PEERING OVER THE MANIFOLD VISTAS: A UNIFIED GEOMETRIC FRAMEWORK FOR COUNTERFACTUAL XAI

Anonymous authors
Paper under double-blind review

ABSTRACT

Counterfactual recourse identifies minimal predictive interventions on complex data manifolds, yet existing methods either yield off-manifold shortcuts via Euclidean descent or incur prohibitive computational overhead via Riemannian curvature pullbacks. Grounded in the local differential structure of deep representations, we propose a unified geometric framework: we derive a metric-weighted local right-inverse of the task differential, prove mathematically that it yields the strictly metric-optimal update direction, and bound manifold distortion between intrinsic pullback and Euclidean extremes. By systematically truncating task-irrelevant singular directions, our solver generates sparse, stable updates that suppress off-target attribute drift. Across cross-domain benchmarks, our method consistently matches or outperforms prior baselines: in latent counterfactual steering, it reduces off-target distraction entropy by 30% and perceptual deviation by 29% relative to explicit Riemannian optimization while running 2.5× faster. Furthermore, on blackbox semiconductor metrology with a tabular foundation model, our formulation resolves nanoscale physical geometries 2.8× faster than standard descent under stochastic zeroth-order Jacobians.

1 INTRODUCTION

Explainable AI (XAI) is the study of *"why"* a model made a prediction to understand its inherent logic (Arrieta et al., 2020; Guidotti, 2024). A prominent class of local explanations characterizes predictions through input sensitivity or gradients, producing saliency or feature-attribution maps (Simonyan et al., 2013; Sundararajan et al., 2017). While such attribution methods identify features associated with a local prediction, they do not generally specify which input changes are sufficient to alter the outcome (Kommiya Mothilal et al., 2021; Karimi et al., 2022). In contrast, *counterfactual* XAI is a growing subfield where instead of asking *"why"*, researchers ask *"what if"* to study the inner workings of the model. In brief, a counterfactual explanation describes the closest input state that satisfies the desired target prediction (Wachter et al., 2018; Chou et al., 2022; Guidotti, 2024). A counterfactual is typically formulated as a nearby or minimum-cost input satisfying a target predictive constraint (Wachter et al., 2018; Guidotti, 2024; Verma et al., 2024). Counterfactual

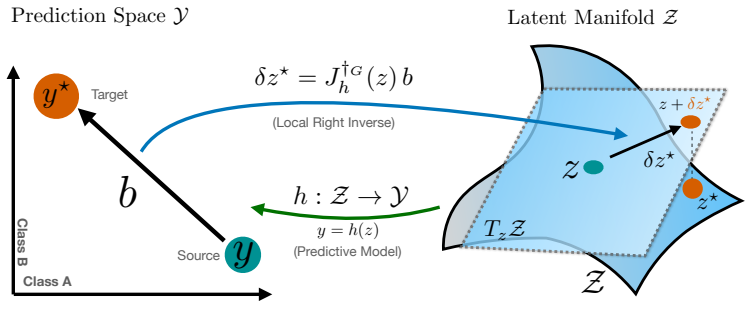

Figure 1: Cartoon overview of geometric counterfactual steering using the local right inverse.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

generation is commonly cast as a constrained optimization problem. In iterative methods, the chosen objective and geometry determine the endpoint and also the trajectory through the representation space (Wachter et al., 2018; Pegios et al., 2025).

In this work, we focus on counterfactual generation over deep generative representations (Joshi et al., 2019; Rodríguez et al., 2021; Khorram & Fuxin, 2022; Augustin et al., 2022). Because the decoder is non-linear, Euclidean latent geometry generally provides a distorted view of displacement in the generated data space (Arvanitidis et al., 2018; Shao et al., 2018). Consequently, counterfactual optimization under flat or misaligned geometries can produce unnatural trajectories, semantic drift, or adversarially aligned perturbations (Pegios et al., 2025; Zaher et al., 2026). However, computing local Riemannian geometry remains computationally prohibitive, particularly for large foundation models (Arvanitidis et al., 2022a; Chen et al., 2019). To bridge this divide, we introduce a framework situated between unconstrained Euclidean descent and explicit Riemannian optimization. Exploiting the local inverse structure available wherever the relevant representation differential has sufficient rank (Kumar et al., 2017), our approach steers counterfactual updates along the model manifold without explicitly constructing the full pullback geometry.

Deep neural networks map high-dimensional domains into complex geometric structures, often characterized as polyhedral complexes under piecewise-linear activations or Riemannian submanifolds under smooth nonlinearities (Montúfar et al., 2014; Chaukair et al., 2023). Emerging theoretical and empirical evidence further suggests that deep representations preserve meaningful local structure, making local inverse mappings tractable even when no global coordinate chart exists (Schulze et al., 2025; Nikolaou et al., 2026). Building on this perspective, we use the local right inverse of the task differential to construct an instantaneous task-relevant tangent space and steer counterfactual updates within it. In total, our contributions are threefold:

- **A geometric formulation of local counterfactual recourse.** We derive the counterfactual update as a metric-constrained inverse problem and prove that the resulting local right-inverse direction is the minimum-cost perturbation under the decoder-induced pullback geometry.
- **A sparse local-inverse optimization framework.** We show that counterfactual updates can be computed directly in the task-relevant subspace, avoiding explicit construction of the full pullback metric or tangent projector. We further derive truncation guarantees that allow non-sensitive directions to be removed, yielding sparse and numerically stable updates.
- **Implicit manifold alignment without explicit Riemannian optimization.** We demonstrate empirically that the resulting trajectories preserve counterfactual efficacy while improving trajectory fidelity and substantially reducing the computational cost of explicit Riemannian optimization across generative representation benchmarks.

The remainder of this paper is organized as follows: Section 2 reviews related literature, followed by Section 3, which establishes the mathematical nomenclature and preliminaries of the problem. In Section 4, we introduce our geometric counterfactual framework and our main algorithm. Section 5 details three empirical experiments evaluating our approach against foundational baselines. Finally, Section 6 provides discussion and concluding remarks.

2 Related Work

Counterfactual explanations. Counterfactual explanations describe how an input must change to obtain a desired prediction from a fixed model. Classical approaches formulate this as an optimization problem balancing target attainment against proximity to the observed input (Wachter et al., 2018; Guidotti, 2024; Verma et al., 2024), with subsequent work incorporating sparsity, diversity, actionability, and causal feasibility (Mothilal et al., 2020; Ustun et al., 2019; Karimi et al., 2021; Dandl et al., 2020; Poyiadzi et al., 2020; Chou et al., 2022; Karimi et al., 2022). These methods motivate our minimum-cost objective, but ambient-space proximity alone does not ensure that perturbations correspond to plausible or semantically meaningful changes.

Generative and manifold-constrained counterfactuals. To improve plausibility, counterfactual search can instead be constrained through learned generative representations (Pawelczyk et al., 2020; Joshi et al., 2019; Rodríguez et al., 2021; Khorram & Fuxin, 2022; Augustin et al., 2022). Optimiz-

ing a latent representation encourages counterfactuals to remain within data-supported regions, but Euclidean latent distance generally does not reflect the distortion induced by a nonlinear decoder.

Riemannian counterfactual geometry. A principled treatment of this distortion equips the latent space with a pullback metric induced by the decoder Jacobian (Tosi et al., 2014; Arvanitidis et al., 2018; Shao et al., 2018; Arvanitidis et al., 2022b;a). Recent counterfactual methods similarly exploit manifold structure, Riemannian latent traversals, or perceptual geodesics to produce geometrically aligned explanations (Tsiourvas et al., 2024; Pegios et al., 2025; Zaher et al., 2026). While existing approaches highlight the role of model-induced geometry, they rely on costly Riemannian solvers or explicit geometric quantities. Instead, we unify Euclidean and Riemannian regimes under a single framework.

