Decoding Vision Transformers: the Diffusion Steering Lens

Supplementary Material

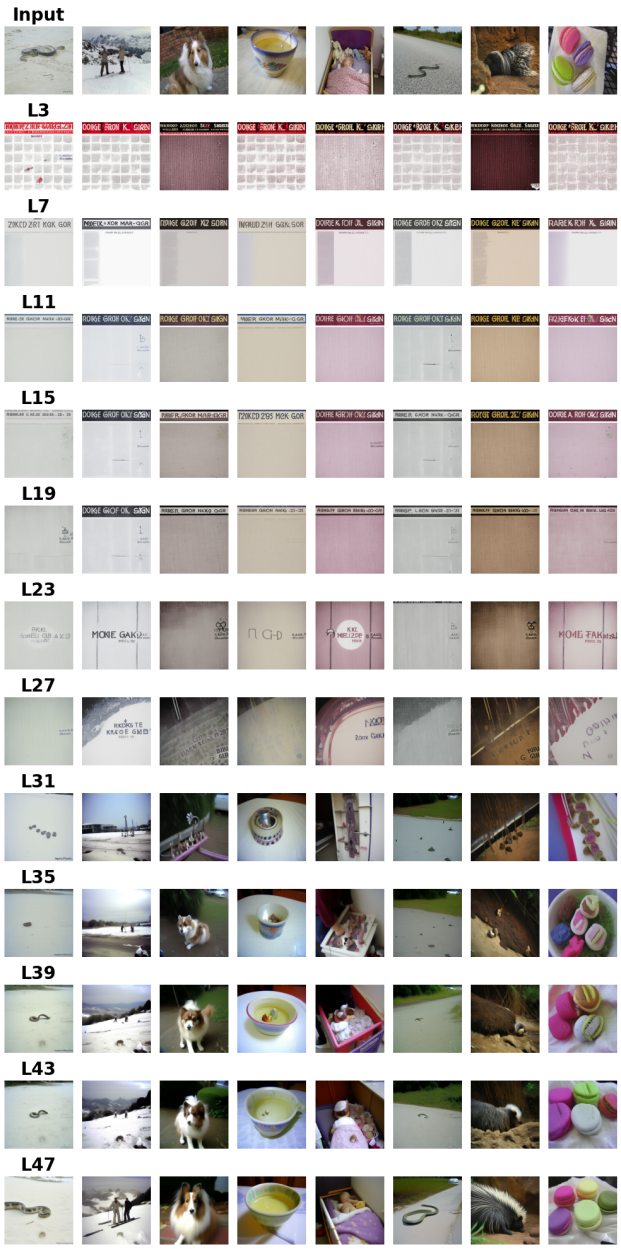

Figure 8. *Resid post visualizations using Diffusion Lens [43] (every 4 layer)*

Figure 9. *MLP out visualizations using Diffusion Lens [43] (every 4 layer)*

