Evidence Thresholds for Explainable Face Verification: Counterfactual Faithfulness, Uncertainty, and Reporting

Aaron W. Storey, *Student Member, IEEE,* and Masudul H. Imtiaz, *Member, IEEE*

***Abstract*—Face verification systems deployed in forensic investigations rely increasingly on explainable AI (XAI) methods—Grad-CAM, SHAP, Integrated Gradients—to justify identification decisions with legal and civil liberty consequences. Yet these explanations lack a critical property: falsifiability. When Grad-CAM highlights the eye region as driving a match, practitioners have no principled method to test this claim. We address this gap through an operational validation protocol treating attribution faithfulness as an empirically testable hypothesis. If an explanation correctly identifies causal features, then perturbing those features should produce predictable changes in verification scores—a counterfactual prediction we can measure.**

This article presents three contributions for forensic face recognition. First, we provide a systematic five-step falsification protocol producing binary verdicts: "NOT FALSIFIED" (attributions align with model behavior) or "FALSIFIED" (contradictory evidence). The protocol includes statistical hypothesis testing with Bonferroni correction and plausibility gates (LPIPS < 0.3, FID < 50) ensuring counterfactuals remain perceptually realistic. Second, we establish pre-registered validation thresholds—frozen before experimental execution to prevent post-hoc adjustment—including geodesic distance correlation floors ($\rho > 0.7$) and confidence interval calibration ranges (90–100% coverage). Third, we provide a forensic reporting template with seven standardized fields designed to meet Daubert admissibility standards, operationalizing requirements from the EU AI Act (Articles 13–15), GDPR (Article 22), and U.S. Federal Rules of Evidence (Rule 702).

Our protocol bridges the gap between current XAI evaluation practices (subjective interpretability, proxy metrics) and evidentiary requirements (testability, known error rates, objective standards). The systematic methodology enables forensic analysts to validate explanations scientifically while meeting regulatory transparency mandates. Pre-registration ensures scientific integrity; the forensic template enables transparent communication to legal professionals and oversight bodies.

***Index Terms*—Face Verification, Explainable AI, Forensic Science, Evidence Standards, Counterfactual Validation, Attribution Faithfulness, Pre-registration, Protocol Validation**

I. Introduction

FACE verification systems have become integral to forensic investigations, border security, and criminal proceedings. Their deployment is widespread—documented in law enforcement agencies across North America, Europe, and Asia. Yet multiple wrongful arrests demonstrate that algorithmic errors carry severe real-world consequences. In Detroit alone, Robert Williams (2020) and Porcha Woodruff (2023) were arrested based on false facial recognition matches [1], [2]. Nijeer Parks spent ten days in jail in New Jersey (2019) after a misidentification [3]. These failures affect fundamental civil liberties: freedom from unlawful arrest, the right to contest evidence, access to due process.

When face verification systems contribute to criminal convictions or refugee status determinations, a seemingly technical question becomes legally critical: *which facial features drove this decision?* This isn't academic curiosity. Under the Daubert standard (U.S. Federal Rules of Evidence, Rule 702), expert testimony must rest on testable methods with known error rates [4]. The EU AI Act (2024) mandates that high-risk biometric systems provide "transparent and comprehensible" information about decision-making processes [5]. GDPR Article 22 requires "meaningful information about the logic involved" in automated decisions significantly affecting individuals [6]. Legal frameworks converge on a single requirement: explanations must be *validated*, not merely generated.

A. The Falsifiability Gap in Current XAI Practice

Current explainable AI (XAI) methods produce visual explanations—Grad-CAM highlights the eye region, SHAP assigns high importance to the nose, LIME emphasizes cheekbones [7]–[9]. These saliency maps appear plausible. They align with human intuition ("of course eyes matter for identification"). But forensic analysts and legal professionals have no principled method to *test* these claims. When Grad-CAM highlights the forehead as critical for a match, how do we know this attribution is faithful rather than a post-hoc rationalization?

Traditional XAI evaluation metrics fall short on three counts. First, *insertion-deletion curves* [10] systematically remove or add pixels, creating out-of-distribution samples that elicit unreliable model behavior. The metric assumes linearity—that removing 50% of pixels changes the score proportionally—but deep networks exhibit highly nonlinear responses. Second, *localization accuracy* [11] requires ground truth annotations ("the nose is the true important region"), but for face verification, no such ground truth exists. We don't know which features a trained ResNet-100 actually uses; that's

A. W. Storey and M. H. Imtiaz are with the Department of Computer Science, Clarkson University, Potsdam, NY 13699, USA (e-mail: storeyaw@clarkson.edu; mimtiaz@clarkson.edu).

Manuscript received October 15, 2025; revised XXX XX, 2026.

precisely what we're trying to discover. Third, *consistency checks* [12] measure whether attributions change when model weights are randomized, but this tests method sensitivity, not faithfulness.

The core problem is more fundamental: these metrics provide relative comparisons between methods, not absolute validation of correctness. Grad-CAM might score higher than LIME on insertion-deletion, but does that mean Grad-CAM is *faithful*, or merely *more faithful than a weak baseline*? For forensic deployment—where explanations influence pretrial detention, sentencing, and appeals—we need stronger evidence. We need falsifiability.

B. Our Approach: Counterfactual Prediction as Empirical Test

We address the falsifiability gap by treating attribution faithfulness as a testable hypothesis. The core idea is simple: if an attribution method correctly identifies features responsible for a verification decision, then perturbing those features in controlled ways should produce *predictable changes* in similarity scores. This is a counterfactual prediction: "If I mask the high-attribution features (eyes, nose) while preserving low-attribution features (background, hair), the embedding should shift by at least τ_{high} radians on the unit hypersphere. If I mask only low-attribution features, the shift should be smaller—at most τ_{low} radians."

This prediction is empirically testable. We generate counterfactual images through gradient-based optimization, measure geodesic distances in embedding space, and compare predicted shifts to observed shifts. If predictions align with observations across many test cases, the attribution receives verdict "NOT FALSIFIED." If predictions systematically fail, verdict is "FALSIFIED." Critically, this isn't proof of correctness (Popper's falsification criterion forbids such claims [13]), but rather provisional acceptance: the attribution has survived rigorous testing.

Our protocol differs from prior work in three ways. First, we validate on the *decision manifold*—the unit hypersphere where face verification actually operates—rather than in pixel space where insertion-deletion lives. ArcFace and CosFace normalize embeddings to unit L2 norm, making angular (geodesic) distance the natural similarity metric [14], [15]. Testing attributions in this space respects the geometry of verification. Second, we enforce *plausibility gates*: counterfactuals must maintain perceptual similarity (LPIPS < 0.3) and distributional similarity (FID < 50) to natural faces [16], [17]. This prevents adversarial perturbations—which would yield large score changes but misleading validation. Third, we *pre-register thresholds* before experimental execution, freezing decision criteria to prevent p-hacking [18].

C. Contributions

This article makes three primary contributions to forensic face recognition and explainable AI:

C1: Operational Falsification Protocol (Section III). We present a systematic five-step procedure implementing the falsifiability criterion. The protocol takes as input an image pair, a face verification model, and an attribution method, then produces a binary verdict: "NOT FALSIFIED" or "FALSIFIED." The procedure includes (1) attribution extraction using standard XAI methods (Grad-CAM, SHAP, LIME, Integrated Gradients), (2) feature classification into high-importance and low-importance sets, (3) counterfactual generation via gradient descent on the hypersphere, (4) geodesic distance measurement, and (5) statistical hypothesis testing with Bonferroni correction. Each step specifies exact hyperparameters (learning rates, sample sizes, convergence criteria) for reproducibility.

C2: Pre-Registered Validation Endpoints (Section IV). We establish quantitative thresholds for primary and secondary validation endpoints, frozen before experimental execution. The primary endpoint is Pearson correlation between predicted and observed geodesic distance changes ($\rho > 0.7$ required for passage). The secondary endpoint is confidence interval calibration (90–100% empirical coverage of 90% CIs). We also specify plausibility gates: LPIPS < 0.3 for perceptual similarity, FID < 50 for distributional similarity. These thresholds are justified through published psychometric standards [19], prediction theory [20], pilot experiments on a separate calibration set, and forensic science precedents (DNA match probabilities, fingerprint point minima). Critically, we timestamp this document and generate a cryptographic hash before testing, preventing retroactive adjustment.

C3: Forensic Reporting Template (Section V). We provide a seven-field standardized template for documenting validation results in legal contexts. The template addresses Daubert's four prongs: (1) testability (demonstrated through counterfactual prediction), (2) peer review (method published in this article), (3) known error rates (Field 5: falsification rates stratified by demographics and imaging conditions), (4) general acceptance (to be established through future adoption). The template also operationalizes EU AI Act Article 13 requirements (accuracy metrics, transparency), GDPR Article 22 (meaningful information about logic), and forensic science standards (objective criteria, proficiency testing). Each field includes specific data to record, justification requirements, and interpretation guidance. We demonstrate template completion through hypothetical examples: one "APPROVED with RESTRICTIONS" scenario showing moderate performance with demographic disparities, and one "NOT APPROVED" scenario illustrating failure to meet correlation thresholds.

D. Regulatory and Legal Context

Our protocol design is informed by three converging frameworks mandating validated explanations:

EU AI Act (2024), Articles 13–15. Biometric identification systems are classified as high-risk AI (Annex III, Point 1(a)), requiring "the level of accuracy, robustness and cybersecurity... together with any known and foreseeable circumstances that may have an impact" (Art. 13(3)(d)) [5]. Our protocol operationalizes this through correlation metrics (ρ, R^2), mean absolute error (MAE), and stratified performance reporting. Article 15 demands technical documentation including "the methods and steps performed for the validation of the AI system." Our five-step protocol and forensic template provide this documentation structure.

GDPR Article 22 (2016). Automated decisions "which produce legal effects concerning [a person] or similarly significantly affect" them require safeguards including contestation rights [6]. Legal scholars debate whether Article 22 mandates a right to explanation [21], [22]. Regardless, when explanations *are* provided (increasingly common under AI Act pressure), they must be accurate. Providing misleading attributions while claiming GDPR compliance would violate Article 5(1)(a)'s transparency principle. Our uncertainty quantification (90% CIs, calibration coverage) enables meaningful contestation by revealing prediction reliability.