Local inverse directions. Local Jacobian structure has also been used to characterize the geometry and controllable directions of deep generative representations (Wang & Ponce, 2021; Voynov & Babenko, 2020; Härkönen et al., 2020). Our framework connects this local inverse perspective with counterfactual optimization: we derive a metric-weighted inverse direction that minimizes decoder-induced cost while realizing a prescribed local prediction change. Its pushforward defines a tangent-aligned counterfactual update, while task-irrelevant directions can be truncated for sparse and numerically stable optimization.

3 PRELIMINARIES

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the input space, let $f : \mathcal{X} \to \mathcal{Y}$ be a predictive model, and let $x \in \mathcal{X}$ be an observed instance. Given a desired target set $\mathcal{Y}^\star \subseteq \mathcal{Y}$ and a cost function $d_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to [0, \infty)$, a *counterfactual explanation* is an input x_{cf} that attains the desired prediction while remaining as close as possible to x . Formally,

$$x_{\mathrm{cf}} \in \arg\min_{x' \in \mathcal{X}} d_{\mathcal{X}}(x, x') \qquad \text{subject to} \qquad f(x') \in \mathcal{Y}^\star. \tag{1}$$

Thus, a counterfactual explanation is a minimum-cost modification of the observed input whose prediction lies in the specified target region (Wachter et al., 2018; Guidotti, 2024).

For high-dimensional or structured data, direct optimization in the ambient input space may produce candidates that satisfy the predictive constraint in Eq. 1 but are implausible or semantically meaningless (Khorram & Fuxin, 2022; Augustin et al., 2022; Zaher et al., 2026). We therefore restrict the counterfactual search to a lower-dimensional latent representation of the data. Let $\mathcal{Z} \subseteq \mathbb{R}^m$, with $m \ll d$, be an open latent domain, and let

$$\phi : \mathcal{X} \to \mathcal{Z}, \qquad \psi : \mathcal{Z} \to \mathcal{X}$$

denote an encoder and decoder, respectively. For an observation $x \in \mathcal{X}$, we write $z_0 := \phi(x)$ for its latent representation. The decoder image $\mathcal{M} := \psi(\mathcal{Z}) \subseteq \mathcal{X}$ serves as a generative approximation of the set of plausible inputs. For the theoretical development, we assume that the observation is represented by the decoder, i.e. $x = \psi(z_0)$. In practice, when reconstruction is imperfect, one may instead use $\hat{x} := \psi(\phi(x))$ as the starting point, provided that the reconstruction preserves the predictive outcome of interest. Restricting the counterfactual search to the decoder image yields the latent optimization problem,

$$z_{\mathrm{cf}} \in \arg\min_{z' \in \mathcal{Z}} d_{\mathcal{X}}\big(\psi(z_0), \psi(z')\big) \qquad \text{subject to} \qquad f\big(\psi(z')\big) \in \mathcal{Y}^\star. \tag{2}$$

The corresponding counterfactual in the original input space is $x_{\mathrm{cf}} := \psi(z_{\mathrm{cf}})$. For convenience, we define the predictive model in latent coordinates as,

$$h := f \circ \psi : \mathcal{Z} \to \mathcal{Y}.$$

Because Euclidean distances in latent space do not generally reflect the perceptual shifts induced by the decoder in input space (Joshi et al., 2019), we equip $\mathcal{Z}$ with the Riemannian metric inherited from the ambient cost $d_{\mathcal{X}}(x, x') = \|x - x'\|_2^2$. Assuming ψ is continuously differentiable and locally immersive ($\operatorname{rank} J_\psi(z) = m$), the decoder image locally forms an m-dimensional manifold with tangent space $T_x\mathcal{M} = \operatorname{Im} J_\psi(z)$. The ambient Euclidean inner product pulls back to a Riemannian metric on $T_z\mathcal{Z}$ via $g_z(u, v) := \langle J_\psi(z)u, J_\psi(z)v \rangle_{\mathbb{R}^d}$, yielding the coordinate tensor,

$$G(z) = J_\psi(z)^\top J_\psi(z).$$

Because $J_\psi(z)$ has full column rank, $G(z)$ is strictly symmetric positive definite. A first-order Taylor expansion of the decoder then gives,

$$\|\psi(z+\delta z)-\psi(z)\|_2^2 = \delta z^\top G(z)\delta z + o\big(\|\delta z\|_2^2\big)\,, \tag{3}$$

confirming that the pullback metric $\|\delta z\|^2_{G(z)} := \delta z^\top G(z)\delta z$ measures latent perturbations directly by the local displacement they produce in input space.

4 Geometric Counterfactual Framework

We now derive our counterfactual construction from the latent formulation of the standard optimization problem in Eq. 2. Instead of solving this generally non-linear problem globally, we consider its local behavior around a current latent representation $z \in \mathcal{Z}$. Throughout this section, let $x = \psi(z)$, $h = f \circ \psi$, and let $G(z) = J_\psi(z)^\top J_\psi(z)$ denote the decoder-induced metric introduced in Section 3. We assume that h is continuously differentiable at the points under consideration. All proofs of the theorem, corollary, and propositions in this section are deferred to Appendix A.

Local Counterfactual Approximation Consider a nearby latent candidate $z' = z + \delta z$, where $\delta z \in T_z\mathcal{Z} \simeq \mathbb{R}^m$ is a small perturbation. By first-order expansion of the latent predictive map,

$$h(z+\delta z) = h(z) + J_h(z)\delta z + o(\|\delta z\|_2), \tag{4}$$

where $J_h(z)$ denotes the Jacobian of h at z. Let $b \in \mathbb{R}^q$ denote a desired local change in the model output (see Appendix B.1 for further remarks on b). To first order, requiring $h(z+\delta z) \approx h(z) + b$ amounts to imposing

$$J_h(z)\delta z = b. \tag{5}$$

At the same time, Eq. 3 shows that the corresponding change in input space is locally measured by $\delta z^\top G(z)\delta z$. Consequently, the local approximation of the latent counterfactual problem is

$$\begin{aligned} \underset{\delta z \in T_z\mathcal{Z}}{\text{minimize}} \quad & \frac{1}{2}\delta z^\top G(z)\delta z \\ \text{subject to} \quad & J_h(z)\delta z = b. \end{aligned} \tag{P1}$$

Problem P1 asks for the smallest decoder-induced perturbation that realizes the desired predictive change to first order. This constrained Problem P1 admits a closed-form solution. Below we present the optimal local counterfactual direction and through it derive our generalized local right-inverse map-based update.

Theorem 1 (Local metric-optimal counterfactual direction)**.** Let $G(z) \succ 0$ and suppose that the desired output displacement b is locally feasible, i.e., $b \in \operatorname{Im} J_h(z)$. Then Problem P1 has the unique solution,

$$\delta z^\star = G(z)^{-1}J_h(z)^\top \left[J_h(z)G(z)^{-1}J_h(z)^\top\right]^\dagger b. \tag{6}$$

Thus, $\delta z^\star$ is the minimum-$G(z)$-norm perturbation that realizes b. Furthermore, the operator

$$J_h^{\dagger_G}(z) := G(z)^{-1}J_h(z)^\top \left[J_h(z)G(z)^{-1}J_h(z)^\top\right]^\dagger, \tag{7}$$

defines the metric-weighted local (pseudo-)inverse of the predictive differential, yielding $\delta z^\star = J_h^{\dagger_G}(z)b$; it is a right inverse, $J_h J_h^{\dagger_G} = I_q$, whenever $\operatorname{rank} J_h(z) = q$. Crucially, Eq. 6 recovers a metric-weighted generalization of the Gauss–Newton step; its value here is not the closed form itself but that it makes the truncation guarantees of Proposition 1, to be presented below, possible.

Corollary 1 (Pullback Tangent Projection of Counterfactual Recourse)**.** Given the setting in Theorem 1 with ambient state $x = \psi(z)$, let $P(z) = J_\psi(z)J_\psi(z)^\dagger$ denote the orthogonal projector onto the ambient tangent space $T_x\mathcal{M}$. Under the composite map $h = f \circ \psi$, the optimal latent update satisfies $\delta z^\star = J_\psi(z)^\dagger\big(J_f(x)P(z)\big)^\dagger b$. Furthermore, if the metric is isotropic, $G(z) = c\,I_m$ with $c > 0$, the update reduces to $\delta z^\star = J_h^\dagger(z)\,b$.

