 2025) combines naive deletion with gradient ascent via importance sampling, balancing forgetting speed and model stability.

SalUn+KL-Away applies the SalUn saliency mask (Fan et al., 2024) to restrict weight updates to parameters most responsible for memorization, combined with KL-Away.

Table 1: Evaluation before and after unlearning. ↑ higher is better, ↓ lower is better.

Method	Non-Anon ↓	UA (%) ↑	ODA (%) ↑	PIR (%) ↓	AUC ↑	FID ↓
Original (no unlearning)	885	N/A	77.9	15.8	0.88	53.5
SISS	**525**	**81.3**	**86.9**	**6.8**	0.52	179.2
KL-Away	602	78.9	84.9	7.2	0.83	84.9
SalUn + KL-Away (50%)	766	76.7	80.9	9.2	**0.85**	**81.2**

3. Experiments

We use MIMIC-CXR-JPG (Johnson et al., 2024), restricted to postero-anterior views (15,223 images from 10,156 patients), with cardiomegaly as the target condition. After removing 222 mislabeled lateral-view images, we train a Medfusion latent diffusion model (Müller-Franzes et al., 2023) and generate 4,000 synthetic images with a balanced class split.

We report anonymization and unlearning metrics: number of non-anonymous images (Non-Anon), Unlearning Accuracy (UA), Overall Dataset Anonymization (ODA), and Patient Identifiability Ratio (PIR). Clinical utility is measured via area under the receiver operating characteristic curve (AUC) of a DenseNet-121 (Huang et al., 2017) cardiomegaly classifier trained on the synthetic data and evaluated on a real held-out test set. Image quality is assessed via Fréchet Inception Distance (FID) (Heusel et al., 2017).

Table 1 summarizes the results. SISS achieves the strongest anonymization (UA 81.3%, PIR 6.8%) but at a severe cost: AUC drops to 0.52 and FID to 179.2, rendering images clinically unusable. KL-Away offers a better balance (PIR 7.2%, AUC 0.83), while SalUn+KL-Away achieves the best utility–privacy compromise (PIR 9.2%, AUC 0.85). Full-parameter methods outperform localized variants, suggesting memorization is distributed across parameters. No method achieves complete anonymization. We attribute this to the high feature entanglement between D_f and D_r (76%), a configuration identified as worst-case for unlearning (Fan et al., 2024), and the inability of a finite synthetic sample to capture all memorized patients, causing exposure of previously undetected identities.

4. Conclusion

We show that patient-level MU in latent diffusion models faces a fundamental privacy–utility trade-off that current gradient-based methods cannot resolve. High feature entanglement between memorized and retained data, combined with distributed memorization in latent space, makes reliable patient-level forgetting challenging. Furthermore, *Forget sets* built from finite synthetic samples cannot capture all memorized patients, meaning unlearning may expose previously undetected identities. These findings suggest that reliable patient-level deletion requires not only new unlearning techniques but also fundamentally different pipelines for identifying memorized patient's PII. Until such methods are developed, alternatives such as disentanglement or differential privacy offer more reliable trade-offs.

Acknowledgments

This work was supported by the Dutch Research Council (NWO) through the AiNED XS Europa project NGF.1609.241.009.

References

Silas Alberti, Kenan Hasanaliyev, Manav Shah, and Stefano Ermon. Data Unlearning in Diffusion Models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=SuHScQv5gP.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pages 141–159. IEEE, 2021.

Filipe Campos, Liliana Petrychenko, Luís F Teixeira, and Wilson Silva. Latent diffusion models for privacy-preserving medical case-based explanations. In *CEUR Workshop Proceedings*, volume 3831. CEUR WS, 2024.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, pages 5253–5270, 2023.

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):10850–10869, 2023.

Viltė K. Dessers and Peggy Valcke. *The Right to Be Forgotten*, chapter 15, pages 179–192. John Wiley & Sons, Ltd, 2025. ISBN 9781394240821. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781394240821.ch15.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=gn0mIhQGNM.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017. doi: 10.1109/CVPR.2017.243.

Alistair Johnson, Matthew Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. MIMIC-CXR-JPG - chest radiographs with structured labels. *PhysioNet*, March 2024. doi: 10.13026/jsn5-t979. URL https://doi.org/10.13026/jsn5-t979. Version 2.1.0.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Guihong Li, Hsiang Hsu, Chun-Fu Chen, and Radu Marculescu. Machine Unlearning for Image-to-Image Generative Models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=9hjVoPWPnh.

Helena Montenegro and Jaime S Cardoso. A literature review on example-based explanations in medical image analysis. *Journal of Healthcare Informatics Research*, pages 1–49, 2025.

Gustav Müller-Franzes, Jan Moritz Niehues, Firas Khader, Soroosh Tayebi Arasteh, Christoph Haarburger, Christiane Kuhl, Tianci Wang, Tianyu Han, Teresa Nolte, Sven Nebelung, et al. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Scientific Reports (Sci Rep)*, 13(1):12098, 2023.