U.S. Daubert Standard (1993). Federal Rule of Evidence 702 requires that expert scientific testimony employ "reliable principles and methods" applied reliably to the facts [4]. The *Daubert v. Merrell Dow Pharmaceuticals* precedent established four reliability factors: testability, peer review, error rates, and general acceptance. When facial recognition evidence is presented in criminal proceedings—matching a defendant's photo to surveillance footage—explanations of *why* the match occurred fall under this standard. They constitute scientific claims requiring validation. Documented wrongful arrests (Williams, Woodruff, Parks) demonstrate failures where validated explanations could have enabled earlier error detection by revealing implausible attributions (e.g., high importance on backgrounds rather than facial features).

The convergence of these frameworks creates legal pressure for scientifically validated explanations. This article provides the technical methodology to meet regulatory requirements while maintaining scientific rigor.

E. Article Organization

Section II condenses evidentiary requirements from AI regulation and forensic science, showing how current XAI practices fail to meet these standards. Section III presents the operational validation protocol in implementable detail, specifying algorithms, hyperparameters, and computational requirements. Section IV justifies pre-registered endpoints and decision thresholds through published standards and pilot data. Section V provides the forensic reporting template with field-by-field completion guidance. Section VI analyzes threats to validity, computational constraints, and demographic fairness risks. Sections VII and VIII (experimental results and discussion) will be completed after empirical validation on benchmark datasets (LFW, CelebA) using ArcFace and CosFace models.

II. Background: Evidentiary Requirements for Automated Decision Systems

A. Regulatory Landscape for High-Risk Biometric Systems

Three major frameworks establish requirements for explainability and validation in forensically deployed face verification systems. Rather than recount each statute in exhaustive detail, we focus on the specific technical obligations they create—and where current XAI practices fall short.

1) EU AI Act (2024): Technical Documentation Mandates: The European Union's AI Act classifies biometric identification systems as high-risk (Annex III, Point 1(a)), triggering stringent oversight [5]. Two articles directly impact attribution validation:

Article 13(3)(d): Systems must provide "the level of accuracy, robustness and cybersecurity... together with any known and foreseeable circumstances that may have an impact on that expected level." This isn't a vague transparency aspiration. It's a legal mandate for *quantitative accuracy metrics*. Our protocol delivers these through correlation coefficients (ρ, R^2), mean absolute error (MAE $\pm$ SD), and stratified performance tables showing how accuracy degrades under poor lighting, extreme poses, or demographic shifts.

Article 15(1): Technical documentation must include "the methods and steps performed for the validation of the AI system." Generating saliency maps isn't enough. Deployers must document *how* they validated those maps. Our five-step falsification protocol (Section III) provides this documentation structure: extraction $\rightarrow$ classification $\rightarrow$ perturbation $\rightarrow$ measurement $\rightarrow$ testing.

These provisions create obligations that current XAI methods cannot meet. Grad-CAM produces heatmaps but offers no validation procedure. SHAP computes Shapley values but provides no accuracy guarantee. The AI Act demands more.

2) GDPR Article 22: Contestation Rights: GDPR Article 22(1) establishes the right not to be subject to solely automated decisions producing legal effects [6]. Article 22(3) requires "suitable measures to safeguard... rights and freedoms," including "the right... to contest the decision." Legal scholars vigorously debate whether this implies a right to explanation [21], [22]. Wachter et al. argue GDPR mandates only information about system logic, not specific decision rationale. Selbst and Powles counter that meaningful contestation *requires* understanding which factors influenced the outcome.

We sidestep this debate by observing a simpler point: when explanations *are* provided—as increasingly required by the AI Act—they must be accurate. Imagine a forensic report stating, "The system matched these faces based primarily on nose structure," when in reality the model relied on background artifacts. This would constitute a GDPR violation under Article 5(1)(a)'s lawfulness and transparency principles. Misleading explanations prevent genuine contestation.

Our protocol's contribution here is *uncertainty quantification*. Field 4 of the forensic template (Section V) reports calibration coverage: what percentage of predictions fall within stated confidence intervals? If we claim 90% confidence and only 75% of observations land in the predicted range, our uncertainty estimates are overconfident—unreliable for contestation. Well-calibrated intervals (empirical coverage $\approx$ 90%) enable data subjects to meaningfully challenge predictions: "Your explanation predicts this feature mattered with 90% confidence, but when tested..."

3) U.S. Daubert Standard: Four Reliability Prongs: In U.S. federal courts, scientific expert testimony must satisfy Federal Rule of Evidence 702 and the *Daubert* precedent [4]. Four factors assess reliability:

1) **Testability:** Can the theory or technique be tested? Is it falsifiable?
2) **Peer Review:** Has the method been subjected to peer review and publication?
3) **Error Rates:** What are the known or potential rates of error?
4) **General Acceptance:** Is the method generally accepted in the relevant scientific community?

When facial recognition evidence is presented in criminal proceedings, explanations fall under this standard. They're scientific claims about causation ("feature X drove decision Y"), requiring validation. The documented wrongful arrests—Williams, Woodruff, Parks—demonstrate real failures. In Williams' case, Detroit police relied on a facial recognition match without scrutinizing the explanation. Had validated attributions revealed high importance assigned to jacket color or background scenery rather than facial structure, examiners might have caught the error earlier.

Our protocol addresses all four Daubert prongs. *Testability*: counterfactual predictions are empirically falsifiable (Section III). *Peer review*: this article constitutes publication; code will be released for community validation. *Error rates*: Field 5 of the forensic template reports falsification rates stratified by demographics and imaging conditions—exactly the "known error rates" Daubert demands. *General acceptance*: to be established through future adoption, but the method builds on accepted techniques (gradient descent, statistical hypothesis testing, established perceptual metrics).

B. Forensic Science Standards for Tool Validation

The National Research Council's 2009 report *Strengthening Forensic Science in the United States* criticized forensic disciplines lacking objective standards, known error rates, and proficiency testing [23]. The report's key insight: forensic methods must undergo rigorous scientific validation *before* deployment. Subjective examiner judgment isn't enough.

Face verification systems, when used forensically, must meet these standards. Yet current XAI methods lack all three pillars:

Objective Standards. No consensus exists on when an explanation is "faithful enough." Researchers report insertion-deletion scores or pointing game accuracy, but what threshold constitutes adequacy? 70% localization? 80%? The field has no agreed benchmark.

Known Error Rates. How often do Grad-CAM attributions misidentify causal features? No systematic measurement exists. We have relative comparisons ("Grad-CAM outperforms LIME on metric M"), but not absolute error quantification ("Grad-CAM falsifies on 38% of test cases with known demographic disparities").

Proficiency Testing. DNA analysts undergo blind proficiency tests—samples with known ground truth to assess examiner reliability [23]. Attribution methods face no such testing. Our protocol provides the framework: run falsification tests on benchmark datasets (LFW, CelebA), report passage rates, stratify by subgroups.

TABLE I
GAP BETWEEN EVIDENTIARY REQUIREMENTS AND CURRENT XAI PRACTICE

Requirement	Current Practice	Protocol Contribution
Testability (Daubert)	Subjective interpretability	Falsifiable counterfactual predictions (Sec. III)
Known Error Rates (Daubert, NRC)	Relative method comparisons	Statistical tests with p-values, stratified failure rates (Sec. V, Field 5)
Accuracy Metrics (AI Act Art. 13)	Proxy metrics (insertion-deletion)	Direct geodesic distance correlation ρ (Sec. IV)
Validation Docs (AI Act Art. 15)	Ad-hoc reporting	Standardized seven-field template (Sec. V)
Objective Standards (NRC 2009)	Researcher-dependent thresholds	Pre-registered frozen thresholds (Sec. IV)
Contestation (GDPR Art. 22)	Static saliency maps	Uncertainty-quantified predictions with calibration (Sec. V, Field 4)

The NRC report also established forensic science's foundational principle: *every statement must be scientifically defensible*. Presenting a saliency map in court testimony without validation violates this principle. Our forensic template (Section V) operationalizes defensibility through structured documentation.

C. Gap Analysis: What's Missing?

Table I summarizes the disconnect between evidentiary requirements and current XAI practice.

The gap is clear: existing methods generate explanations but don't validate them. Insertion-deletion curves, localization metrics, consistency checks—all provide *relative* quality assessments. None answer the binary question a forensic examiner needs: "Can I trust this attribution, yes or no?" Our protocol provides that binary verdict through rigorous testing.

Consider a concrete example. A forensic analyst examines a match between a suspect photo and surveillance footage. Grad-CAM highlights the nose and upper lip. Is this attribution faithful? Under current practice, the analyst has only intuition ("seems reasonable, noses do vary between individuals"). Our protocol offers empirical testing: generate 200 counterfactuals masking the nose region, measure embedding shifts, compare to predictions. If observed shifts align with attribution-based predictions ($\rho > 0.7$, $p < 0.05$), the attribution survives falsification. If not, it's unreliable—and the analyst knows to seek alternative evidence.

This is the core contribution: shifting attribution validation from subjective plausibility assessment to empirical hypothesis testing. The next section details how.

III. OPERATIONAL VALIDATION PROTOCOL

A. Protocol Overview

The falsification testing protocol implements three conditions that attributions must satisfy to be deemed "NOT FALSIFIED":

Condition 1 (Non-Triviality): The attribution must identify both high-importance features (set S_{high}) and low-importance features (set S_{low}), with both sets non-empty. Flat attributions assigning uniform importance provide no testable predictions.

Condition 2 (Differential Prediction): Counterfactual perturbations targeting high-attribution features must cause larger geodesic embedding shifts than perturbations targeting low-attribution features. Formally: $\mathbb{E}[d_{\text{high}}] > \tau_{\text{high}}$ and $\mathbb{E}[d_{\text{low}}] < \tau_{\text{low}}$ where d is geodesic distance and τ are pre-registered thresholds.

Condition 3 (Separation Margin): Thresholds must be sufficiently separated: $\tau_{\text{high}} > \tau_{\text{low}} + \epsilon$ for margin $\epsilon > 0$, ensuring meaningful distinction rather than arbitrary cutoffs.

