Counterfactual Generation Example: Eye Masking for Falsification Testing
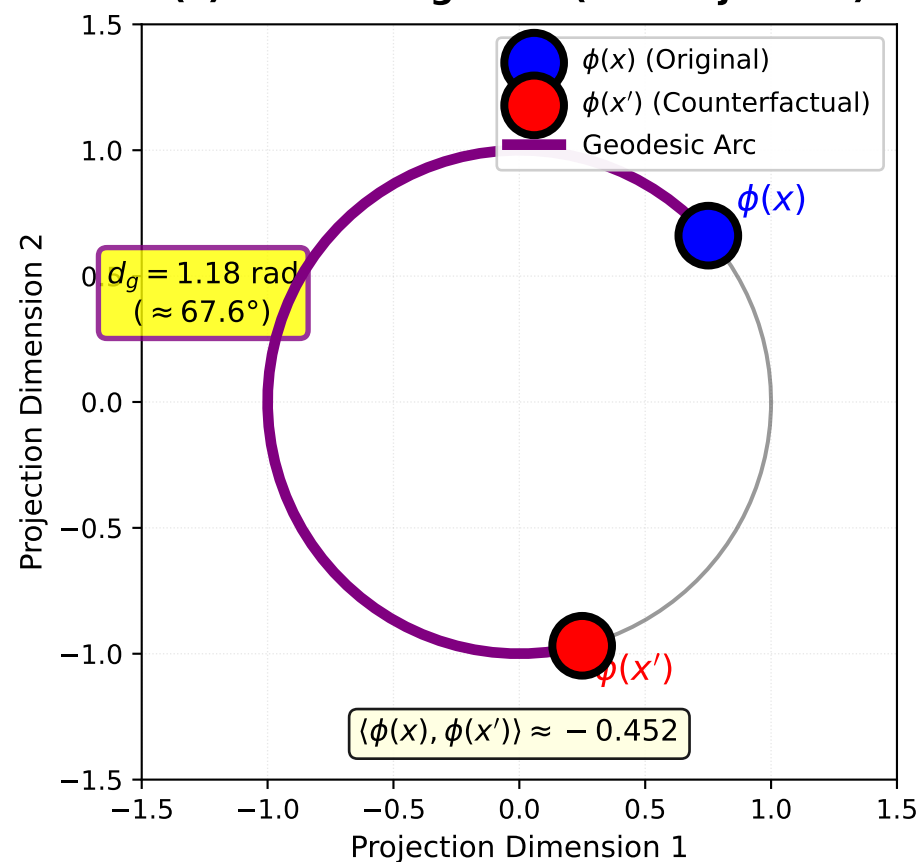
(a) Original Image with Grad-CAM

Eyes Region

$\phi_{eyes} = 0.82$ **(HIGH)**
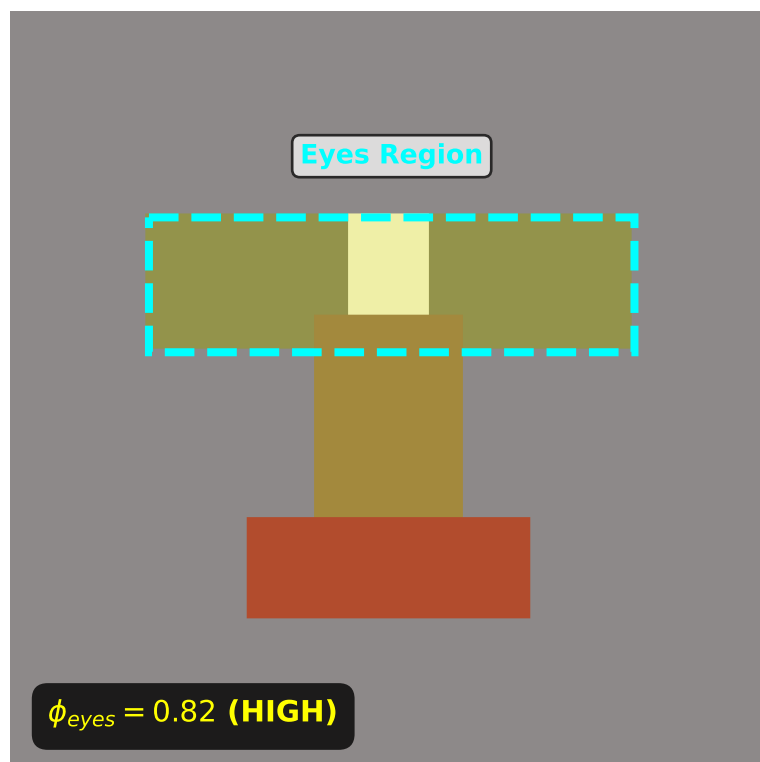
(b) Counterfactual (Eyes Masked)

$\|x' - x\|_2 = 0.145$

$d_g = 1.18$ **radians**
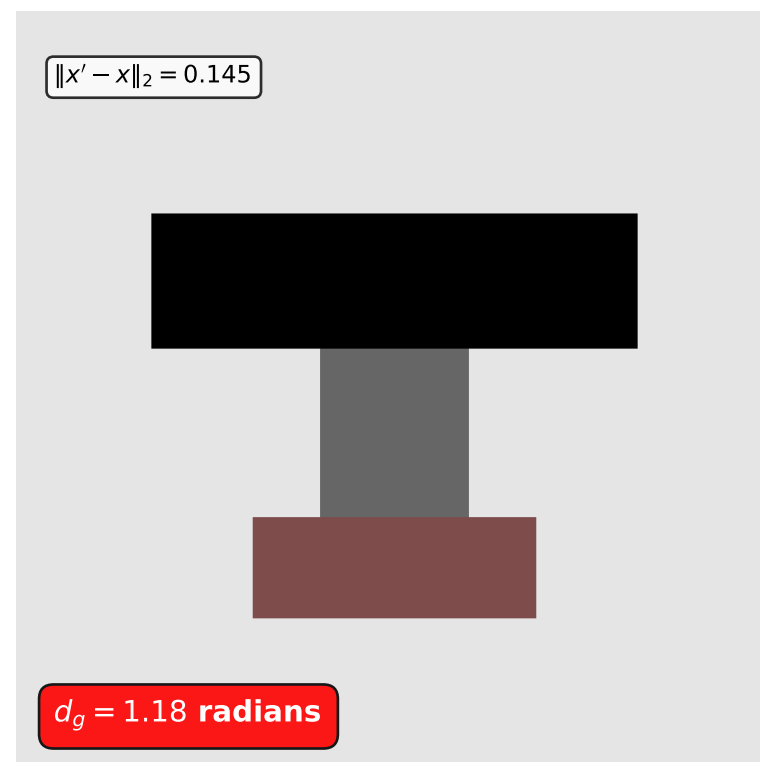
(c) Embedding Shift (2D Projection)

- $\phi(x)$ (Original)
- $\phi(x')$ (Counterfactual)
- Geodesic Arc

$\phi(x)$

$d_g = 1.18$ rad ($\approx 67.6°$)

$\phi(x')$

$\langle \phi(x), \phi(x') \rangle \approx -0.452$

Projection Dimension 1 / Projection Dimension 2

(d) Convergence of Algorithm 3.1

- $d_g(\phi(x), \phi(x'^{(t)}))$
- Target $\delta_{target} = 0.8$ rad
- Convergence ($t = 82$)
- Tolerance $\varepsilon_{tol} = 0.01$ rad

Converged $d_g = 0.802$

Geodesic Distance d_g (radians) / Iteration (t)