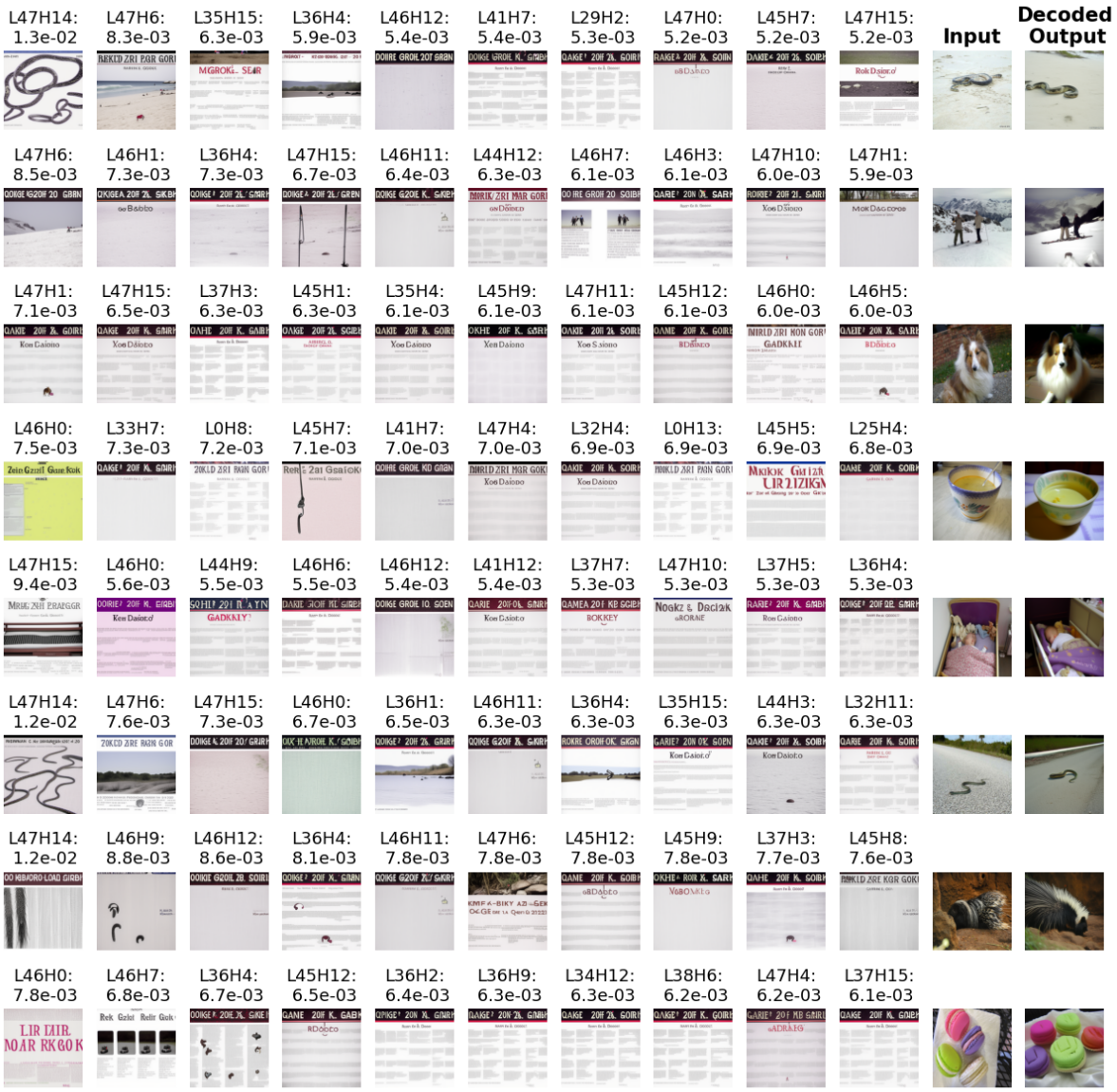

Figure 10. *Top 10 heads in similarity with input when using Diffusion Lens.*

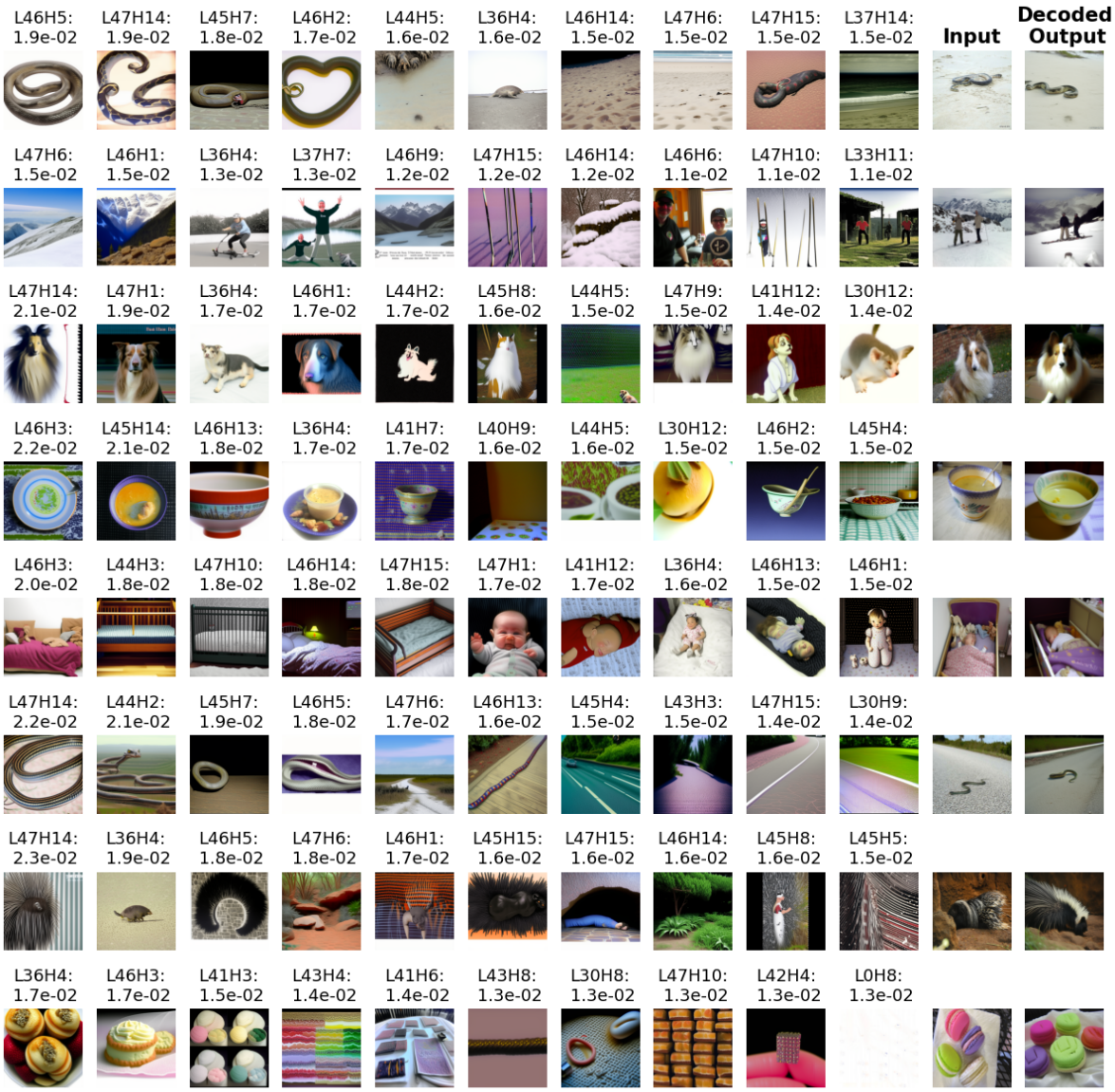

Figure 11. *Top 10 heads in similarity with input when using Diffusion Steering Lens.*