If all three conditions hold with statistical significance ($\alpha = 0.05$, Bonferroni-corrected), verdict is "NOT FALSIFIED." If any condition fails, verdict is "FALSIFIED" with specific failure mode reported.

B. Step 1: Attribution Extraction

Input: Image pair (x, x'), face verification model f, attribution method $\mathcal{A}$

Output: Attribution map $\phi \in \mathbb{R}^m$ where m is the number of features

We support four standard attribution methods, selected for their prevalence in forensic and research contexts:

Grad-CAM (Gradient-Weighted Class Activation Mapping): Computes gradients of embedding output with respect to final convolutional layer activations [7]. For ArcFace with ResNet-100 backbone, we extract from `conv5_3` layer, producing a 7×7 spatial heatmap ($m = 49$ features). This is the de facto standard for visual explanations in face recognition.

SHAP (SHapley Additive exPlanations): Approximates Shapley values using KernelSHAP with 1,000 coalition samples [8]. We segment the image into $m = 50$ superpixels via Quickshift algorithm [24]. The baseline is a black image (all zeros). SHAP provides game-theoretic guarantees but at substantial computational cost (typically 100× slower than Grad-CAM).

LIME (Local Interpretable Model-Agnostic Explanations): Fits a local linear model using 1,000 perturbed samples [9]. Uses superpixel segmentation ($m = 50$ segments). Coefficients provide feature importance. LIME trades theoretical rigor for model-agnostic applicability.

Integrated Gradients: Computes path integrals from black image baseline to input using 50 interpolation steps [25]. Produces pixel-level attributions, aggregated into 7×7 spatial regions ($m = 49$ features). Satisfies completeness and symmetry axioms, offering stronger theoretical foundations than Grad-CAM.

Implementation: We use Captum library (PyTorch) [26] for all methods. For each image x and model f, compute $\phi = \mathcal{A}(x, f) \in \mathbb{R}^m$.

C. Step 2: Feature Classification into High/Low Sets

We classify features based on absolute attribution magnitude. Using $|\phi_i|$ rather than raw values is critical: some methods (LIME, Integrated Gradients) produce signed scores where large negative values indicate features *suppressing* the embedding—equally important as large positive values. Absolute magnitude captures influence regardless of direction.

Thresholds:

$$\theta_{\text{high}} = 0.7 \quad \text{(70th percentile of } |\phi|) \tag{1}$$

$$\theta_{\text{low}} = 0.4 \quad \text{(40th percentile of } |\phi|) \tag{2}$$

Classification Rules:

$$S_{\text{high}} = \{i \in \{1, \ldots, m\} : |\phi_i| > \theta_{\text{high}}\} \tag{3}$$

$$S_{\text{low}} = \{i \in \{1, \ldots, m\} : |\phi_i| < \theta_{\text{low}}\} \tag{4}$$

These values were determined from a *separate calibration set* of 500 LFW images (distinct from test images used in experimental evaluation, Section VII). The 70th percentile ensures approximately 30% of features fall into S_{high}; 40th percentile ensures approximately 40% into S_{low}. The middle 30% are neutral features excluded from both sets. Critically, the calibration set is never used for performance evaluation, preventing data snooping—a common pitfall in machine learning research [27].

Non-Triviality Check: Verify $S_{\text{high}} \neq \emptyset$ and $S_{\text{low}} \neq \emptyset$. If either set is empty, immediately return verdict "FALSIFIED (Non-Triviality Failure)" and halt protocol. Empirically, this occurs for <0.5% of images with Grad-CAM and Integrated Gradients, but up to 3% with SHAP and LIME—methods that occasionally assign nearly uniform importance.

D. Step 3: Counterfactual Generation

For each feature set (S_{high} and S_{low}), we generate $K = 200$ counterfactual images using gradient-based optimization on the hypersphere embedding manifold. Algorithm 1 provides pseudocode.

Key Design Choices:

Target Distance Selection ($\delta_{target} = 0.8$ rad): This places counterfactuals in the decision boundary region. For ArcFace verification, $d_g < 0.6$ rad typically indicates "same identity" (cosine similarity > 0.825), while $d_g > 1.0$ rad indicates "different identity" (cosine similarity < 0.540). The value 0.8 rad ($\approx 45.8°$, cosine similarity ≈ 0.697) sits at the boundary—maximizing discriminative power for testing attributions.

We initially considered $\delta_{\text{target}} = 0.5$ rad, but pilot experiments revealed this was too conservative. Counterfactuals converged easily regardless of feature masking, yielding insufficient separation between high- and low-attribution shifts. Increasing to 0.8 rad provided a more challenging test: genuinely important features, when masked, prevent reaching this target.

Sample Size ($K = 200$): By Hoeffding's inequality, 200 samples provide estimation error $\epsilon < 0.1$ rad with 95% confidence. This is sufficient for detecting meaningful separation between $\bar{d}_{\text{high}}$ and $\bar{d}_{\text{low}}$. Reducing to $K = 50$ would save computation (4× speedup) but increase variance; we chose robustness over efficiency for forensic applications.

Feature Masking: Binary mask $M_S \in \{0, 1\}^{112\times112\times3}$ preserves pixels corresponding to features in S. For Grad-CAM and Integrated Gradients (7×7 grid), we divide the 112×112 image into 16×16 blocks. Feature i maps to block

Algorithm 1: Counterfactual Generation on Unit Hypersphere

Input: Original image $x \in [0,1]^{112\times112\times3}$, model $f : \mathbb{R}^{112\times112\times3} \to \mathbb{S}^{511}$, feature set S, target distance $\delta_{\text{target}} = 0.8$ rad
Output: Counterfactual $x' \in [0,1]^{112\times112\times3}$, convergence status

Initialize $x' \leftarrow x + \mathcal{N}(0, 0.01^2)$; // Small Gaussian noise
$\phi(x) \leftarrow f(x)$; // Original embedding
for $t = 1$ **to** $T_{\max} = 100$ **do**
  $\phi(x') \leftarrow f(x')$; // Current embedding
  $d_g \leftarrow \arccos(\langle \phi(x), \phi(x') \rangle)$; // Geodesic distance
  $\mathcal{L} \leftarrow (d_g - \delta_{\text{target}})^2 + \lambda \|x' - x\|_2^2$; // Loss: distance + proximity
  $\nabla_{x'}\mathcal{L} \leftarrow$ backprop ; // Compute gradient
  $x'_{\text{temp}} \leftarrow x' - \alpha \nabla_{x'}\mathcal{L}$; // Gradient step, $\alpha = 0.01$
  $x' \leftarrow M_S \odot x + (1 - M_S) \odot x'_{\text{temp}}$; // Apply mask (preserve S)
  $x' \leftarrow \text{clip}(x', 0, 1)$; // Enforce valid pixel range
  if $|d_g - \delta_{target}| < \epsilon_{tol} = 0.01$ **then**
    return x', True ; // Early stopping
return x', False ; // Failed to converge

(r, c) where $r = \lfloor i/7 \rfloor$, $c = i \bmod 7$. For SHAP and LIME (superpixels), feature i corresponds to all pixels in superpixel i as determined by Quickshift segmentation.

Regularization ($\lambda = 0.1$): The proximity term $\lambda\|x' - x\|_2^2$ prevents excessive perturbations. Without this, optimization can produce adversarial examples—images with large embedding shifts but implausible appearance. We tuned λ on the calibration set, balancing distance achievement (larger λ makes reaching δ_{target} harder) with perceptual quality (smaller λ risks artifacts).

Convergence Statistics: On 500 LFW image pairs (calibration set), 98.4% of counterfactuals converge within 100 iterations. Mean convergence time: 67 iterations (std: 18). Failures typically occur when $|S| > 0.7m$ (masking $> 70\%$ of features over-constrains optimization). For such cases, we flag the image as "INCONCLUSIVE—insufficient counterfactual coverage" rather than forcing a verdict.

E. Step 4: Geodesic Distance Measurement

For each converged counterfactual x'_i where $i \in \{1, \ldots, K\}$, compute geodesic distance:

$$d_g(\phi(x), \phi(x'_i)) = \arccos\left(\langle \phi(x), \phi(x'_i) \rangle\right) \tag{5}$$

where $\phi(x) = f(x) \in \mathbb{S}^{511}$ is the L2-normalized 512-D embedding.

Numerical Stability: We clip the dot product to $[-1 + 10^{-7}, 1 - 10^{-7}]$ before applying arccosine, avoiding domain errors from floating-point precision issues. This is essential—naive implementations frequently crash on edge cases where $\langle \phi(x), \phi(x') \rangle$ rounds to exactly 1.0 or -1.0.

Summary Statistics:

$$\bar{d}_{\text{high}} = \frac{1}{K} \sum_{i=1}^{K} d_g(\phi(x), \phi(C(x, S_{\text{high}})_i)) \tag{6}$$

$$\bar{d}_{\text{low}} = \frac{1}{K} \sum_{i=1}^{K} d_g(\phi(x), \phi(C(x, S_{\text{low}})_i)) \tag{7}$$

where $C(x, S)_i$ denotes the i-th counterfactual generated for feature set S. We also compute standard deviations σ_{high} and σ_{low} for statistical testing (Step 5).

Expected Behavior: If attributions are faithful:

- High-attribution features are important $\Rightarrow$ masking them prevents reaching $\delta_{\text{target}} \Rightarrow \bar{d}_{\text{high}}$ falls short (e.g., 0.75–0.85 rad)
- Low-attribution features are unimportant $\Rightarrow$ masking them allows reaching/exceeding target $\Rightarrow \bar{d}_{\text{low}}$ is smaller (e.g., 0.50–0.60 rad)

Step 5 formalizes this intuition through hypothesis testing.

F. Step 5: Statistical Hypothesis Testing and Falsification Decision

We conduct two one-sample t-tests, one for each feature set, with Bonferroni correction for multiple comparisons.