Consequently, the counterfactual step restricts the predictive differential $J_f(x)$ to the generative tangent space $T_x\mathcal{M}$. In the isotropic case, it is evaluated directly via the task pseudoinverse $J_h^\dagger(z)$

without materializing $G(z)$ or $P(z)$; in general, $J_h^\dagger(z)\, b$ is the Euclidean-minimal latent step rather than the $G(z)$-minimal one. Pushing the optimal latent update $\delta z^\star$ through the decoder yields an ambient tangent direction $\delta x^\star = J_\psi(z)\delta z^\star \in T_x\mathcal{M}$ aligned with the target b. To ensure numerical stability against ill-conditioned directions and high-dimensional noise, we establish theoretical bounds that justify systematically truncating non-sensitive coordinates.

Proposition 1 (Sparsity-Driven Active Class Truncation Error Bound)**.** Let $\mathcal{A} \subset \{1, \ldots, q\}$ be the active class index set with inactive complement $\mathcal{I} = \{1, \ldots, q\} \setminus \mathcal{A}$. Let $\widetilde{J}(z) = J_h(z)G(z)^{-1/2}$ have nonzero rows $\tilde{g}_c^\top$, $c = 1, \ldots, q$ (for $G = I$, $\tilde{g}_c = \nabla_z h_c(z)$), and let $\widetilde{J}_\mathcal{A}$ be the rows indexed by $\mathcal{A}$. Normalize $\widehat{J} = D^{-1/2}\widetilde{J}$ with $D = \mathrm{diag}(\|\tilde{g}_c\|_2^2)$, so that $\widehat{J}\widehat{J}^\top = I_q + E$ with E symmetric and zero on the diagonal (its entries are the cosines between gradients). Assume the inter-class coupling condition $\rho(E) \le \rho_0 < 1$, and set $C_0 = (1 - \rho_0)^{-1}$ and $\hat{b}_c = b_c/\|\tilde{g}_c\|_2$. Then the active step $\delta z^\star_\mathcal{A} = G(z)^{-1/2}\widetilde{J}_\mathcal{A}^\dagger\, b_\mathcal{A}$ satisfies

$$\|\delta z^\star - \delta z^\star_\mathcal{A}\|_{G(z)} \le C_0^{1/2} \left(\sum_{j \in \mathcal{I}} \frac{|b_j|}{\|\tilde{g}_j\|_2} + \rho_0\, C_0\, \|\hat{b}_\mathcal{A}\|_2 \right). \tag{8}$$

The second term (active-to-inactive cross-talk) vanishes if $E_{\mathcal{I}\mathcal{A}} = 0$ or $b_\mathcal{A} = 0$.

We further derive a complementary bound governed by the spectral decomposition of the whitened Jacobian (Proposition 2, Appendix A), which controls the error from retaining only the top-k singular directions. Proposition 1 shows that truncating updates to active task coordinates incurs an error controlled by the inactive targets and inter-class coupling. By confining updates to $\mathrm{Im}(J_\mathcal{A}^\top)$, this active projection eliminates unconstrained drift along task-irrelevant singular modes, providing implicit manifold regularization without materializing the full pullback $G(z)$. Combining these bounds, we present our generalized stepping scheme in Algorithm 1.

Algorithm 1 Sparsity-Driven Local Inverse Counterfactual Update

Require: Initial $z_0 \in \mathbb{R}^m$, map $h = f \circ \psi$, target $y^\star$, threshold τ, step size η, max steps T
Ensure: Counterfactual z_{cf} and $x_{\text{cf}} = \psi(z_{\text{cf}})$

$z \leftarrow z_0$
for $t = 0, 1, \ldots, T-1$ **do**
  $b \leftarrow \textsc{project}(y^\star, h(z))$ ▷ Last layer direction correction (Appendix B.1 Eq.25)
  $\mathcal{A}_t \leftarrow \textsc{TargetSelector}(h(z), y^\star, \tau)$ ▷ Task-dependent active indices (Appendix B.2)
  Form shift $b_{\mathcal{A}_t} \in \mathbb{R}^{|\mathcal{A}_t|}$ and sub-Jacobian $J_{\mathcal{A}_t}(z) = \nabla_z h_{\mathcal{A}_t}(z) \in \mathbb{R}^{|\mathcal{A}_t| \times m}$
  if $G(z)$ is given **then**
    $\delta z^\star \leftarrow J_h^{\dagger_G}(z)\, b$ ▷ $J_h^{\dagger_G}$ as given in Eq. 7
  else
    Compute SVD: $J_{\mathcal{A}_t}(z) = U_{\mathcal{A}_t}\Sigma_{\mathcal{A}_t}V_{\mathcal{A}_t}^\top$
    Compute pseudoinverse step: $\delta z^\star \leftarrow J_{\mathcal{A}_t}(z)^\dagger b_{\mathcal{A}_t} = V_{\mathcal{A}_t}\Sigma_{\mathcal{A}_t}^\dagger U_{\mathcal{A}_t}^\top b_{\mathcal{A}_t}$
  end if
  $z \leftarrow z + \eta\, \delta z^\star$
end for
return $z_{\text{cf}} = z$, $x_{\text{cf}} = \psi(z)$

5 Experiments

To evaluate the out-of-the-box efficacy and generality of our framework, we conducted three computational experiments using frozen, zero-shot backbones without task-specific fine-tuning. We benchmarked our sparsity-driven local inverse counterfactual update (***Ours***, Algorithm 1) against two foundational paradigms: unconstrained Euclidean gradient descent (***Wachter et al.***, (Wachter et al., 2018)) and Riemannian Stochastic Gradient Descent (***RSGD***, (Pegios et al., 2025)). Because existing latent steering algorithms largely represent variants of these Euclidean and Riemannian formulations, this comparison directly isolates the fundamental mechanics of latent trajectory optimization. Our base formulation is denoted ***Ours***, while its Riemannian metric-weighted counterpart is

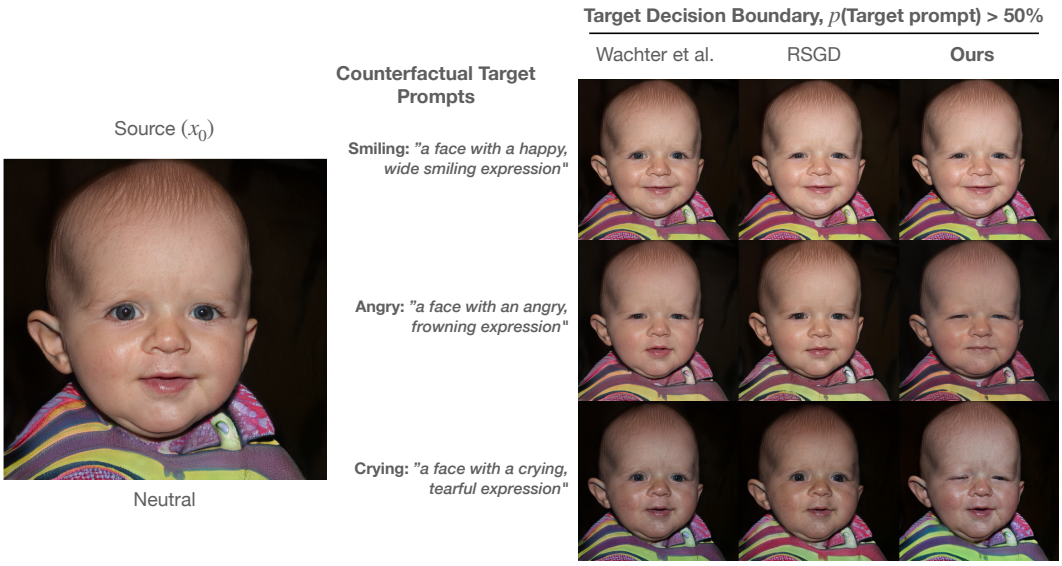

Figure 2: **Qualitative comparison of counterfactual generation at the target decision boundary.** (Left) Initial face generated by StyleGAN2 (seed 1000). (Right) Target steering of counterfactual explanations is driven by a zero-shot CLIP classifier (ViT-B/32) matching the synthesized images against target text prompts (rows: Smiling, Angry, Crying). The columns display counterfactuals produced by the methods: (*Wachter et al.*), (*RSGD*), and (*Ours*) evaluated at ($p \geq 50\%$).

denoted *Ours (+G)*. The particular parameter configurations are given in Appendix C. All pipelines were implemented in PyTorch and executed on an NVIDIA H100 GPU.