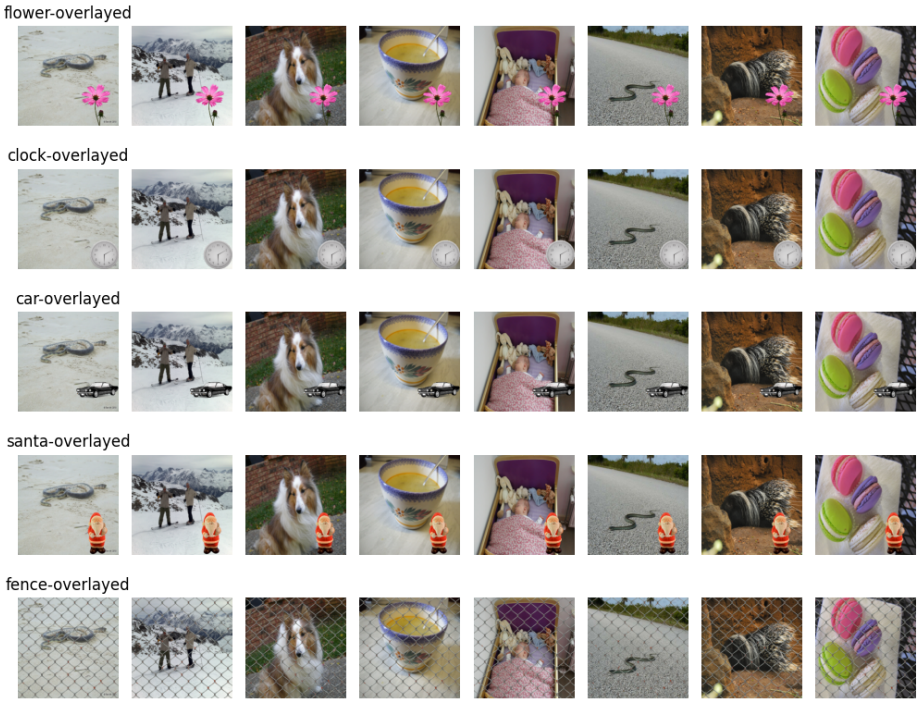

Figure 12. *Images with overlays.*

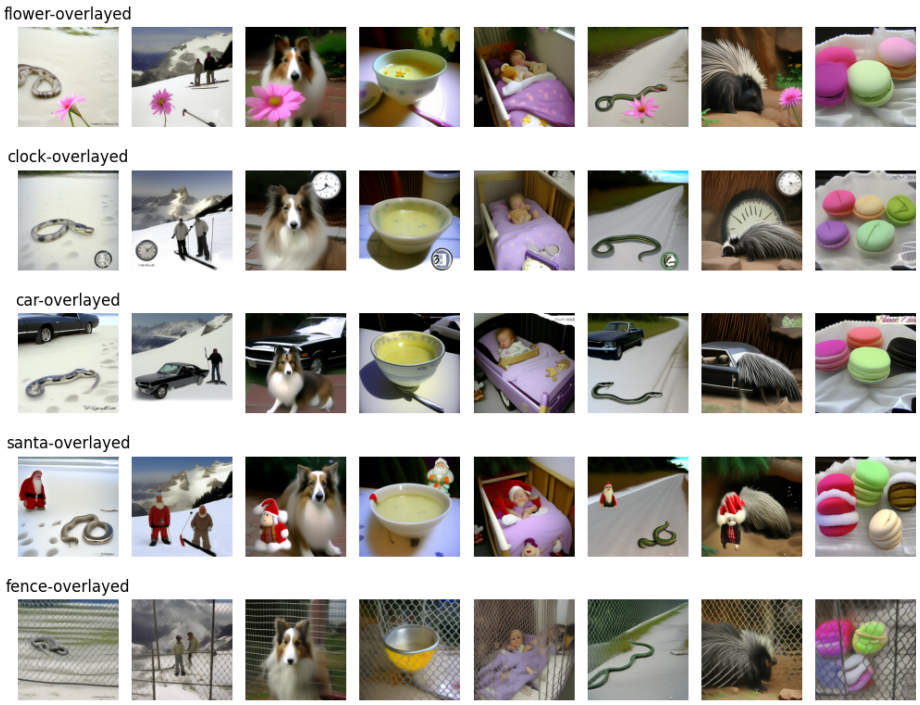

Figure 13. *Reconstruction results of images with overlays.*