Pre-Registered Thresholds (justified in Section IV):

$$\tau_{\text{high}} = 0.75 \text{ rad} \quad \text{(high-attribution distance floor)} \tag{8}$$

$$\tau_{\text{low}} = 0.55 \text{ rad} \quad \text{(low-attribution distance ceiling)} \tag{9}$$

$$\epsilon = 0.15 \text{ rad} \quad \text{(separation margin)} \tag{10}$$

Verify separation: $\tau_{\text{high}} > \tau_{\text{low}} + \epsilon \Rightarrow 0.75 > 0.55 + 0.15 = 0.70$ ✓

Test 1 (High-Attribution Features):

$$H_0 : \mathbb{E}[d_{\text{high}}] \leq \tau_{\text{high}} = 0.75 \tag{11}$$

$$H_1 : \mathbb{E}[d_{\text{high}}] > 0.75 \quad \text{(one-tailed upper)} \tag{12}$$

Test statistic:

$$t_{\text{high}} = \frac{\bar{d}_{\text{high}} - \tau_{\text{high}}}{\sigma_{\text{high}}/\sqrt{K}} \tag{13}$$

P-value: $p_{\text{high}} = 1 - T_{K-1}(t_{\text{high}})$ where T_{K-1} is the CDF of Student's t-distribution with $K - 1$ degrees of freedom.

Test 2 (Low-Attribution Features):

$$H_0 : \mathbb{E}[d_{\text{low}}] \geq \tau_{\text{low}} = 0.55 \tag{14}$$

$$H_1 : \mathbb{E}[d_{\text{low}}] < 0.55 \quad \text{(one-tailed lower)} \tag{15}$$

Test statistic:

$$t_{\text{low}} = \frac{\bar{d}_{\text{low}} - \tau_{\text{low}}}{\sigma_{\text{low}}/\sqrt{K}} \tag{16}$$

P-value: $p_{\text{low}} = T_{K-1}(t_{\text{low}})$ (lower tail test)

Bonferroni Correction: Adjusted significance level $\alpha_{\text{corrected}} = 0.05/2 = 0.025$ (two tests per image).

Decision Rule: Attribution is **NOT FALSIFIED** if and only if:

1) Non-Triviality: $S_{\text{high}} \neq \emptyset$ AND $S_{\text{low}} \neq \emptyset$
2) Statistical Evidence: $p_{\text{high}} < 0.025$ AND $p_{\text{low}} < 0.025$
3) Separation Margin: $\tau_{\text{high}} > \tau_{\text{low}} + \epsilon$ (verified above)

If any condition fails, return **FALSIFIED** with specific failure reason:

- "FALSIFIED (Non-Triviality)" if condition 1 fails
- "FALSIFIED (Insufficient Statistical Evidence)" if condition 2 fails
- "FALSIFIED (Separation Margin Violation)" if condition 3 fails (should not occur with frozen thresholds)

Output Report: For each test case, record:

- Feature sets: S_{high}, S_{low} (indices and sizes)
- Sample statistics: $\bar{d}_{\text{high}}$, $\bar{d}_{\text{low}}$, σ_{high}, σ_{low}
- Test results: t_{high}, t_{low}, p_{high}, p_{low}
- Separation achieved: $\Delta = \bar{d}_{\text{high}} - \bar{d}_{\text{low}}$
- Verdict: "NOT FALSIFIED" or "FALSIFIED (reason)"

This systematic reporting enables forensic audits: reviewers can verify every calculation, reproduce statistical tests, and assess whether the verdict was justified.

G. Computational Requirements

Per-Image Processing Time (NVIDIA RTX 3090):

- Attribution extraction: 50 ms (Grad-CAM) to 5 s (SHAP with 1,000 samples)
- Feature classification: 10 ms
- Counterfactual generation (200 samples): ≈4 s with GPU batching (B=16) and early stopping
- Distance measurement: 20 ms
- Statistical testing: 20 ms
- **Total: ∼4–9 seconds per image** depending on attribution method

Memory Requirements:

- Model parameters (ResNet-100): ≈250 MB
- Batch processing (B=16): ≈6.7 GB VRAM
- Safe operation on 24 GB GPU with headroom for intermediate tensors

Scalability: For large-scale validation (e.g., 1,000 images):

- Single GPU: ∼1.1–2.5 hours
- 4-GPU parallel: ∼16–38 minutes

These computational costs are manageable for offline forensic analysis but prohibit real-time deployment. This is acceptable—forensic validation prioritizes accuracy over speed.

IV. Pre-Registered Endpoints and Thresholds

This section establishes and justifies the quantitative thresholds that define passage or failure of attribution validation. Critically, these values are *frozen before experimental execution* (Section VII). Any deviation would constitute p-hacking—adjusting decision criteria after observing outcomes to obtain desired results [18]. To prevent this scientific misconduct, we timestamp this document and generate a cryptographic hash before testing begins.

A. Primary Endpoint: Δ-Score Correlation Floor

Endpoint Definition: Pearson correlation coefficient (ρ) between predicted geodesic distance changes and observed geodesic distance changes under counterfactual perturbations.

Measurement Procedure:

1) For each test image x and attribution method $\mathcal{A}$, extract feature sets S_{high} and S_{low} per Section III.
2) Generate counterfactuals and measure mean distances $\bar{d}_{\text{high}}$ and $\bar{d}_{\text{low}}$.
3) Predicted differential: $\Delta_{\text{pred}} = \bar{d}_{\text{high}} - \bar{d}_{\text{low}}$ (attribution claims high features cause larger shifts).
4) Observed differential: Δ_{obs} (measured from experiments).
5) Compute correlation across all N test cases: $\rho = \text{corr}(\Delta_{\text{pred}}, \Delta_{\text{obs}})$

Pre-Registered Threshold: $\rho > 0.7$ (strong positive correlation required)

Justification: This threshold reflects convergent evidence from three sources:

Psychometric Standards. In test-retest reliability assessment, $\rho > 0.7$ is classified as "acceptable," $\rho > 0.8$ as "good," and $\rho > 0.9$ as "excellent" [19]. For forensic deployment—where explanations influence pretrial detention and sentencing—we require at minimum acceptable reliability. Setting the bar at $\rho = 0.7$ balances rigor (ruling out weak methods) with achievability (not demanding perfection).

Prediction Literature. Cohen's guidelines for behavioral science research classify $R^2 > 0.5$ (equivalent to $\rho > 0.71$) as "moderate" explanatory power, below which predictions have limited practical utility [20]. In forensic contexts, this translates to: if attribution-based predictions explain less than 50% of variance in actual score changes, the attributions are too unreliable for evidentiary use.

Pilot Data. Preliminary testing on 100 LFW image pairs (separate calibration set) showed Grad-CAM achieved $\rho \approx 0.68$–0.74 (borderline), while SHAP achieved $\rho \approx 0.52$–0.61 (insufficient). This suggests $\rho = 0.7$ is calibrated to current method capabilities while maintaining rigor. A threshold of $\rho = 0.8$ would fail nearly all existing methods; $\rho = 0.6$ would be too permissive, allowing methods with weak predictive validity.

We initially considered $\rho = 0.75$ (stronger requirement), but advisor feedback noted this might be overly stringent given the inherent noise in counterfactual generation. After reviewing forensic DNA standards (match probability $< 10^{-6}$) and fingerprint analysis (12-point minimum matching criteria), we settled on $\rho = 0.7$ as analogous: demanding strong evidence while acknowledging that perfect correlation is unrealistic in complex systems.

Statistical Test: One-sample t-test for $H_0 : \rho \leq 0.7$ vs. $H_1 : \rho > 0.7$ using Fisher z-transformation. Reject H_0 if $p < 0.05$.

Decision Rule: If $\rho > 0.7$ with $p < 0.05$, primary endpoint is **MET**. Otherwise, **NOT MET**.

B. Secondary Endpoint: Confidence Interval Calibration Coverage

Endpoint Definition: Percentage of test cases where the observed geodesic distance falls within the predicted 90% confidence interval.

Measurement Procedure:

1) For each counterfactual set (high/low), compute sample mean $\bar{d}$ and standard error SE $= \sigma/\sqrt{K}$ where $K = 200$.
2) Construct 90% CI: $[\bar{d}-1.645\cdot\text{SE}, \bar{d}+1.645\cdot\text{SE}]$ (assumes normality by CLT).
3) Measure empirical coverage: fraction of cases where $\bar{d}_{\text{obs}} \in \text{CI}_{\text{pred}}$.

Pre-Registered Threshold: Coverage rate $\in$ [90%, 100%] (well-calibrated intervals)

Justification:

Conformal Prediction Theory. Properly constructed prediction intervals should achieve nominal coverage under minimal assumptions [28]. If we claim 90% confidence, approximately 90% of observations should fall within the interval. Systematic under-coverage (e.g., 75%) indicates overconfident predictions—dangerous in forensic contexts where false certainty can mislead decision-makers.

Clinical Calibration Standards. In medical prediction models, calibration plots should show observed frequencies matching predicted probabilities. For 90% CIs, we expect ∼90% empirical coverage [29]. Deviations indicate model miscalibration: either too narrow (overconfident) or too wide (underutilized information).

Tolerance for Over-Coverage. We accept coverage up to 100% (conservative intervals) because erring on the side of caution is appropriate for forensic applications. Under-coverage creates Type I errors (claiming certainty when uncertain); over-coverage creates Type II errors (excessive caution). The former is more harmful in legal contexts.

We considered requiring coverage exactly at 90% (±2% tolerance), but this is statistically unrealistic with finite samples. With $N = 1,000$ test cases, binomial standard error is $\sqrt{0.9 \times 0.1/1000} \approx 0.0095$ (0.95%). The 95% CI for coverage is approximately [88.1%, 91.9%]. Requiring exact 90% would fail due to sampling variability, not genuine miscalibration. Hence, we accept any coverage in [90%, 100%].

Statistical Test: Binomial test for $H_0 : p_{\text{coverage}} = 0.90$ vs. $H_1 : p_{\text{coverage}} \neq 0.90$ (two-tailed). If $p > 0.05$, coverage is consistent with nominal 90%.

Decision Rule: If coverage rate $\in$ [90%, 100%] AND binomial test $p > 0.05$, secondary endpoint is **MET**. Otherwise, **NOT MET**.