5.1 Qualitative Evaluation at the Decision Boundary

We begin with a simple experiment to generate counterfactual facial expressions for human-perceivable qualitative analysis. The counterfactual optimization was conducted directly over the latent space ($\mathcal{W}^+$) of a pretrained StyleGAN2 generator, initialized from a fixed neutral latent code w_0 (seed 1000, baby with a neutral face). We built our classifier by composing the StyleGAN2 generator with a multi-class zero-shot classifier built upon a frozen CLIP (ViT-B/32) visual encoder. The classifier computed scaled cosine similarities (see Appendix C.1) against descriptive prompts, such as, "a face with a happy, wide smiling expression", "a face with an angry, frowning expression", and "a face with a crying, tearful expression"; followed by a softmax over all candidate classes. Hence, this descriptive classifier could directly steer the optimization path through the latent space. In this experiment, we choose no metric to highlight the target-specific sparsity.

When comparing the resulting counterfactual images in Figure 2, we observe that while Euclidean gradient descent (*Wachter et al.*) and Riemannian gradient descent (*RSGD*) both reach the target boundary in latent space, their trajectories induce diffuse, low-frequency artifacts across the facial canvas and global lighting without engaging distinct facial musculature (see Appendix Fig. A1, for corresponding pixel attribution saliency maps). In contrast, our method (*Ours*) constrains updates to the span of the task Jacobian, producing localized structural edits: forehead tightening for anger, characteristic squints for anger and crying, and a realistic mouth contraction, all while preserving the identity of the base subject. Crucially, this fidelity is achieved entirely implicitly: without explicit identity regularizers, discriminator penalties, or costly metric pullbacks, our sparsity-driven inverse restricts optimization strictly to the decision-relevant coordinates, steering semantic emotion while keeping the subject's identity untouched.

Table 1: **Quantitative Benchmark across Target Breeds.** Metrics report mean ± standard deviation ($N = 200$ trajectories per solver). Target probability and flip rate assess counterfactual efficacy within 500 steps. Realism is measured via discriminator score $\mathcal{D}(x_{\text{cf}})$ and perceptual distance (LPIPS). Bold indicates best performance (excluding $\|\Delta w\|_2$). Metrics detailed in Appendix C.2

Target	Method	Target Prob. (%) ↑	Flip Rate (%) ↑	Steps to 50% ↓	$\|\Delta w\|_2$	$\mathcal{D}(x_{\text{cf}})$ ↑	$\Delta\mathcal{D}$ ↑	LPIPS ↓	Time (s) ↓
Cougar	Wachter et al.	25.0 ± 26.6	22.0%	454 ± 103	3.4 ± 0.4	-8.6 ± 1.0	-0.0 ± 0.6	**0.20 ± 0.06**	**33**
	RSGD	8.8 ± 16.4	8.0%	490 ± 45	2.5 ± 0.6	-8.6 ± 1.1	-0.0 ± 0.7	0.30 ± 0.10	200
	Ours	**32.6 ± 28.5**	24.0%	454 ± 91	5.1 ± 0.6	-8.7 ± 0.9	-0.1 ± 0.6	0.22 ± 0.05	92
	Ours (+G)	30.3 ± 27.1	**26.0%**	**432 ± 116**	4.7 ± 0.9	**-8.4 ± 0.9**	**0.1 ± 1.0**	0.50 ± 0.10	260
Lynx	Wachter et al.	88.7 ± 16.3	96.0%	103 ± 110	3.4 ± 0.4	-8.2 ± 0.9	0.3 ± 0.7	**0.25 ± 0.08**	**32**
	RSGD	71.8 ± 28.1	80.0%	222 ± 176	2.7 ± 0.5	**-8.0 ± 0.7**	0.5 ± 0.8	0.37 ± 0.08	201
	Ours	**92.6 ± 11.8**	**98.0%**	110 ± 97	4.7 ± 0.5	-8.2 ± 0.9	0.3 ± 0.6	0.26 ± 0.07	68
	Ours (+G)	83.8 ± 10.9	**98.0%**	**86 ± 104**	3.1 ± 0.8	**-8.0 ± 0.7**	**0.6 ± 0.8**	0.41 ± 0.08	239
Persian	Wachter et al.	70.4 ± 28.0	80.0%	188 ± 183	3.1 ± 0.4	-8.2 ± 0.9	0.4 ± 0.7	**0.26 ± 0.08**	**32**
	RSGD	57.7 ± 26.0	66.0%	268 ± 190	2.7 ± 0.5	-8.2 ± 0.8	0.3 ± 0.9	0.37 ± 0.08	201
	Ours	**77.1 ± 23.5**	88.0%	198 ± 153	4.5 ± 0.5	-8.3 ± 0.8	0.3 ± 0.7	**0.26 ± 0.07**	76
	Ours (+G)	71.9 ± 15.2	**92.0%**	**123 ± 139**	3.2 ± 0.9	**-8.1 ± 0.8**	**0.4 ± 1.0**	0.43 ± 0.11	241
Siamese	Wachter et al.	58.8 ± 30.6	66.0%	275 ± 186	3.5 ± 0.5	-8.4 ± 0.9	0.2 ± 0.7	**0.25 ± 0.08**	**33**
	RSGD	38.5 ± 28.9	46.0%	386 ± 146	2.8 ± 0.5	-8.5 ± 1.0	0.1 ± 0.8	0.36 ± 0.10	201
	Ours	**69.4 ± 30.9**	**80.0%**	267 ± 151	4.9 ± 0.5	-8.4 ± 0.9	0.1 ± 0.6	**0.25 ± 0.07**	85
	Ours (+G)	59.0 ± 23.5	**80.0%**	**227 ± 166**	4.0 ± 0.9	-8.4 ± 1.0	**0.2 ± 0.9**	0.47 ± 0.10	253
Overall	Wachter et al.	60.7 ± 34.7	66.0%	255 ± 199	3.4 ± 0.4	-8.3 ± 0.9	0.2 ± 0.7	**0.24 ± 0.08**	**33**
	RSGD	44.2 ± 34.6	50.0%	341 ± 182	2.7 ± 0.6	-8.3 ± 0.9	0.2 ± 0.8	0.35 ± 0.10	201
	Ours	**68.0 ± 33.1**	72.5%	257 ± 178	4.8 ± 0.6	-8.4 ± 0.9	0.1 ± 0.7	0.25 ± 0.07	80
	Ours (+G)	61.2 ± 27.4	**74.0%**	**217 ± 186**	3.7 ± 1.1	**-8.2 ± 0.9**	**0.3 ± 0.9**	0.45 ± 0.10	248

5.2 Quantitative Analysis Across Multi-Target Latent Trajectories, with and without Metric

To evaluate our framework both with and without metric conditioning (*Ours (+G)* and *Ours*) against standard non-metric (*Wachter et al.*) and Riemannian (*RSGD*) baselines, we generated multi-target counterfactual recourse trajectories across the AFHQ-cat manifold in StyleGAN2. Using rejection sampling, we sampled 50 distinct domestic tabby cat latents from the extended latent space $\mathcal{W}^+$, keeping all generated instances and backgrounds uncurated and unmodified (see Appendix C.2). Treating each sample as a fixed factual origin w_0, we computed counterfactual trajectories across all four optimization formulations toward four distinct target classes—Cougar, Lynx, Persian cat, and Siamese cat—guided by a frozen ResNet-50 classifier. This setup yielded 200 matched trajectory suites for comparative analysis (Table 1).

