EGLOCE: Training-Free Energy-Guided Latent Optimization for Concept Erasure

Junyeong Ahn*, Seojin Yoon*, Sunyong Baik

01 Background

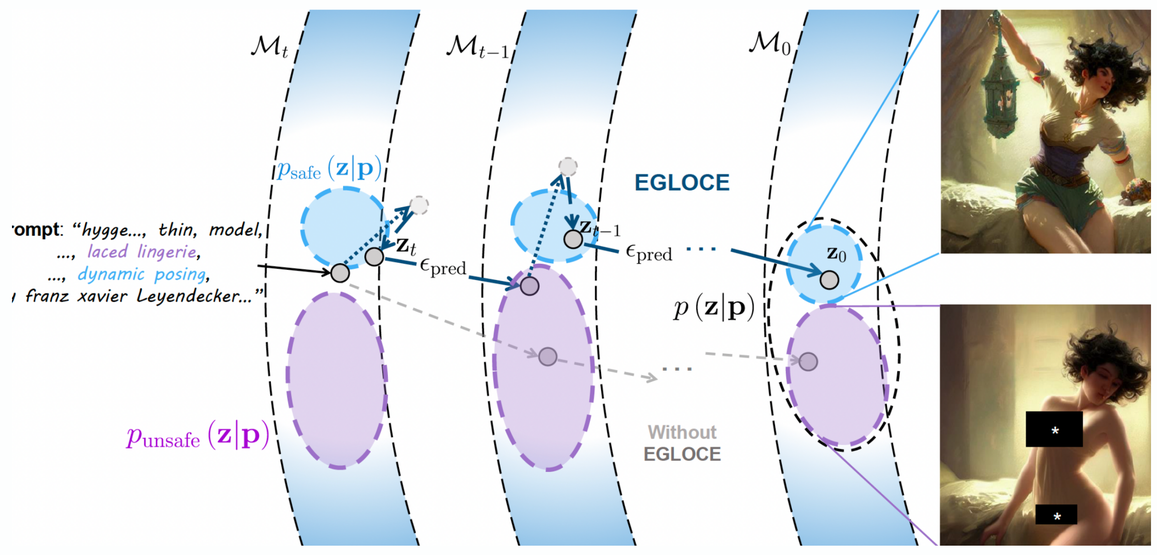
Unlearning is an AI technique that **removes unwanted concepts** from a model. In an era where cencorship concerns are increasingly important, it provides an efficient way to adjust and control pre-trained models. It can **reduce the influence of copyright data** and **remove harmful content** such as NSFW material, resulting in safer models. Research on unlearning also helps us better understand how models generate different types of content.

02 Objectives

- Unlearning must **remove targeted components** while **preserving the model's overall performance**. There is a fundamental trade-off between these two objectives.
- Even when a concept seems to be completely removed, it may still be **reconstructed through adversarial attacks**. Also, some factors are difficult to clearly characterize or summarize.
- Taking these factors into account, we mainly study unlearning in **T2I models**.

03 Motivation

Inference Trajectory Correction



- We focus on **correcting the diffusion sampling trajectory** at inference time, steering it away from unwanted concept regions while keeping it close to input prompt regions.

*Black boxes are manually added to generated images to cover explicit content.