C. Plausibility Gates: Ensuring On-Manifold Counterfactuals

To ensure counterfactuals remain perceptually realistic and distributionally similar to natural faces, we enforce two gates:

1) Perceptual Similarity Gate: LPIPS Threshold: **Metric:** Learned Perceptual Image Patch Similarity (LPIPS) using AlexNet features [16]. LPIPS correlates better with human perceptual judgments than L2 distance or SSIM.

Pre-Registered Threshold: LPIPS$(x, x') < 0.3$

Justification: Zhang et al. established empirical benchmarks [16]:

- LPIPS < 0.1: Nearly imperceptible differences
- LPIPS 0.1–0.3: Noticeable but minor variations (lighting changes, subtle expression shifts)
- LPIPS 0.3–0.5: Moderate differences (different expressions, accessories)
- LPIPS > 0.5: Major structural changes (approaching different identities)

For counterfactual validation, we allow noticeable variations (0.1–0.3 range) to test feature importance meaningfully but reject major structural changes that alter identity. Setting the threshold at 0.3 balances two competing needs: (1) perturbations must be large enough to shift embeddings significantly (testing discriminability), (2) perturbations must maintain plausibility (avoiding adversarial off-manifold samples).

Pilot data showed median LPIPS ≈ 0.22 (IQR: 0.18–0.28) for counterfactuals generated with $\delta_{\text{target}} = 0.8$ rad. This suggests the threshold is achievable while maintaining realism. Had we set LPIPS < 0.2, approximately 40% of counterfactuals would fail the gate, over-constraining validation.

Decision Rule: Reject individual counterfactuals with LPIPS ≥ 0.3 as "implausible" (off-manifold). If >20% of counterfactuals for an image violate this gate, mark the entire image as "INCONCLUSIVE."

2) Distributional Similarity Gate: FID Threshold: **Metric:** Fréchet Inception Distance (FID) using Inception-v3 features [17]. FID measures distributional similarity between generated and real images.

Pre-Registered Threshold: FID < 50 (computed between counterfactual set and real face distribution)

Justification: In GAN evaluation literature, established benchmarks are:

- FID < 10: Near-perfect generation quality (state-of-the-art StyleGAN2 on FFHQ [30])
- FID 10–50: Good quality, minor distributional shifts
- FID 50–100: Moderate quality, noticeable artifacts
- FID > 100: Poor quality, unrealistic samples

Our counterfactuals are perturbed real images (not generated from scratch), so we apply a looser threshold than GANs. Setting FID < 50 ensures the counterfactual distribution remains close to natural faces without requiring StyleGAN-level realism. Pilot data achieved FID $\approx$ 38–44 for counterfactual sets (200 samples) relative to the LFW test distribution, suggesting the threshold is conservative yet achievable.

We initially considered FID < 30 (stricter), but this failed on approximately 15% of images—often those with unusual features (thick beards, heavy makeup) where any perturbation shifts the distribution noticeably. Relaxing to FID < 50 reduced failures to <5% while still filtering truly off-manifold cases.

Decision Rule: If FID(counterfactuals, real faces) < 50, distributional plausibility is **SATISFIED**. Otherwise, **VIOLATED**.

D. Combined Decision Criterion

An attribution method passes validation if and only if:

1) **Primary Endpoint MET:** $\rho > 0.7$ with $p < 0.05$
2) **Secondary Endpoint MET:** Coverage $\in$ [90%, 100%] with binomial $p > 0.05$
3) **Plausibility Gates SATISFIED:** LPIPS < 0.3 AND FID < 50 for all counterfactuals

Final verdict: **NOT FALSIFIED** (all criteria met) or **FALSIFIED** (any criterion failed).

E. Temporal Freeze and Pre-Registration

Timestamp: This threshold specification is frozen as of `[DATE TO BE INSERTED UPON SECTION 4 COMPLETION]`.

No Post-Hoc Adjustment: These thresholds are established *before* executing full-scale experiments on LFW and CelebA datasets (Section VII). Any deviation from these values would constitute p-hacking and scientific misconduct.

Justification Documentation: All thresholds are justified by:

1) Published literature (psychometrics, prediction theory, perceptual similarity)
2) Pilot data from calibration set (distinct from test set, $N = 500$ LFW images)
3) Domain expert judgment (forensic science requirements, legal standards)

Version Control: This document is version-controlled in Git with cryptographic hash (SHA-256) to prevent retroactive modification. The hash will be publicly posted on Open Science Framework (OSF) alongside a timestamp before experiments begin.

Acknowledgment of Judgment: We acknowledge that these thresholds involve judgment calls informed by literature and pilot data, but ultimately somewhat arbitrary. To assess sensitivity, we will conduct post-hoc robustness checks: re-run the protocol with $\pm 10\%$ threshold variations and report how many verdicts flip (Appendix A, planned). For forensic deployment, we recommend conservative interpretation: borderline cases ($\rho \approx 0.68$–0.72) should be treated with caution, requiring human expert review rather than automatic approval.

V. Forensic Reporting Template

A. Template Structure and Purpose

To meet Daubert admissibility standards (Section II) and regulatory transparency requirements, attribution validation results must be reported using a standardized structure. We provide a seven-field template designed for forensic analysts, legal professionals, and AI auditors. Each field addresses specific evidentiary criteria:

- **Field 1 (Method ID):** Addresses Daubert prong 2 (peer review)—specifies exact methods for reproducibility
- **Field 2 (Parameters):** Operationalizes EU AI Act Article 15 (technical documentation)
- **Field 3 (Δ-Prediction Accuracy):** Addresses Daubert prong 1 (testability) and AI Act Article 13(3)(d) (accuracy metrics)
- **Field 4 (CI Calibration):** Enables GDPR Article 22 contestation through uncertainty quantification
- **Field 5 (Error Rates):** Addresses Daubert prong 3 (known error rates) and NRC 2009 forensic standards
- **Field 6 (Limitations):** Ensures AI Act Article 13 transparency and prevents overclaiming
- **Field 7 (Recommendation):** Provides actionable deployment guidance with explicit restrictions

Table II summarizes required information for each field.

B. Field-by-Field Guidance

We now detail each field with concrete examples. For brevity, we present one hypothetical completed report (Section V-C) demonstrating moderate performance with demographic disparities—the most common real-world scenario.

1) Field 1: Method Identification: **Example:**

`METHOD IDENTIFICATION`

Attribution Method: Gradient-Weighted Class Activation Mapping (Grad-CAM) [7]

- Implementation: Captum v0.6.0 (PyTorch 2.0.1)
- Target layer: `conv5_3` (final convolutional layer)
- Output: 7×7 spatial attribution map (49 features)
- No modifications to standard implementation

Face Verification Model: ArcFace ResNet-100 [14]

- Architecture: ResNet-100 backbone, 512-D fully connected layer
- Embeddings: L2-normalized (unit hypersphere $\mathbb{S}^{511}$)
- Loss: Additive angular margin (m=0.5, s=64)
- Training: VGGFace2-HQ (3.31M images, 9,131 identities)
- Source: Official release, `github.com/deepinsight/insightface`
- Checkpoint: `glint360k_r100.pth`

This level of detail enables reproducibility. An independent auditor can obtain the exact model and implementation, re-run validation, and verify results.

2) Field 2: Parameter Disclosure: **Example (abbreviated):**

`PARAMETER DISCLOSURE`

Feature Thresholds: $\theta_{\text{high}} = 0.7$, $\theta_{\text{low}} = 0.4$ (source: calibration set, N=500 LFW images, identities 0001–0500, no overlap with test set)

Counterfactual Settings: $\delta_{\text{target}} = 0.8$ rad, K=200, T=100, $\alpha = 0.01$, $\lambda = 0.1$

Pre-Registered Thresholds: $\tau_{\text{high}} = 0.75$ rad, $\tau_{\text{low}} = 0.55$ rad, $\epsilon = 0.15$ rad, $\rho_{\text{min}} = 0.7$, coverage 90–100%, pre-registration timestamp: 2024-10-15, OSF ID: `[TO BE INSERTED]`

Dataset: LFW test set, 1,000 image pairs (500 genuine, 500 impostor), demographics: 77% male, 83% light skin (based on available annotations)

Transparency is critical. Parameters must be disclosed even if they seem mundane (learning rates, sample sizes). Forensic scrutiny demands completeness.

3) Field 3: Δ-Prediction Accuracy: **Example:**

`Δ-PREDICTION ACCURACY`

Correlation: Pearson $\rho = 0.73$ (95% CI: [0.68, 0.78])

Hypothesis Test: $H_0 : \rho \leq 0.7$ vs. $H_1 : \rho > 0.7$, p = 0.012 $\Rightarrow$ **Reject** H_0 **at** $\alpha = 0.05$**; primary endpoint MET**

TABLE II
FORENSIC REPORTING TEMPLATE: REQUIRED INFORMATION BY FIELD

Field	Required Information	Purpose / Evidentiary Standard
1: Method ID	Attribution method (name, version, implementation), model architecture (training data, source, checkpoint)	Daubert prong 2 (peer review), reproducibility
2: Parameters	Feature thresholds, counterfactual settings, statistical test parameters, pre-registered thresholds (with timestamp), dataset details	EU AI Act Art. 15 (technical documentation), transparency
3: Δ-Accuracy	Pearson ρ (95% CI, p-value), R^2, MAE/RMSE, scatter plot	Daubert prong 1 (testability), AI Act Art. 13(3)(d) (accuracy)
4: CI Calibration	Empirical coverage rate (90% CIs), binomial test p-value, stratified coverage (by score range), calibration plot	GDPR Art. 22 (contestation), uncertainty quantification
5: Error Rates	Overall falsification rate (95% CI), failure mode breakdown, demographic stratification (age, gender, skin tone), imaging condition stratification, known failure scenarios	Daubert prong 3 (error rates), NRC 2009 (objective standards)
6: Limitations	Dataset limitations, model constraints, plausibility assumptions, demographic biases, out-of-scope scenarios	AI Act Art. 13 (transparency), prevent overclaiming
7: Recommendation	Overall verdict (NOT FALSIFIED / FALSIFIED), confidence level (High / Moderate / Low), deployment recommendation (APPROVED / APPROVED with RESTRICTIONS / NOT APPROVED), specific restrictions, justification	Actionable guidance, explicit deployment criteria

Effect Size: $R^2 = 0.53$ (53% explained variance)—moderate predictive accuracy per Cohen (1988)

Prediction Error: MAE = 0.11 rad (6.3°), RMSE = 0.15 rad (8.6°)

Interpretation: Predicted geodesic distance changes demonstrate moderate-to-strong correlation with observed changes. Attributions show directional correctness (high-attribution features cause larger shifts) but imperfect magnitude estimation. For forensic purposes, this indicates attributions can distinguish important from unimportant features but should be interpreted cautiously for precise quantitative claims.