When examining the aggregate results across all 800 trajectories (Table 1, Overall; metrics detailed in Appendix C.2), our formulation effectively balances the strengths of unconstrained Euclidean descent and Riemannian manifold optimization even in the metric-free regime. Across the global metrics, *Ours (+G)* achieves the highest overall flip rate (74.0%) and fastest convergence (217 ± 186 steps), outperforming RSGD by 24.0 percentage points while improving discriminator realism ($\Delta\mathcal{D} = +0.3$ vs. $+0.2$). Crucially, even without explicit metric evaluation, *Ours* significantly surpasses both foundational baselines: it achieves a 72.5% flip rate and the top overall target probability ($68.0 \pm 33.1\%$), while matching the minimal perceptual distortion of Euclidean descent (LPIPS $= 0.25 \pm 0.07$ vs. 0.35 ± 0.10 for RSGD). By eliminating stochastic Hutchinson projections, *Ours* converges in 80 s, yielding a $2.5\times$ speedup over RSGD and $3.1\times$ over *Ours (+G)*, confirming that active subspace projection alone acts as an effective implicit manifold prior without explicit Riemannian compute.

To contextualize the elevated LPIPS of *Ours (+G)* (Table 1), we inspected representative counterfactuals across all four target classes (Figure 3a). While baselines attain lower LPIPS via diffuse, superficial texture perturbations, *Ours (+G)* navigates coherent, large-scale morphological shifts (e.g., cougar jaw restructuring) along the data manifold. To quantify off-target disruption across the trajectory, we tracked the distraction entropy over optimization steps (Figure 3b). Notably, *Ours* suppresses off-target entropy most aggressively after 150 steps(light-green curve). While *Ours (+G)* exhibits an early transient entropy rise, it rapidly descends and settles well below *RSGD* (dark-green curve). This confirms that across both variants, our framework mitigates diffuse off-manifold wandering. Conceptually, formulating the update via an active local right inverse restricts traversal

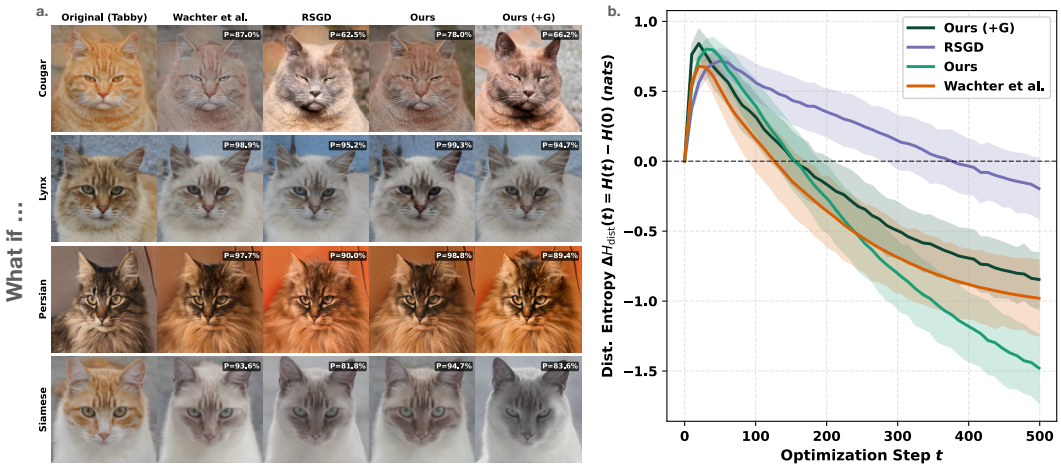

Figure 3: **Qualitative comparisons and latent optimization dynamics.** (a) Counterfactuals generated from a shared tabby source (first column) across four target classes with target probabilities $p(y_{\text{target}})$. Columns left to right: *Wachter et al.*, *RSGD*, *Ours*, and *Ours (+G)*. (b) Mean trajectory distraction entropy (N=800) with the 95% confidence interval sleeve.

strictly to directions the classifier attends to, altering only task-critical attributes. Enforcing sparsity on this update keeps the computation well-conditioned and lightweight, bypassing expensive metric pullbacks during standard shifts. Crucially, our generalized framework unifies the metric and metric-free regimes under a single mathematical formulation, establishing an upper- and lower-bound operating envelope for counterfactual recourse that span compute-optimal local edits, *Ours*, to curvature-regularized global transformations *Ours (+G)*.

5.3 Counterfactual Recourse Solves High-Dimensional Physics Inversion on Blackbox Foundation Models

To demonstrate cross-modal generality, we evaluate our framework on physical parameter inversion in semiconductor manufacturing: reconstructing nanoscale silicon grating profiles from optical scatterometry spectra (Figure 4a,b). Traditionally, this requires prohibitive numerical Maxwell solvers or lookup libraries. Instead, we instantiate the forward operator through a frozen, blackbox Tabular Foundation Model (TabFM), conditioned in-context on 32 simulated samples. Given target spectrum y^*, we reframe parameter retrieval as counterfactual recourse from an arbitrary initial grating x_0 to x^*. Treating TabFM as a parameter-agnostic surrogate, our method reconstructs the 6-parameter cross-sectional geometry without retraining or evaluating expensive numerical physics engines (Appendix C.3, D).

Because the physical design parameters are mutually independent, an orthogonal SVD transformation induces an isometric latent space ($G(z) = I$), under which RSGD collapses to standard Euclidean gradient descent. Since TabFM outputs spectra sequentially, we estimate local Jacobians via stochastic finite differences (Appendix B.5). Initializing across multiple seeds from the worst-case anchor, our sparse local inverse attains spectral convergence (MSE $\leq 5 \times 10^{-5}$) in an average of 69 ± 52 steps, achieving a $2.8\times$ speedup over RSGD (198 ± 141 steps; Figure 4c).

Tracking the effective condition number ($\tilde{\kappa}_{\text{eff}} = \sigma_{\max}/(\sigma_{\min} + \tau)$) reveals the underlying mechanism (Figure 4d), while both methods exhibit comparable median conditioning (~ 12.6), our method traverses a wider conditioning variance ($\text{std} = 1.41$). Rather than stalling along high-curvature, weakly sensitive optical modes, our adaptive channel masking and damped pseudoinversion isolate critical singular directions, suppressing stochastic noise and enabling large, stable updates through narrow loss corridors. Consequently, while RSGD exhibits severe trajectory jitter and slow descent, our solver converges rapidly without oscillatory drift (Figure 4e). Inspecting the recovered cross-sections confirms that both methods reach the identical physical profile (Seed 47, Figure 4f,g), demonstrating that our sparse local inverse navigates non-convex physical landscapes by an order of magnitude faster (19 vs. 87 steps, seed 47) without sacrificing geometric fidelity.