04 Method

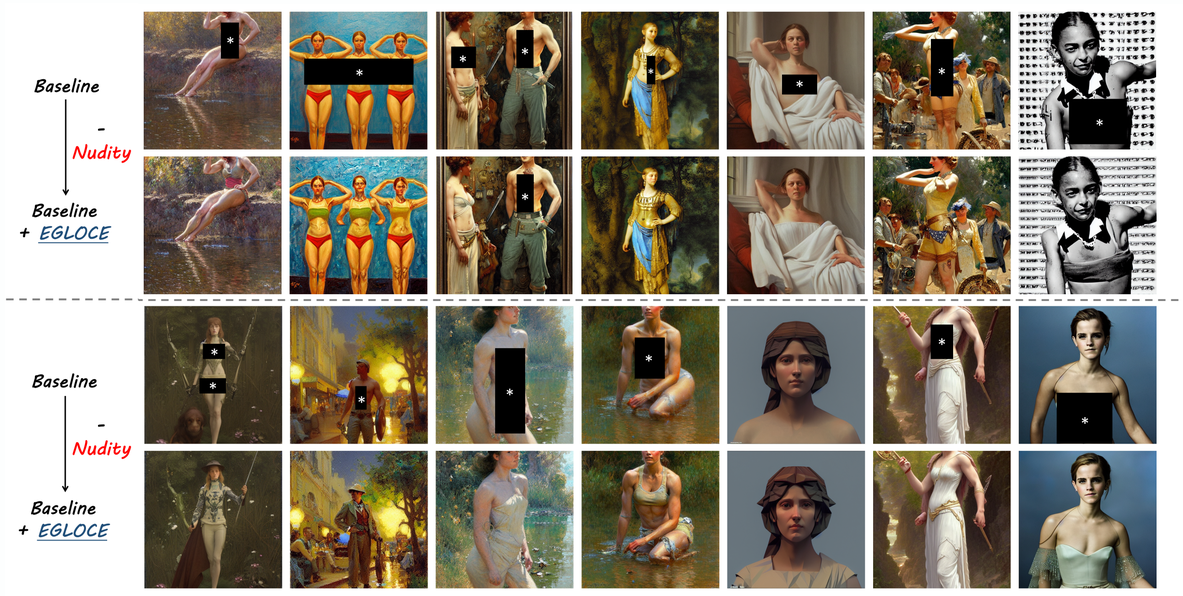
Energy-Guided Erasure Sampling

- Inspired by energy-based diffusion model (EDM) variants, which have only focused on better condition reflection, we introduce **repelling energy and retention energy**, where the former erases target concepts and the latter prevents generation from deviating from originally expected images.
- Through an extensive ablation over application timesteps, EGLOCE shows that optimizing latents in **the mid-to-late diffusion stages** is more effective than the early-to-mid stage application used in previous EDM variants.
- K-times iterations for optimization convergence based on *fixed-point iteration theory* in numerical analysis.

05 Results

We show that our method can be **plugged into SOTA baseline models** and consistently **improves performance**.