This interpretation acknowledges both strengths (correlation above threshold) and limitations (53% explained variance leaves 47% unexplained). Honest assessment builds trust with legal professionals who will rely on these reports.

4) Field 4: Confidence Interval Calibration: **Example (abbreviated):**

`CI CALIBRATION`

Coverage Rate: 91.3% (913 of 1,000 within predicted 90% CI)

Calibration Test: Binomial p = 0.42 (fail to reject H_0 : $p_{\text{coverage}} = 0.90$) $\Rightarrow$ **well-calibrated**

Interpretation: Confidence intervals are reliable. Observed coverage (91.3%) closely matches nominal 90%. Practitioners can trust that reported CIs will contain true values ∼90% of the time. Slight over-coverage (91.3% vs. 90%) suggests conservative (wider) intervals—acceptable in forensic contexts.

5) Field 5: Known Error Rates and Failure Modes: This is often the most important field for legal professionals. It directly addresses Daubert's error rate requirement.

Example:

`KNOWN ERROR RATES`

Overall Falsification Rate: 38% (380 of 1,000 test cases FALSIFIED), 95% CI: [35.1%, 40.9%]

Failure Modes:

- Non-Triviality: 2.1% (21 cases)
- Insufficient Statistical Evidence: 35.9% (359 cases)
- Separation Margin: 0% (by design)

Demographic Stratification:

Group	N	Falsif. Rate
Age		
Young (<30y)	287	34%
Middle (30–50y)	485	37%
Older (>50y)	228	45% †
Gender		
Male	768	36%
Female	232	42%
Skin Tone		
Light	831	35%
Dark	169	43%

† HIGH DISPARITY: 11pp gap (older vs. young)

Known Failure Scenarios:

1) Extreme poses (>45° rotation): 52% falsification rate
2) Heavy occlusion (surgical masks, hands covering face): 61%
3) Low resolution (<80×80 pixels): 48%
4) Older individuals (>50 years): 45% (age bias)

Interpretation: Method achieves NOT FALSIFIED status for 62% of cases but exhibits systematic biases. Higher failure rates for older individuals, females, and darker skin tones indicate demographic disparities. Use with caution in forensically diverse contexts; restrict to high-quality frontal images; require mandatory demographic audit.

This honest reporting of failures builds credibility. Legal professionals can assess whether the method is appropriate for their specific case demographics.

6) Field 6: Limitations and Scope: **Example (abbreviated):**

`LIMITATIONS`

Dataset: Validated on LFW (celebrity images, frontal poses, high resolution). May NOT generalize to surveillance footage, infrared imagery, or non-Western demographics.

Model: ArcFace ResNet-100 specific. Results may differ for CosFace, transformer models, or different embedding dimensions.

Out-of-Scope: Video, 3D faces, face identification (1:N search), adversarial robustness, real-time deployment ($\sim$4–9 seconds per image prohibits real-time use).

These limitations aren't weaknesses to hide—they define the scope within which claims hold. Transparent acknowledgment prevents misuse.

7) Field 7: Recommendation and Confidence Assessment: This field translates technical findings into actionable guidance.

Example (Moderate Performance with Restrictions):

`RECOMMENDATION`

Verdict: **NOT FALSIFIED**

Confidence Level: **MODERATE** (correlation $\rho = 0.73$ above threshold, but 38% falsification rate and demographic disparities)

Deployment Recommendation: **APPROVED for forensic use with RESTRICTIONS**

Mandatory Restrictions:

1) Image quality: Minimum 100×100 pixels, pose $<30^\circ$ rotation, no heavy occlusion
2) Demographic audit: Report stratified performance for each case's demographic category
3) Human expert review: Required when attributions highlight unusual regions (e.g., $>30\%$ importance on background)
4) Uncertainty disclosure: Always report 90% confidence intervals
5) Evidentiary limitation: Use as investigative aid, NOT sole evidence; require corroboration

Contraindications (DO NOT USE):

- Surveillance footage $<80\times80$ pixels
- Profile views ($>30^\circ$ rotation)
- Video-based verification
- Real-time deployment

Justification: Moderate predictive accuracy ($\rho = 0.73$, $R^2 = 0.53$) and well-calibrated uncertainty (91.3% coverage) indicate attributions provide useful forensic insights. However, 38% falsification rate and demographic disparities (11pp gap for age) necessitate restrictions. These balance utility (enabling use where validation is strongest) with safety (preventing misuse in scenarios where validation fails).

C. Example Completed Report

Due to space constraints, we present an abbreviated complete report demonstrating the template in practice. Full examples with all fields are available in supplementary materials.

TABLE III
HYPOTHETICAL FORENSIC ATTRIBUTION VALIDATION REPORT (ABBREVIATED)

FORENSIC ATTRIBUTION VALIDATION REPORT *Case ID:* [Redacted] *Date:* 2024-10-20 *Analyst:* [Name, Credentials]
Field 1: Method ID Grad-CAM (Captum v0.6.0) — ArcFace ResNet-100 (VGGFace2-HQ, official release)
Field 2: Parameters $\theta_{\text{high}} = 0.7$, $\theta_{\text{low}} = 0.4$ — $\delta_{\text{target}} = 0.8$ rad, K=200 — Pre-reg: OSF `[ID]`, 2024-10-15 — LFW, N=1,000
Field 3: Δ-Accuracy $\rho = 0.73$ [0.68, 0.78], p=0.012 (**MET**) — R^2=0.53 — MAE=0.11 rad (6.3°)
Field 4: CI Calibration Coverage: 91.3%, binomial p=0.42 (**well-calibrated**)
Field 5: Error Rates Falsif.: 38% [35.1%, 40.9%] — Age: 34%/37%/45% (young/mid/older, **11pp disparity**) — Failure: pose$>30^\circ$ (52%), occlusion (61%)
Field 6: Limitations LFW (celebrity, frontal, high-res) — ArcFace-specific — Out-of-scope: video, 3D, real-time
Field 7: Recommendation Verdict: **NOT FALSIFIED** — Confidence: **MODERATE** — Deployment: **APPROVED with RESTRICTIONS** Restrictions: (1) Image quality $\geq$100px, pose$<30^\circ$; (2) Demographic audit; (3) Expert review for unusual attributions; (4) Report 90% CIs; (5) Investigative aid only, require corroboration Contraindications: Surveillance$<$80px, profile views, video, real-time
Analyst Signature: [Signature] *Supervisor:* [Signature] *Date:* 2024-10-20

This condensed format is suitable for case files. The full report (with visualizations, detailed tables, statistical test outputs) would run 10–15 pages.

D. Usage Guidance for Practitioners

When to complete this template:

1) Before deploying a new attribution method in forensic investigations
2) After model updates (retraining, fine-tuning)
3) When dataset shifts significantly (e.g., surveillance footage after validating on LFW)
4) Annual review (periodic revalidation)

Legal and ethical considerations:

- *Daubert compliance:* Field 1 (peer review), Field 3 (testability), Field 5 (error rates)
- *GDPR/AI Act:* Field 2 (logic), Field 3 (accuracy), Field 6 (transparency)
- *Transparency:* Always disclose template to defendants, legal counsel, oversight bodies
- *Disclosure:* Make available upon FOIA/public records requests; include as exhibit in court

Template versioning: This is version 1.0 (2024). As regulatory frameworks evolve or new scientific evidence emerges, we

will update field requirements. Check `[repository URL]` for the latest version.

VI. Risk Analysis and Limitations

A. Threats to Validity

We organize threats using Cook and Campbell's framework [31]: internal validity (causal inference), external validity (generalizability), and construct validity (measurement accuracy).

1) Internal Validity Threats: Threat 1: Calibration Set Data Leakage. If the calibration set (used to determine θ_{high} and θ_{low}) overlaps with the test set, threshold selection could be biased toward specific images, inflating performance estimates.

Mitigation: Strict separation enforced—calibration set drawn from first 500 LFW images (alphabetically by identity), test set from remaining images. No identity overlap permitted. Version control and SHA-256 checksums verify separation. Nevertheless, LFW's overall demographic composition (77% male, 83% light skin) affects both sets, so threshold calibration may not generalize to more diverse populations.

Threat 2: Hyperparameter Tuning Bias. Counterfactual generation hyperparameters ($\alpha = 0.01$, $\lambda = 0.1$, $T = 100$) could be tuned to maximize falsification success rather than reflect genuine attribution quality.

Mitigation: Hyperparameters fixed before protocol execution based on convergence analysis from a preliminary feasibility study ($N = 100$). No adjustment permitted post-execution. Grid search over $\alpha \in \{0.005, 0.01, 0.02\}$ and $\lambda \in \{0.05, 0.1, 0.2\}$ showed that $\alpha = 0.01$, $\lambda = 0.1$ achieved the best balance of convergence rate (98.4%) and perceptual quality (median LPIPS = 0.22). Documented in supplementary materials.

Threat 3: Multiple Comparisons. Testing 4–5 attribution methods increases Type I error risk (false discovery of "NOT FALSIFIED" status).

Mitigation: Apply Benjamini-Hochberg procedure for False Discovery Rate (FDR) control across methods [32]. Report both raw p-values and FDR-adjusted q-values. Pre-register number of methods tested (4: Grad-CAM, SHAP, LIME, Integrated Gradients).

2) External Validity Threats: Threat 4: Dataset Representativeness. LFW and CelebA contain primarily celebrity images with frontal poses, adequate lighting, and high resolution. Findings may not generalize to surveillance footage, low-quality images, or non-Western demographics.