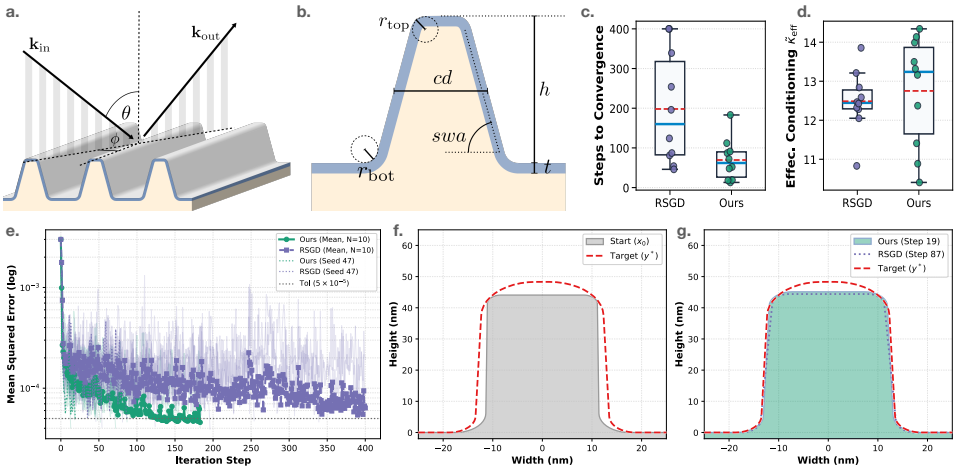

Figure 4: **Physical metrology inversion on a blackbox foundation model.** **(a)** Optical scatterometry schematic with wavevectors k_{in}, k_{out}, and polarization angles. **(b)** 6-parameter cross-section profile (cd, h, swa, t, r_{top}, r_{bot}). **(c)** Optimization steps to spectral convergence (MSE $\leq 5 \times 10^{-5}$) and **(d)** per-run median condition numbers $\tilde{\kappa}_{\text{eff}}$ across $N = 10$ seeds (solid blue: median, dashed red: mean). **(e)** Spectral MSE trajectories; solid curves show means ($N = 10$), faint dotted curves show Seed 47, and dotted threshold marks early stopping. **(f)** Initial profile (x_0, grey) against target (x^*, dashed red). **(g)** Final reconstructed cross-section for Seed 47 (Ours: green, RSGD: purple dotted, target: dashed red).

6 Discussion and Conclusion

We cast counterfactual recourse as a local inverse problem, proving the metric-optimal path reduces to a metric-weighted local right inverse (Theorem 1). More crucially, by pairing this inverse with truncation bounds for weakly sensitive coordinates (Propositions 1, 2), our Sparsity-Driven Local Inverse Update (Algorithm 1) achieves manifold-preserving recourse even without explicit metric evaluations or ambient tangent projections: restricting updates to the active differential $\text{Im}(J_{\mathcal{A}}^{\top})$ inherently projects out off-manifold drift. Unlike RSGD, which entangles all outputs through scalar loss reduction, our right-inverse permits uncoupled inactive dimensions ($E_{\mathcal{I}\mathcal{A}} \approx 0$) to be truncated directly prior to inversion. Hence, this formulation interfaces directly with frozen foundation representations, avoiding the bias or distortion of surrogate retraining and latent fine-tuning. Across three frozen backbones, our framework demonstrates superior efficiency and fidelity: steering StyleGAN2 facial expressions without identity penalties, minimizing distraction entropy while running $2.5\times$ faster in the active regime on cat transitions, and converging $2.8\times$ faster on ill-conditioned physics inversion.

Nonetheless, several limitations warrant future work. First, relying on local Jacobian approximations leaves the solver vulnerable to stochastic noise and local minima in high-curvature regions. Second, underdetermined target displacements b in classification can bias trajectory selection toward distinct local optima. Finally, while our framework exhibits empirical stability across seeds, comprehensive adversarial and distributional robustness evaluations were outside our scope; formalizing perturbation bounds under severe distribution shift remains an important future direction.

As opaque foundation models become central to high-stakes scientific, commercial, and decision-critical pipelines, interrogating their internal logic remains a paramount safety challenge. By bypassing costly geometric pullbacks, our framework democratizes actionable, manifold-preserving recourse across complex black-box architectures, establishing a scalable foundation for transparent model control. This property makes it applicable across modalities, including the emerging class of tabular foundation models, where counterfactuals exist as post-hoc procedures rather than geometry-aware intrinsic operations. A key future direction in this setting is the domain constraints, routinely imposed on tabular counterfactuals, which operate in the input space. Addressing these gaps across modalities is essential as mathematically grounded tools build the foundation for scalable governance in AI systems.

AI USE STATEMENT

In this work, we used generative AI tools for text editing, grammatical refinement, and prototyping implementation code for the empirical evaluation pipelines. We have not used generative AI tools for conceiving research ideas, formulating hypotheses, or deriving mathematical proofs; all theoretical propositions, lemmas, and theorems were formulated and proven solely by the authors. Other tasks involving automated generation of conceptual claims or literature discovery are not applicable to this work.

Additionally, we used generative AI tools for assisting with drafting and refining experimental visualization scripts. All written text was originally drafted by the authors, with generative AI serving exclusively as an editorial tool to enhance conciseness, formatting, and stylistic clarity. All AI-assisted code was manually inspected, debugged, validated, and executed by the authors to guarantee computational correctness and reproducibility across all benchmark runs. We take full responsibility for the final content of this work, including all text, mathematical claims, code, and empirical results.

ETHICS STATEMENT

This work evaluates latent counterfactual recourse using synthetic imagery synthesized by pretrained generative models (e.g., StyleGAN2) and frozen zero-shot classifiers (e.g., CLIP). Generative foundation models trained on web-scale datasets frequently inherit systemic societal biases, stereotyping, and skewed representation across demographic attributes, which can influence both the distribution of sampled starting points and the morphological trajectories traversed during latent optimization. The inclusion of these models serves exclusively as a controlled benchmark to evaluate the mathematical geometry and fidelity of optimization algorithms; it does not constitute an endorsement of the underlying data distributions, representation biases, or implicit associations present within these models. Our evaluations rely strictly on publicly available, pretrained checkpoints and synthetic artifacts without involving human subjects, personally identifiable data, or physical deployment, and we advocate for careful bias audits when deploying counterfactual steering in real-world decision-making systems.

REPRODUCIBILITY STATEMENT

To ensure complete empirical and theoretical reproducibility, all mathematical claims, formal propositions, and proofs are detailed with their foundational assumptions in Appendix A. Comprehensive descriptions of experimental protocols, surrogate conditioning configurations, optimization hyperparameter selections, and baseline implementations are provided in Appendix C and Appendix D. All experiments rely on deterministic random seeding and publicly available model checkpoints. An anonymized code repository containing the complete PyTorch implementation of our Sparsity-Driven Local Inverse Counterfactual Update (Algorithm 1), evaluation pipelines, and trajectory visualization scripts is provided in the attached supplementary material and will be made publicly available upon publication.

AUTHOR CONTRIBUTIONS

To be included in the camera-ready version.

ACKNOWLEDGMENTS

To be included in the camera-ready version.

REFERENCES

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: On the curvature of deep generative models. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SJzRZ-WCZ.

Georgios Arvanitidis, Bogdan M. Georgiev, and Bernhard Schölkopf. A prior-based approximate latent riemannian metric. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4634–4658. PMLR, 28–30 Mar 2022a.

Georgios Arvanitidis, Miguel González-Duque, Alison Pouplin, Dimitrios Kalatzis, and Soren Hauberg. Pulling back information geometry. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4872–4894. PMLR, 28–30 Mar 2022b.

Maximilian Augustin, Valentyn Boreiko, Francesco Croce, and Matthias Hein. Diffusion visual counterfactual explanations. *Advances in Neural Information Processing Systems*, 35:364–377, 2022.

Mustafa Chaukair, Christof Schütte, and Vikram Sunkara. On the activation space of relu equipped deep neural networks. *Procedia Computer Science*, 222:624–635, 2023.

Nutan Chen, Francesco Ferroni, Alexej Klushyn, Alexandros Paraschos, Justin Bayer, and Patrick van der Smagt. Fast approximate geodesics for deep generative models. In *International Conference on Artificial Neural Networks*, pp. 554–566. Springer, 2019.

Yu-Liang Chou, Catarina Moreira, Peter Bruza, Chun Ouyang, and Joaquim Jorge. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion*, 81:59–83, 2022.

Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. Multi-objective counterfactual explanations. In *International conference on parallel problem solving from nature*, pp. 448–469. Springer, 2020.

