Co-occurring Associated REtained concepts in Diffusion Unlearning

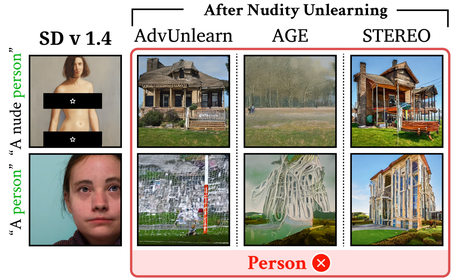
Miso Kim[1], Georu Lee[1], Yunji Kim[1], Hoki Kim[2*], Jinseong Park[3*], Woojin Lee[1*]
[1]Dongguk University, [2]Chung-Ang University, [3]Korea Institute for Advanced Study

ICLR

Motivation & Problem

Challenges of Existing Methods:

- Prior works focus on the trade-off between unlearning performance and generation quality
- However, they overlook that target concepts are removed together with co-occurring concepts



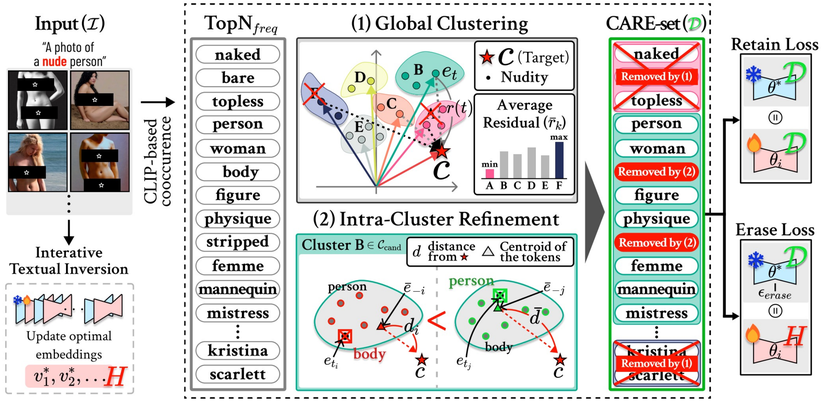
< Example: Unlearning '*nudity*' >

What is CARE? Co-occurring Associated REtained concepts

Removed Target	Co-occurence words	Benign co-occurrences
1 Nudity	~~Nude~~, ~~Naked~~, ~~Bare~~, Person, Woman, Figure, ..	Person, Woman, Figure, ..
2 VanGogh	~~Gogh~~, ~~Vincent~~, Stars, Galaxy, Moon, ..	Stars, Galaxy, Moon, ..
3 Tench	~~TincaTinca~~, ~~Doctorfish~~, Freshwater, Male, Sturgeon, ..	Freshwater, Male, Sturgeon, ..

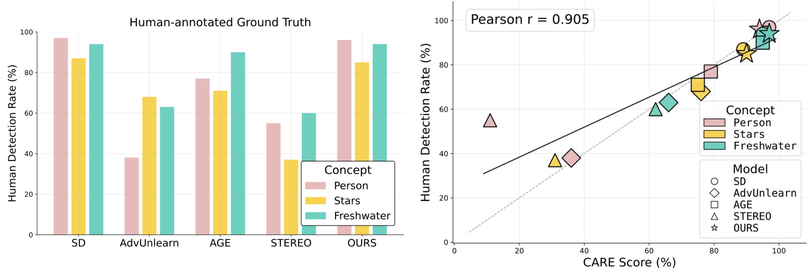
ReCARE Framework

ReCARE: Robust erasure for CARE



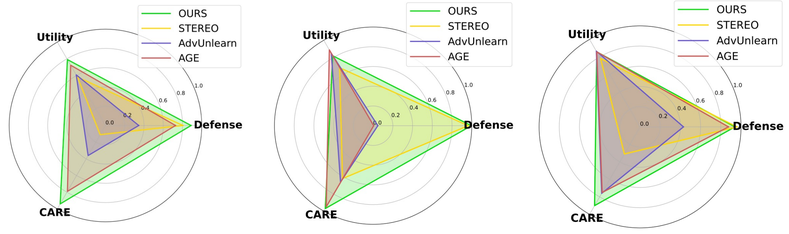
CARE Score: A New Evaluation Metric

- Measures CARE retention from generated outputs
- Strong correlation with human annotated ground truth



Evaluation & Results

Balanced Utility - Defense - CARE Trade-off



Quantitative Results