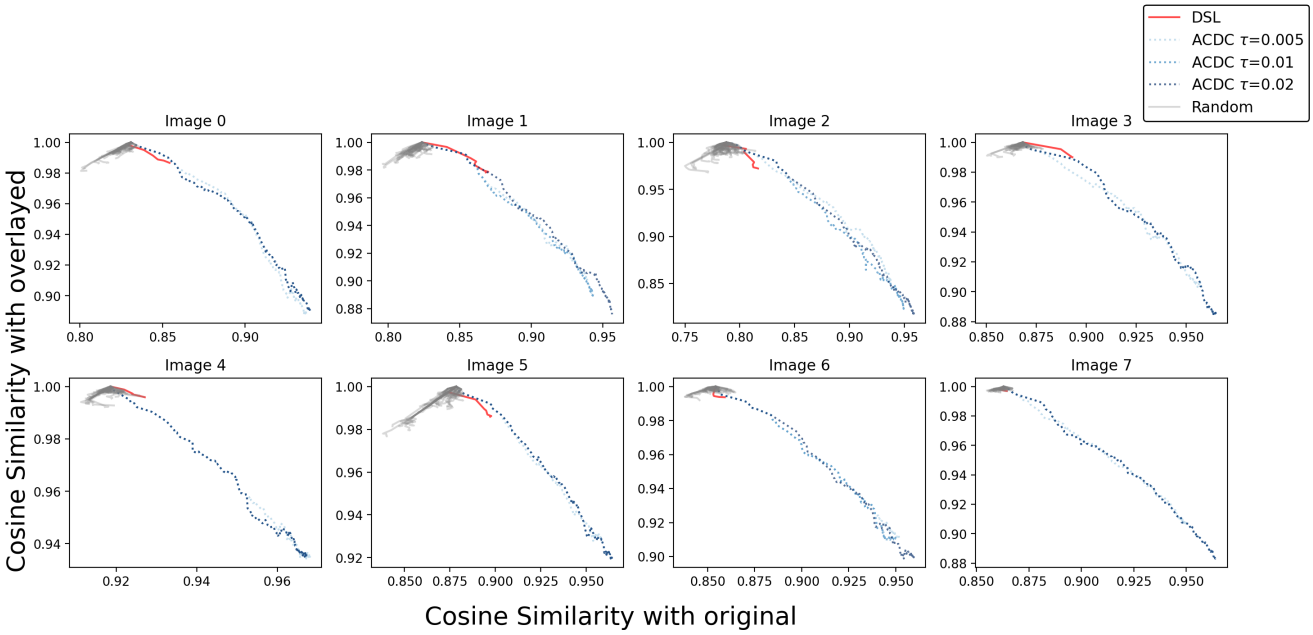

Figure 14. *Comparison of DSL, ACDC-like, and random ablation strategies for removing overlay information from clock-overlayed images.*