Nando Farchmin, Martin Hammerschmidt, Philipp-Immanuel Schneider, Matthias Wurm, Bernd Bodermann, Markus Bär, and Sebastian Heidenreich. Efficient global sensitivity analysis for silicon line gratings using polynomial chaos. In Bernd Bodermann, Karsten Frenner, and Richard M. Silver (eds.), *Modeling Aspects in Optical Metrology VII*, volume 11057 of *Proc. SPIE*, pp. 110570J. International Society for Optics and Photonics, SPIE, 2019. doi: 10.1117/12.2525978. URL https://doi.org/10.1117/12.2525978.

Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking: R. guidotti. *Data mining and knowledge discovery*, 38(5):2770–2824, 2024.

Martin Hammerschmidt, Martin Weiser, Xavier Garcia Santiago, Lin Zschiedrich, Bernd Bodermann, and Sven Burger. Quantifying parameter uncertainties in optical scatterometry using Bayesian inversion. In Bernd Bodermann, Karsten Frenner, and Richard M. Silver (eds.), *Modeling Aspects in Optical Metrology VI*, volume 10330 of *Proc. SPIE*, pp. 1033004. International Society for Optics and Photonics, SPIE, 2017. doi: 10.1117/12.2270596. URL https://doi.org/10.1117/12.2270596.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in neural information processing systems*, 33:9841–9850, 2020.

JCMwave. Jcmsuite: Finite-element simulation software for nano-optics and photonics. URL https://jcmwave.com/jcmsuite/. Accessed: 2026-09-18.

Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019. URL https://arxiv.org/abs/1907.09615.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: From counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 353–362. Association for Computing Machinery, 2021. doi: 10.1145/3442188.3445899. URL https://doi.org/10.1145/3442188.3445899.

Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5):1–29, 2022. doi: 10.1145/3527848. URL https://dl.acm.org/doi/10.1145/3527848.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8110–8119, 2020.

Saeed Khorram and Li Fuxin. Cycle-consistent counterfactuals by latent transformations. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10193–10202. IEEE, 2022.

Ramaravind Kommiya Mothilal, Divyat Mahajan, Chenhao Tan, and Amit Sharma. Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 652–663, 2021.

Abhishek Kumar, Prasanna Sattigeri, and Tom Fletcher. Semi-supervised learning with gans: Manifold invariance with improved inference. *Advances in neural information processing systems*, 30, 2017.

Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. *Advances in neural information processing systems*, 27, 2014.

Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 607–617. Association for Computing Machinery, 2020. doi: 10.1145/3351095.3372850. URL https://doi.org/10.1145/3351095.3372850.

Giorgos Nikolaou, Tommaso Mencattini, Donato Crisostomi, Andrea Santilli, Yannis Panagakis, and Emanuele Rodolà. Language models are injective and hence invertible. In *International Conference on Learning Representations*, volume 2026, pp. 57927–57984, 2026.

Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of The Web Conference 2020*, pp. 3126–3132. Association for Computing Machinery, 2020. doi: 10.1145/3366423.3380087. URL https://doi.org/10.1145/3366423.3380087.

Paraskevas Pegios, Aasa Feragen, Andreas Abildtrup Hansen, and Georgios Arvanitidis. Counterfactual explanations via riemannian latent space traversal. In *NeurIPS 2024 Workshop on Symmetry and Geometry in Neural Representations*, 2025. URL https://openreview.net/forum?id=wskfAkRK8M.

Rafael Poyiadzi, Kacper Sokol, Raúl Santos-Rodríguez, Tijl De Bie, and Peter Flach. Face: Feasible and actionable counterfactual explanations. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 344–350. Association for Computing Machinery, 2020. doi: 10.1145/3375627.3375850. URL https://dl.acm.org/doi/10.1145/3375627.3375850.

Pau Rodríguez, Massimo Caccia, Alexandre Lacoste, Lee Zamparo, Issam Laradji, Laurent Charlin, and David Vázquez. Beyond trivial counterfactual explanations with diverse valuable explanations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1056–1065, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Rodriguez_Beyond_Trivial_Counterfactual_Explanations_With_Diverse_Valuable_Explanations_ICCV_2021_paper.html.

Philipp Schneider. Parameter reconstruction 3: Bayesian least-square reconstruction, aug 2021a. URL https://jcmwave.com/blog/bayesian-least-square/.

Philipp Schneider. Parameter reconstruction 1: Theoretical introduction, aug 2021b. URL https://jcmwave.com/blog/introduction-parameter-reconstruction/.

Philipp Schneider. Parameter reconstruction 2: Variance-based sensitivity analysis, aug 2021c. URL https://jcmwave.com/blog/sensitivity-analysis/.

Philipp-Immanuel Schneider, Martin Hammerschmidt, Lin Zschiedrich, and Sven Burger. Using Gaussian process regression for efficient parameter reconstruction. In Vladimir A. Ukraintsev and Ofer Adan (eds.), *Metrology, Inspection, and Process Control for Microlithography XXXIII*, volume 10959 of *Proc. SPIE*, pp. 1095911. International Society for Optics and Photonics, SPIE, 2019. doi: 10.1117/12.2513268. URL https://doi.org/10.1117/12.2513268.

Kenrick Schulze, Felix Peppert, Christof Schütte, and Vikram Sunkara. Chimeric u-net–modifying the standard u-net towards explainability. *Artificial Intelligence*, 338:104240, 2025.

Hang Shao, Abhishek Kumar, and P. Thomas Fletcher. The riemannian geometry of deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 428–436, 2018.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Alessandra Tosi, Søren Hauberg, Alfredo Vellido, and Neil D. Lawrence. Metrics for probabilistic geometries. In Nevin L. Zhang and Jin Tian (eds.), *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pp. 800–808. AUAI Press, 2014.

Asterios Tsiourvas, Wei Sun, and Georgia Perakis. Manifold-aligned counterfactual explanations for neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 3763–3771. PMLR, 2024.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 10–19. Association for Computing Machinery, 2019. doi: 10.1145/3287560.3287566. URL https://doi.org/10.1145/3287560.3287566.

Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan Hines, John Dickerson, and Chirag Shah. Counterfactual explanations and algorithmic recourses for machine learning: A review. *ACM Computing Surveys*, 56(12):1–42, 2024.

Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *International conference on machine learning*, pp. 9786–9796. PMLR, 2020.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31 (2):841–887, 2018. URL https://ssrn.com/abstract=3063289.

Binxu Wang and Carlos R. Ponce. The geometry of deep generative image models and its applications. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=GH7QRzUDdXG.

Matthias Wurm, Stefanie Bonifer, Bernd Bodermann, and Jan Richter. Deep ultraviolet scatterometer for dimensional characterization of nanostructures: system improvements and test measurements. *Meas. Sci. Technol.*, 22(9):094024, aug 2011. doi: 10.1088/0957-0233/22/9/094024. URL https://doi.org/10.1088/0957-0233/22/9/094024.

Eslam Zaher, Dr Maciej Trzaskowski, Quan Nguyen, and Fred Roosta. Counterfactual explanations on robust perceptual geodesics. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=B5RBKM4vPY.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

A Proofs and Additional Results

Throughout, $B^\dagger$ denotes the Moore–Penrose pseudoinverse and $\|\delta z\|_G^2 = \delta z^\top G\, \delta z$.

Lemma 1 (Pseudoinverse facts)**.** Let $B \in \mathbb{R}^{p\times n}$. (i) $B^\dagger = B^\top (BB^\top)^\dagger$. (ii) $BB^\dagger b = b$ if and only if $b \in \operatorname{Im} B$. (iii) For every b, $B^\dagger b \in \operatorname{Im} B^\top = (\ker B)^\perp$, and $B^\dagger b$ is the minimum-norm minimizer of $\|Bw - b\|_2$.

Proof. These are standard properties of the Moore–Penrose pseudoinverse. □

Throughout the remainder of this appendix, fix z and write $G := G(z)$, $J := J_h(z)$.

