System Architecture: Falsification Testing Pipeline

7-layer architecture processing face images through attribution, counterfactual generation, and falsification testing

Layer 1-3:
Input & Attribution

Attribution Extraction
$\phi = A(x)$
Grad-CAM, SHAP,
LIME, IG

Attribution\n$\phi$

Feature Classification
S_{high}, S_{low}
$\theta_{high} = 0.7$
$\theta_{low} = 0.4$

Sets\n$S_h, S_l$

Layer 4:
Decision Gate

Non-triviality Check
$S_{high} \neq \varnothing$?
$S_{low} \neq \varnothing$?

PASS

FAIL

Layer 5-7:
Testing Pipeline

Counterfactual Generation
Algorithm 3.1
$K = 200$ **samples**
GPU batching

Counterfactuals\n$x', x''$

Geodesic Distance
Measurement
$\bar{d}_{high}, \bar{d}_{low}$
$d_g = \arccos(\langle \phi, \phi' \rangle)$

Distances\n$\bar{d}_h, \bar{d}_l$

Falsification Test
Theorem 3.5
3 conditions
statistical eval

Layer 8:
Verdict

Verdict
NOT FALSIFIED
or FALSIFIED

Runtime: ~4s per image (NVIDIA RTX 3090) | Bottleneck: Counterfactual generation (97.5%)