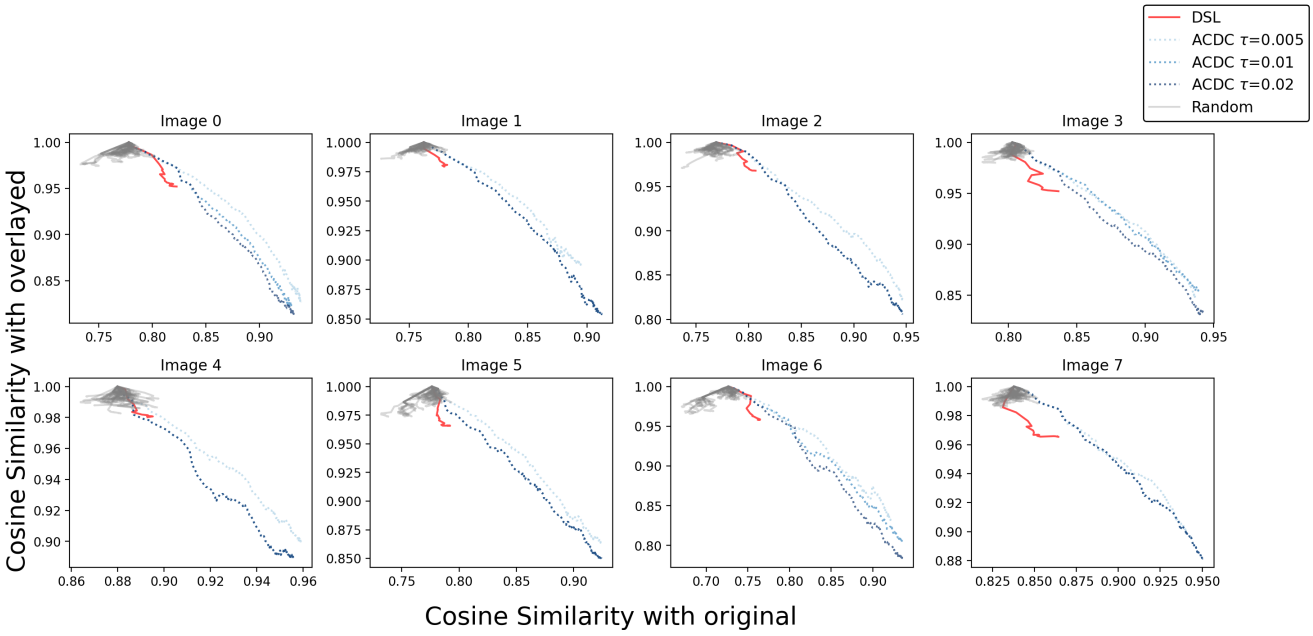

Figure 15. *Comparison of DSL, ACDC-like, and random ablation strategies for removing overlay information from car-overlayed images.*