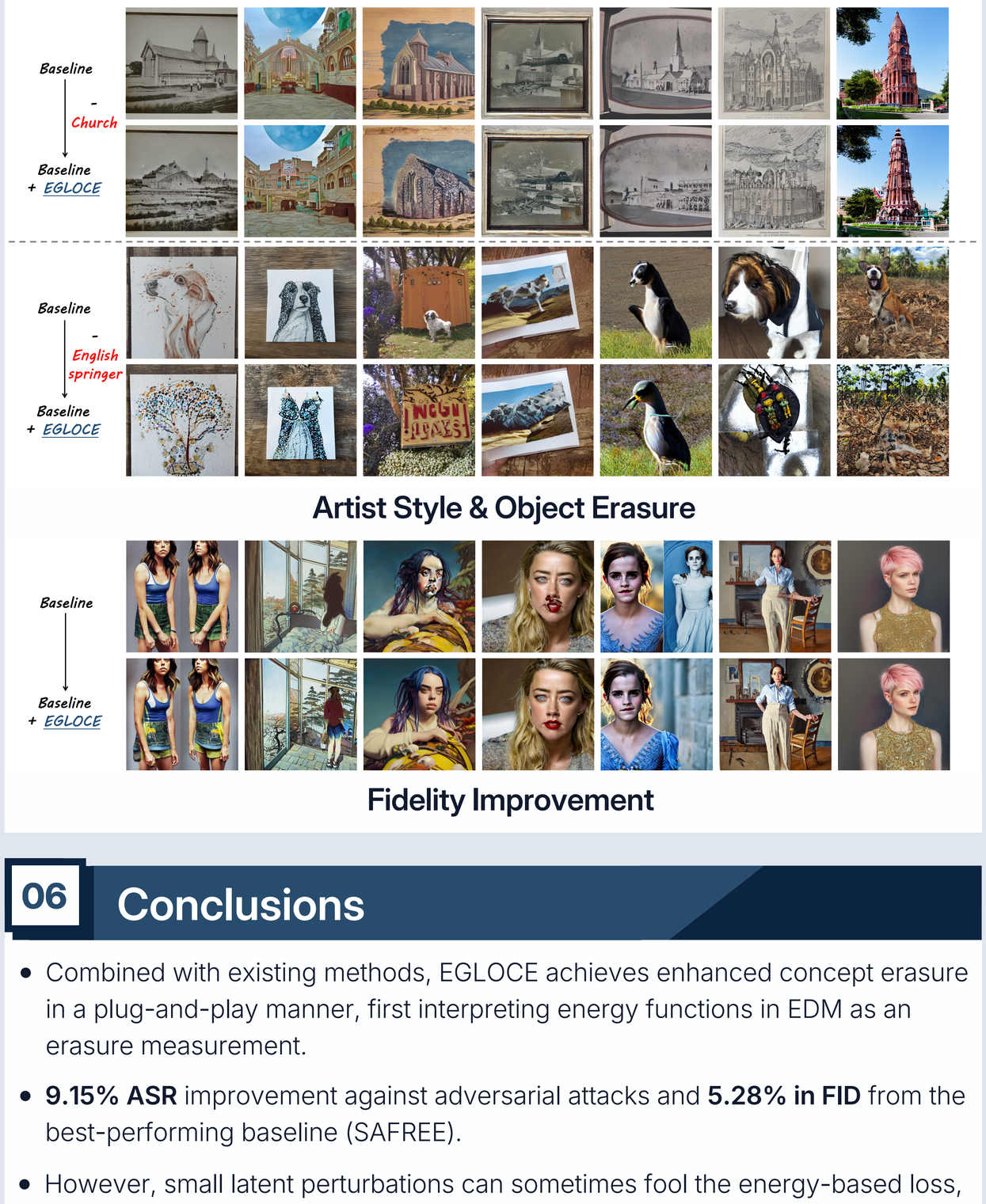
Method	Adversarial					COCO	
	I2P ↓	P4D ↓	Ring-A-Bell ↓	MMA-Diffusion ↓	UnlearnDiffAtk ↓	FID ↓	CLIP ↑
SLD [41]	0.396	0.920	0.557	0.912	**0.570**	18.77	30.79
SLD + Ours	**0.374**	**0.900**	**0.519**	**0.892**	**0.570**	**17.40**	**31.56**
RECE [15]	0.141	0.427	0.127	0.557	0.148	15.07	30.95
RECE + Ours	**0.092**	**0.313**	**0.025**	**0.489**	**0.120**	**13.65**	**31.56**
SAFREE [45]	0.106	0.400	0.165	0.529	0.211	17.12	30.90
SAFREE + Ours	**0.084**	**0.360**	**0.127**	**0.496**	**0.176**	**16.26**	**31.60**



Nudity Erasure

Method	Remove "Van Gogh"			Remove "Kelly McKernan"		
	LPIPS_e ↑	Acc_e ↓	Acc_u ↑	LPIPS_e ↑	Acc_e ↓	Acc_u ↑
SD-v1.4	–	1.00	1.00	–	1.00	1.00
SLD	0.62	1.00	1.00	0.54	1.00	1.00
SLD + Ours	0.61	1.00	1.00	0.53	1.00	1.00
RECE	0.59	1.00	1.00	0.55	1.00	1.00
RECE + Ours	0.60	1.00	1.00	0.57	0.50	1.00
SAFREE	0.69	0.05	0.85	0.59	0.20	1.00
SAFREE + Ours	0.68	0.00	0.88	0.61	0.30	1.00

Method	Church		English Springer	
	Acc_e ↓	Acc_u ↑	Acc_e ↓	Acc_u ↑
SLD	0.90	**1.00**	0.40	**1.00**
SLD + Ours	**0.70**	**1.00**	**0.10**	**1.00**
RECE	**0.00**	0.98	**0.00**	**0.93**
RECE + Ours	**0.00**	**1.00**	**0.00**	**0.93**
SAFREE	**0.60**	**1.00**	0.10	**0.93**
SAFREE + Ours	**0.60**	**1.00**	**0.00**	**0.93**



Artist Style & Object Erasure

Fidelity Improvement

06 Conclusions

- Combined with existing methods, EGLOCE achieves enhanced concept erasure in a plug-and-play manner, first interpreting energy functions in EDM as an erasure measurement.
- **9.15% ASR** improvement against adversarial attacks and **5.28% in FID** from the best-performing baseline (SAFREE).
- However, small latent perturbations can sometimes fool the energy-based loss, leaving the target concept visually intact. In future work, we plan to explore more robust energy functions to mitigate this issue and ensure faithful removal of the target concept.