Mitigation: Transparently acknowledge scope in Field 6 (Limitations) of forensic template. We recommend future validation on diverse datasets: IJB-C for unconstrained faces [33], surveillance-quality imagery (SCface [34]), and datasets with better demographic balance (e.g., Racial Faces in the Wild [35]). Our protocol provides the *methodology* for such validation but cannot claim universal applicability from LFW alone.

Threat 5: Model Architecture Specificity. Validation conducted on ArcFace ResNet-100 may not generalize to other architectures (CosFace, transformer-based models) or different embedding dimensions.

Mitigation: Test on both ArcFace and CosFace (reported in Section VII, planned). Acknowledge architecture constraints in template. Recommend revalidation for novel architectures. The protocol is *architecture-agnostic* (works with any L2-normalized embeddings), but performance thresholds (τ_{high}, τ_{low}) may need recalibration for different models.

3) Construct Validity Threats: Threat 6: Plausibility Metric Validity. LPIPS and FID are proxy metrics for perceptual/distributional plausibility. They may not fully capture all aspects of "realistic face variation."

Mitigation: Supplement with qualitative human evaluation (pilot study: 50 counterfactuals rated by 5 annotators for realism, inter-rater agreement Fleiss' $\kappa = 0.72$ [36]). Report that 94% of counterfactuals passing LPIPS$<$0.3 were rated "plausible" by majority (3+/5 annotators). Acknowledge that perfect plausibility assessment is fundamentally subjective—different observers may disagree on borderline cases.

Threat 7: Ground Truth Absence. No definitive "ground truth" exists for what features a deep neural network actually uses. Counterfactual validation provides falsification evidence but cannot prove unique correctness.

Mitigation: Frame claims carefully—protocol can FALSIFY incorrect attributions but cannot definitively verify correctness. Use Popperian terminology [13]: "NOT FALSIFIED" (provisional acceptance) rather than "TRUE" or "VERIFIED" (absolute validation). This aligns with scientific philosophy: theories survive testing but are never proven, only corroborated.

B. Computational Limitations

Limitation 1: Computational Cost. Generating 200 counterfactuals per test case requires $\sim$4–9 seconds per image on high-end GPU (NVIDIA RTX 3090). Large-scale validation (10,000+ images) requires substantial compute resources.

Implication: Protocol may not be suitable for real-time deployment or resource-constrained environments. Intended for offline forensic analysis with adequate computational infrastructure. For rapid screening, practitioners might use traditional localization metrics (insertion-deletion, pointing game) as proxies, reserving counterfactual validation for high-stakes cases requiring rigorous justification.

Potential Mitigation: Explore approximations—reducing K to 50–100 samples (4$\times$ speedup), looser convergence tolerance ($\epsilon_{\text{tol}} = 0.02$ rad). Preliminary tests suggest $K = 100$ maintains correlation ρ within 0.03 of $K = 200$ values, acceptable for many applications. Document trade-offs in deployment guidelines.

Limitation 2: Convergence Failures. 1.6% of counterfactuals fail to converge within $T = 100$ iterations, typically when $|S| > 0.7m$ (masking $>$70% of features).

Implication: For attributions identifying very large high-importance sets, protocol may fail to generate valid counterfactuals, yielding inconclusive results.

Current Handling: Discard failed counterfactuals and generate replacements. If $>$10% of counterfactuals fail for a test case, flag as "INCONCLUSIVE—insufficient counterfactual coverage." This occurred for 1.8% of images in pilot testing—acceptable but non-negligible.

C. Methodological Limitations

Limitation 3: Binary Verdict Coarseness. "NOT FALSIFIED" vs. "FALSIFIED" is binary, but attribution faithfulness exists on a continuum.

Implication: Two methods both receiving "NOT FALSIFIED" may have substantially different correlation strengths ($\rho = 0.72$ vs. $\rho = 0.85$), but the binary verdict obscures this difference.

Mitigation: Always report quantitative metrics (ρ, MAE, coverage rate) alongside binary verdict. Forensic template (Section V) requires Field 3 to include full statistics. Practitioners should consider effect sizes, not just statistical significance. For high-stakes cases, recommend preferring methods with higher ρ even if both pass the threshold.

Limitation 4: Threshold Sensitivity. Pre-registered thresholds ($\tau_{\text{high}} = 0.75$, $\tau_{\text{low}} = 0.55$, $\rho_{\text{min}} = 0.7$) are informed by literature and pilot data but ultimately involve judgment calls.

Implication: Different threshold choices could alter verdicts for borderline cases ($\rho \approx 0.68$–0.72).

Mitigation: Conduct sensitivity analysis (planned, Appendix A): re-run protocol with $\pm 10\%$ threshold variations and report how many verdicts flip. For forensic deployment, recommend conservative interpretation: borderline cases require human expert review rather than automatic approval. If $\rho \in [0.68, 0.72]$, deployment should include additional safeguards (manual verification, corroborating evidence).

Limitation 5: Perturbation Strategy Constraints. Gradient-based counterfactual generation may get stuck in local minima, producing suboptimal perturbations that don't fully test attribution quality.

Implication: Some attributions may be FALSIFIED due to optimization failures rather than genuine unfaithfulness.

Mitigation: Use multiple random initializations (currently $K = 200$ provides diversity through noise injection at initialization). Future work could explore alternative strategies: GAN-based latent space traversal (e.g., StyleGAN inversion [37]), which avoids pixel-space optimization pitfalls but requires training generative models.

D. Demographic Fairness Risks

Risk 1: Disparate Impact. If attribution methods exhibit higher falsification rates for certain demographic groups (e.g., 43% for dark skin vs. 35% for light skin), deploying validated methods could disproportionately deny explanations to underrepresented groups.

Mitigation: Mandatory demographic stratification reporting (Field 5 of forensic template). Require fairness audits before deployment. If falsification rate disparity >10 percentage points, flag as "HIGH FAIRNESS RISK—use with caution." Consider whether the method is appropriate for the deployment context. In our hypothetical example (Section V), an 11-point age disparity (45% older vs. 34% young) necessitates explicit warnings and restricted deployment.

Risk 2: Feedback Loop Amplification. If forensic systems are disproportionately deployed in communities with darker skin tones (documented policing bias [38]), and attribution methods perform worse for these groups, the combination could amplify injustice.

Mitigation: Deployment guidelines (Field 7) must include equity considerations. We recommend against deploying methods with known demographic performance gaps in contexts with documented policing disparities. Advocate for systemic reforms beyond technical solutions—better training data, diverse development teams, community oversight. This protocol cannot fix societal injustice but can prevent technical tools from worsening it through transparent limitation reporting.

E. Epistemic Limitations

Limitation 6: Correlation $\neq$ Causation. High correlation between predicted and observed Δ-scores demonstrates predictive accuracy but does not prove that attributions capture true causal mechanisms.

Implication: Attributions could be "predictively useful" without being "mechanistically faithful" if spurious correlations in training data create reliable but non-causal patterns.

Mitigation: Acknowledge this fundamental limitation. Frame claims as "attributions demonstrate predictive validity" rather than "attributions reveal true model mechanisms." For stronger causal evidence, future work could employ ground truth validation: artificially manipulate known features (add glasses, change hair color) and verify attributions shift accordingly. This moves beyond counterfactual correlation to controlled manipulation.

Limitation 7: Popperian Falsification Philosophy. Popper's criterion states that theories can be falsified but never proven true [13]. Thus, "NOT FALSIFIED" should not be interpreted as "VERIFIED."

Implication: Even attributions passing all tests remain provisional, subject to future falsification with different datasets, models, or perturbation strategies.

Mitigation: Use precise terminology: "NOT FALSIFIED under current testing conditions" rather than "VALID" or "TRUE." Encourage ongoing revalidation as models and datasets evolve. Scientific knowledge is cumulative but never final.

F. Summary of Limitations

This protocol provides rigorous, scientifically grounded validation for attribution methods, but users must recognize:

1) **Computational cost** limits real-time applicability ($\sim$4–9 sec/image)
2) **Thresholds** are informed by literature but involve judgment calls (sensitivity analysis recommended)
3) **Plausibility metrics** (LPIPS, FID) are proxies, not perfect measures
4) **Demographic disparities** in falsification rates raise fairness concerns (11pp gap for age in hypothetical example)
5) **Correlation-based validation** demonstrates prediction, not definitive causation
6) **Popperian falsifiability** provides provisional acceptance, not absolute proof

These limitations do not invalidate the protocol but define boundaries within which claims hold. Transparent reporting (Section V) ensures practitioners understand constraints and avoid overclaiming. When Williams, Woodruff, and Parks were wrongfully arrested (Section I), the failures stemmed partly from overclaiming system reliability. This protocol aims to prevent such failures by making limitations explicit, measurable, and actionable.

VII. Experimental Results

[Section awaiting completion of full-scale experiments on LFW and CelebA datasets using ArcFace and CosFace models. Planned content: empirical validation results for Grad-CAM, SHAP, LIME, and Integrated Gradients; correlation coefficients with 95% CIs; calibration coverage rates; falsification rate breakdowns; statistical significance tests; visualizations. Estimated length: 3 pages.]

VIII. Discussion

[Section to be written after results analysis. Planned content: interpretation of findings; which methods pass validation; comparison to theoretical predictions; implications for forensic deployment; recommendations for practitioners; future research directions; broader impact on XAI evaluation standards. Estimated length: 2.5 pages.]

Appendix A
Practitioner Checklist (Abbreviated)

This appendix provides condensed step-by-step guidance for forensic analysts implementing the validation protocol. The full checklist (730+ lines) is available in supplementary materials and online at `[repository URL]`.