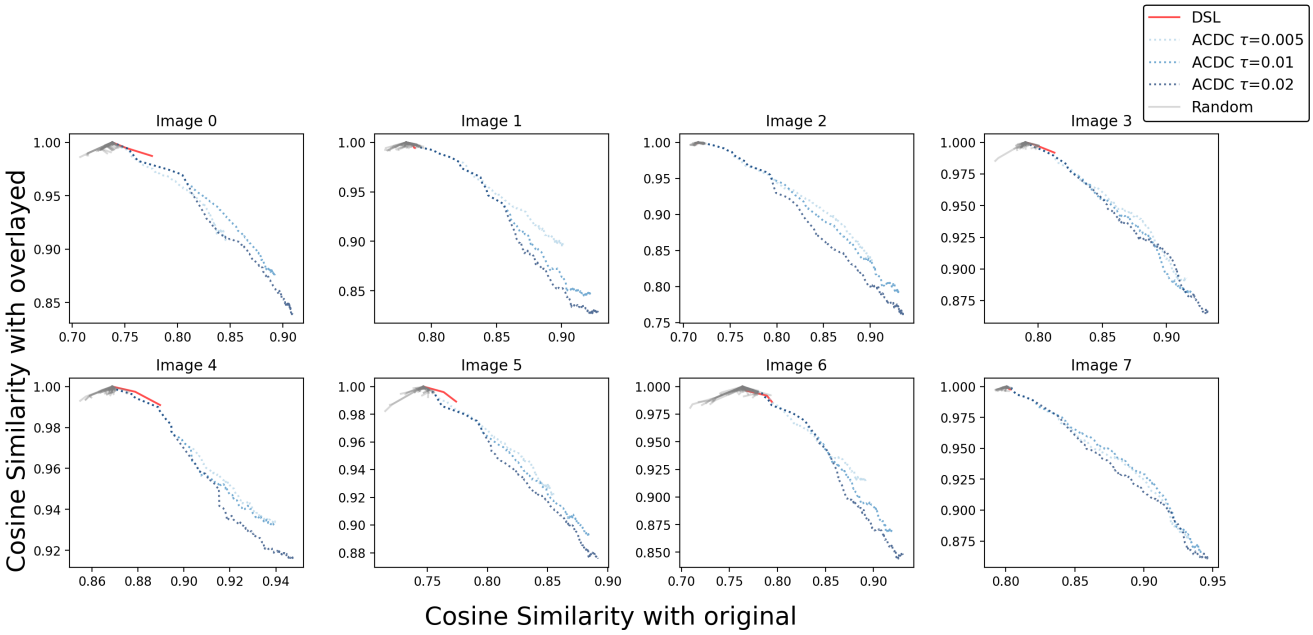

Figure 16. *Comparison of DSL, ACDC-like, and random ablation strategies for removing overlay information from santa-overlayed images.*

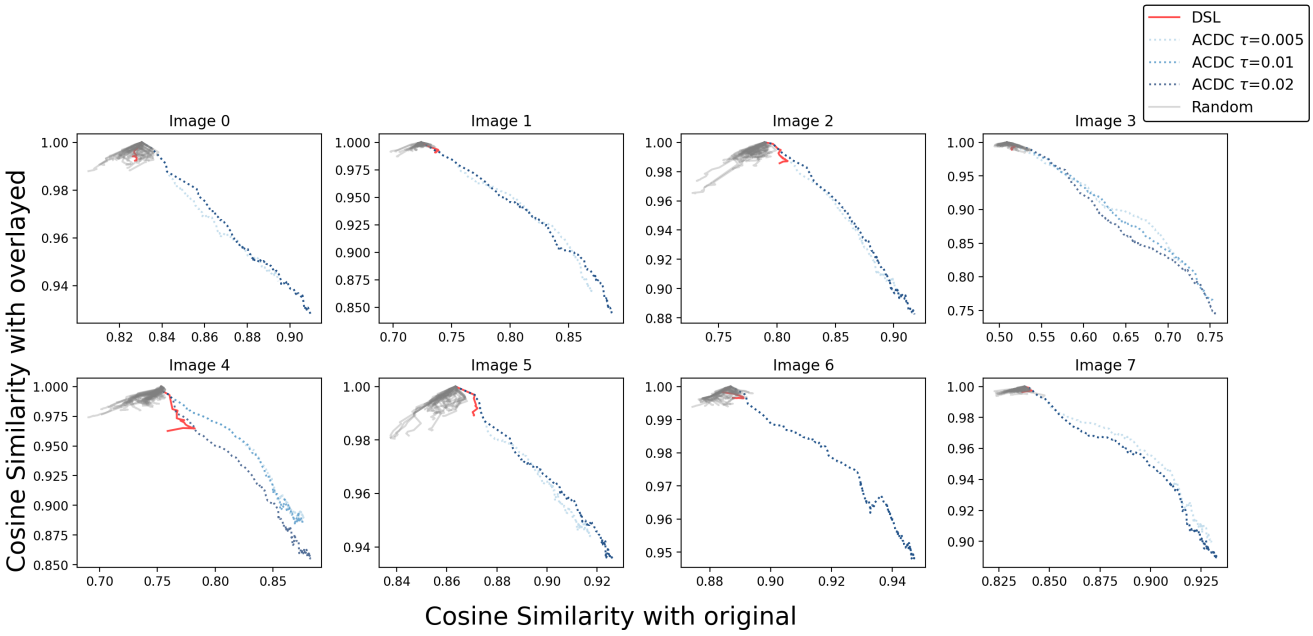

Figure 17. *Comparison of DSL, ACDC-like, and random ablation strategies for removing overlay information from fence-overlayed images.*