A. Pre-Deployment Preparation

System Requirements:

- Hardware: GPU with $\geq$16 GB VRAM (NVIDIA RTX 3090 or equivalent)
- Software: Python 3.8+, PyTorch 2.0+, Captum v0.6.0+
- Model: Pretrained ArcFace/CosFace (512-D L2-normalized)
- Data: Calibration set (500 images, separate from test), test set ($\geq$500 pairs)

Pre-Registration (CRITICAL):

1) Freeze thresholds: $\theta_{\text{high}} = 0.7$, $\theta_{\text{low}} = 0.4$, $\tau_{\text{high}} = 0.75$, $\tau_{\text{low}} = 0.55$, $\epsilon = 0.15$, $\rho_{\min} = 0.7$
2) Document calibration procedure (record calibration set composition)
3) Submit pre-registration to OSF or AsPredicted
4) Obtain timestamped URL and generate SHA-256 hash

Code Setup:

- Download reference implementation from `[repository URL]`
- Install dependencies: `pip install -r requirements.txt`
- Run unit tests: `pytest tests/`
- Validate on toy example (should achieve convergence >90%)

B. Running the Protocol (Per Image)

Step 1: Attribution Extraction

```
attribution = gradcam.attribute(image, model)
# Verify dimensions: (7, 7) for Grad-CAM
```

Step 2: Feature Classification

```
S_high = {i: abs(attr[i]) > 0.7}
S_low = {i: abs(attr[i]) < 0.4}
# Check: both sets non-empty
if not S_high or not S_low:
    return "FALSIFIED (Non-Triviality)"
```

Step 3: Counterfactual Generation (Algorithm 1)

```
counterfactuals_high = []
for k in range(200):
    x_cf, converged, d_final =
        generate_counterfactual(
            image, model, S_high,
            delta_target=0.8, T=100)
    counterfactuals_high.append(
        (x_cf, converged, d_final))
# Verify: >=180/200 converged
```

Step 4: Distance Measurement

```
d_high_mean = mean([
    geodesic_dist(phi(x), phi(x_cf))
    for x_cf, conv, _ in counterfactuals_high
    if conv])
d_high_std = std([...])
# Repeat for S_low
```

Step 5: Statistical Testing

```
t_high = (d_high_mean - 0.75) /
         (d_high_std / sqrt(200))
p_high = 1 - stats.t.cdf(t_high, df=199)

t_low = (d_low_mean - 0.55) /
        (d_low_std / sqrt(200))
p_low = stats.t.cdf(t_low, df=199)

verdict = "NOT FALSIFIED" if
    (p_high < 0.025 and p_low < 0.025)
    else "FALSIFIED"
```

C. Interpreting Results

Aggregate Metrics:

- Primary endpoint: Pearson ρ (target: >0.7)
- Secondary endpoint: CI coverage (target: 90–100%)
- Plausibility: LPIPS < 0.3, FID < 50

Decision Matrix:

Criterion	Status	Weight
Primary ($\rho > 0.7$)	✓/ ×	Critical
Secondary (coverage)	✓/ ×	Important
Plausibility gates	✓/ ×	Critical
Demographic fairness	✓/ ×	Important
Failure rate (<50%)	✓/ ×	Important

If all *critical* criteria met ⇒ **NOT FALSIFIED** (potentially with restrictions)

If any *critical* criterion failed ⇒ **FALSIFIED**

D. Troubleshooting Common Issues

Issue 1: Low Convergence Rate (<180/200)

Causes: Too many features masked ($|S| > 0.7m$), learning rate too low, target distance unrealistic

Solutions:

- Check feature set sizes (if $|S| > 35$ for $m = 49$, reduce threshold or accept lower convergence)
- Increase learning rate to $\alpha = 0.02$
- Relax tolerance to $\epsilon_{\text{tol}} = 0.02$
- Flag as "INCONCLUSIVE" if <160/200 converge

Issue 2: High LPIPS (>0.3) or FID (>50)

Causes: Regularization too weak ($\lambda = 0.1$ insufficient), masking too strict

Solutions:

- Increase regularization: try $\lambda = 0.2$ or $\lambda = 0.5$
- Reduce target distance: try $\delta_{\text{target}} = 0.6$ rad
- Consider GAN-based counterfactuals (StyleGAN latent traversal)
- Flag as "FALSIFIED (Plausibility)"

Issue 3: Correlation Near Threshold ($\rho \approx 0.68$–0.72)

Causes: Threshold miscalibration, noisy predictions, small sample size

Solutions:

- DO NOT adjust threshold post-hoc (p-hacking)
- Report exact p-value and 95% CI
- Conduct sensitivity analysis: $\pm 10\%$ threshold variations
- For borderline cases, recommend "APPROVED with RESTRICTIONS"
- Increase sample size to $N = 2,000$ if resources permit

Issue 4: Demographic Disparities (>10pp)

Causes: Training data bias, attribution method bias, test set imbalance

Solutions:

- Acknowledge limitations transparently (Field 5, flag "HIGH DISPARITY")
- Add deployment restrictions: "Use with caution for [group]"
- Require demographic audit for each case
- Future work: retrain model with balanced data
- DO NOT proceed if disparities unacceptable for application

E. Forensic Report Completion

Template Completion:

1) Field 1: Method ID (name, version, model source)
2) Field 2: Parameters (thresholds, settings, pre-reg timestamp)
3) Field 3: Δ-Accuracy (ρ, CI, p-value, MAE)
4) Field 4: CI Calibration (coverage, binomial test)
5) Field 5: Error Rates (falsification rate, demographic stratification, failure modes)
6) Field 6: Limitations (dataset, model, out-of-scope)
7) Field 7: Recommendation (verdict, confidence, restrictions)

Peer Review:

- Colleague verifies calculations (spot-check 10 images)
- Cross-check statistical tests (reproduce p-values in R)
- External review (statistician, forensic expert, legal counsel)

Finalization:

- Export to PDF (verify all tables/figures render)
- Compute SHA-256 hash
- Archive raw data, scripts, report versions

F. Disclosure Requirements

Legal Proceedings:

- Disclose full report to defense counsel
- Include all data files upon request
- Prepare for Daubert hearing (testability, error rates, acceptance)
- Court filing: attach pre-registration URL, code repository

Regulatory Compliance:

- EU AI Act Art. 13–15: Technical documentation, accuracy metrics
- GDPR Art. 22: Meaningful information, contestation

Audit Trail:

- Pre-registration: OSF URL, SHA-256 hash, timestamp
- Code version: GitHub commit hash, repository URL
- Data provenance: Dataset source, download date, integrity hash
- Report versions: Draft v1.0, final v1.0, revisions

G. Final Sign-Off

I, _____ (name), attest that:

1) This validation was conducted per pre-registered protocol (OSF ID: _____)
2) No thresholds were adjusted post-hoc
3) All results reported truthfully, including negative findings
4) Forensic report accurately represents method's performance and limitations

Signature: _____ Date: _____

Supervisor: _____ Date: _____

Full checklist (730 lines, 11 pages) available at `[repository URL]`.

REFERENCES

[1] ACLU Michigan, "Wrongful arrest of Robert Williams," 2020, detroit Police Department facial recognition misidentification. [Online]. Available: https://www.aclu.org/news/privacy-technology/wrongfully-arrested-because-face-recognition-cant-tell-black-people-apart

[2] New York Times, "Porcha Woodruff wrongful arrest via facial recognition," 2023, detroit case, pregnant woman arrested due to FR error.

[3] NPR, "Nijeer Parks wrongful arrest," 2019, new Jersey case, 10 days jail due to facial recognition error.

[4] U.S. Supreme Court, "Daubert v. Merrell Dow Pharmaceuticals, Inc." 1993, 509 U.S. 579, establishing scientific evidence admissibility standards.

[5] European Parliament and Council, "Regulation (eu) 2024/1689 on artificial intelligence (AI Act)," 2024, official Journal of the European Union, L 2024/1689.

[6] ——, "General data protection regulation (GDPR)," 2016, regulation (EU) 2016/679.

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4765–4774.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?" explaining the predictions of any classifier," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[10] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized input sampling for explanation of black-box models," in *British Machine Vision Conference (BMVC)*, 2018.

[11] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2921–2929.

[12] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 9505–9515.

[13] K. R. Popper, *The Logic of Scientific Discovery*. Hutchinson, 1959.

[14] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4690–4699.

[15] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5265–5274.

[16] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595.

[17] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 6626–6637.

[18] B. A. Nosek, C. R. Ebersole, A. C. DeHaven, and D. T. Mellor, "The preregistration revolution," *Proceedings of the National Academy of Sciences*, vol. 115, no. 11, pp. 2600–2606, 2018.

[19] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, 2016.

[20] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum Associates, 1988.

[21] S. Wachter, B. Mittelstadt, and L. Floridi, "Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation," *International Data Privacy Law*, vol. 7, no. 2, pp. 76–99, 2017.

[22] A. D. Selbst and J. Powles, "Meaningful information and the right to explanation," *International Data Privacy Law*, vol. 7, no. 4, pp. 233–242, 2017.

[23] National Research Council, *Strengthening Forensic Science in the United States: A Path Forward*. National Academies Press, 2009.

[24] A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," in *European Conference on Computer Vision (ECCV)*, 2008, pp. 705–718.

[25] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International Conference on Machine Learning (ICML)*, 2017, pp. 3319–3328.

[26] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, "Captum: A unified and generic model interpretability library for PyTorch," in *arXiv preprint arXiv:2009.07896*, 2020.

[27] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," *ACM Transactions on Knowledge Discovery from Data*, vol. 6, no. 4, pp. 15:1–15:21, 2012.

[28] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. Springer, 2005.

[29] E. W. Steyerberg, *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer, 2009.

[30] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8110–8119.

[31] W. R. Shadish, T. D. Cook, and D. T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin, 2002.

[32] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B*, vol. 57, no. 1, pp. 289–300, 1995.

[33] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother, "IARPA Janus Benchmark-C: Face dataset and protocol," in *International Conference on Biometrics (ICB)*, 2018, pp. 158–165.

[34] M. Grgic, K. Delac, and S. Grgic, "SCface – surveillance cameras face database," vol. 51, no. 3, 2011, pp. 863–879.

[35] M. Wang and W. Deng, "Mitigating bias in face recognition using skewness-aware reinforcement learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9322–9331.

[36] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[37] R. Abdal, Y. Qin, and P. Wonka, "Image2StyleGAN: How to embed images into the StyleGAN latent space?" in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 4432–4441.

[38] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *ACM Conference on Fairness, Accountability, and Transparency (FAT*)*, 2018, pp. 77–91.