Proof of Theorem 1. Write $\widetilde{J} := JG^{-1/2}$. Since $G \succ 0$, its positive square root $G^{1/2}$ is symmetric, positive definite, and invertible. Introduce the whitened variable $w := G^{1/2}\delta z$, so $\delta z = G^{-1/2}w$, $\frac{1}{2}\delta z^\top G \delta z = \frac{1}{2}\|w\|_2^2$, and the output constraint $J\delta z = b$ becomes $\widetilde{J}w = b$; thus Problem P1 is equivalent to

$$\min_{w\in\mathbb{R}^m} \tfrac{1}{2}\|w\|_2^2 \qquad \text{subject to} \qquad \widetilde{J}w = b. \tag{9}$$

Because $G^{-1/2}$ is invertible, $\operatorname{Im}\widetilde{J} = \operatorname{Im} J$, so $b \in \operatorname{Im} J$ guarantees feasibility, and by Lemma 1(ii) $w_0 := \widetilde{J}^\dagger b$ is feasible. Every feasible w can be written $w = w_0 + v$ with $v \in \ker \widetilde{J}$. Since $w_0 \in (\ker \widetilde{J})^\perp$ (Lemma 1(iii)), $\|w\|_2^2 = \|w_0\|_2^2 + \|v\|_2^2 \geq \|w_0\|_2^2$, with equality iff $v = 0$; hence $w^\star = \widetilde{J}^\dagger b$ is the unique minimizer. Transforming back gives $\delta z^\star = G^{-1/2}\widetilde{J}^\dagger b$. By Lemma 1(i), together with $\widetilde{J}^\top = G^{-1/2}J^\top$ and $\widetilde{J}\widetilde{J}^\top = JG^{-1}J^\top$,

$$\delta z^\star = G^{-1/2}\widetilde{J}^\top(\widetilde{J}\widetilde{J}^\top)^\dagger b = G^{-1}J^\top\left(JG^{-1}J^\top\right)^\dagger b, \tag{10}$$

that is, $\delta z^\star = G(z)^{-1}J_h(z)^\top\left[J_h(z)G(z)^{-1}J_h(z)^\top\right]^\dagger b$. Finally, $\|\delta z\|_G = \|w\|_2$, so the minimum Euclidean-norm property of $w^\star$ is exactly the minimum-$G(z)$-norm property of $\delta z^\star$. Hence $\delta z^\star$ is the unique minimum-$G(z)$-norm latent perturbation. □

Remark 1. Recall the operator $J_h^{\dagger_G}(z)$ from Eq. 7. Let $B := J_h(z)G(z)^{-1/2}$, so $J_hG^{-1}J_h^\top = BB^\top$ and $\operatorname{rank} B = \operatorname{rank} J_h(z)$ (as $G(z)^{-1/2}$ is invertible). By Lemma 1(i), $BB^\dagger = BB^\top(BB^\top)^\dagger = (BB^\top)(BB^\top)^\dagger$, hence $J_h(z)J_h^{\dagger_G}(z) = BB^\dagger$. By Lemma 1(ii), $BB^\dagger$ is the orthogonal projector onto $\operatorname{Im} B$, so $J_h(z)J_h^{\dagger_G}(z) = I_q$ if and only if $\operatorname{rank} J_h(z) = q$.

Remark 2. If $b \notin \operatorname{Im} J_h(z)$, the same formula gives, by Lemma 1(iii) applied to $\widetilde{J}$, the minimum-G-norm minimizer of $\|J_h\delta z - b\|_2$. Feasibility is therefore only needed for the constraint $J_h\delta z = b$ to hold exactly.

Proof of Corollary 1. From Theorem 1, $\delta z^\star = G^{-1}J_h^\top\left[J_hG^{-1}J_h^\top\right]^\dagger b$ with $G = J_\psi^\top J_\psi$. Because ψ is locally immersive ($\operatorname{rank} J_\psi = m$), G is non-singular and $J_\psi^\dagger = (J_\psi^\top J_\psi)^{-1}J_\psi^\top = G^{-1}J_\psi^\top$. The projector $P = J_\psi J_\psi^\dagger$ satisfies $P = P^\top = P^2$ and $J_\psi^\dagger P = J_\psi^\dagger J_\psi J_\psi^\dagger = J_\psi^\dagger$. By the chain rule $J_h = J_fJ_\psi$, so

$$G^{-1}J_h^\top = G^{-1}J_\psi^\top J_f^\top = J_\psi^\dagger J_f^\top, \tag{11}$$

$$J_hG^{-1}J_h^\top = J_fJ_\psi G^{-1}J_\psi^\top J_f^\top = J_fPJ_f^\top. \tag{12}$$

Hence $\delta z^\star = J_\psi^\dagger J_f^\top\left[J_fPJ_f^\top\right]^\dagger b$. Set $B := J_f(x)P(z)$. Then $BB^\top = J_fPJ_f^\top$ and $J_\psi^\dagger B^\top = J_\psi^\dagger PJ_f^\top = J_\psi^\dagger J_f^\top$, so by Lemma 1(i)

$$\delta z^\star = J_\psi^\dagger B^\top(BB^\top)^\dagger b = J_\psi(z)^\dagger\left(J_f(x)P(z)\right)^\dagger b. \tag{13}$$

For the pushforward, $J_\psi J_\psi^\dagger = P$ and $\operatorname{Im} B^\dagger = \operatorname{Im} B^\top = \operatorname{Im}(PJ_f^\top) \subseteq \operatorname{Im} P$, so $J_\psi \delta z^\star = PB^\dagger b = B^\dagger b \in T_x\mathcal{M}$. Finally, if $G(z) = c\, I_m$ with $c > 0$, then

$$G^{-1}J_h^\top \left[J_h G^{-1} J_h^\top\right]^\dagger = c^{-1} J_h^\top \left[c^{-1} J_h J_h^\top\right]^\dagger = J_h^\top \left[J_h J_h^\top\right]^\dagger = J_h^\dagger, \tag{14}$$

by Lemma 1(i); the scale cancels, so $\delta z^\star = J_h(z)^\dagger b$. □

Proof of Proposition 1. Write $\widetilde{J} := JG^{-1/2}$ with rows $\tilde{g}_c^\top$, let $D = \operatorname{diag}(\|\tilde{g}_c\|_2^2)$, $\widehat{J} = D^{-1/2}\widetilde{J}$ and $\widehat{\Gamma} = \widehat{J}\widehat{J}^\top = I_q + E$. Since E is symmetric, $\|E\|_2 = \rho(E) \le \rho_0 < 1$; hence the eigenvalues of $\widehat{\Gamma}$ lie in $[1-\rho_0, 1+\rho_0]$, $\widehat{\Gamma} \succ 0$ (so $\widetilde{J}$ has full row rank), $\|\widehat{\Gamma}^{-1}\|_2 \le C_0$, and the principal block $\widehat{\Gamma}_{\mathcal{A}\mathcal{A}} = \widehat{J}_{\mathcal{A}}\widehat{J}_{\mathcal{A}}^\top$ satisfies $\|\widehat{\Gamma}_{\mathcal{A}\mathcal{A}}^{-1}\|_2 \le C_0$.

Let $w = G^{1/2}\delta z$, so $\|\delta z\|_G = \|w\|_2$. With $\widetilde{J} = D^{1/2}\widehat{J}$ and full row rank,

$$w^\star = \widetilde{J}^\top(\widetilde{J}\widetilde{J}^\top)^{-1} b = \widehat{J}^\top D^{1/2}\big(D^{1/2}\widehat{\Gamma}D^{1/2}\big)^{-1} b = \widehat{J}^\top \widehat{\Gamma}^{-1}\hat{b}, \tag{15}$$

and in the same way $w_{\mathcal{A}} := \widetilde{J}_{\mathcal{A}}^\dagger b_{\mathcal{A}} = \widehat{J}_{\mathcal{A}}^\top \widehat{\Gamma}_{\mathcal{A}\mathcal{A}}^{-1}\hat{b}_{\mathcal{A}}$, so that $\delta z_{\mathcal{A}}^\star = G^{-1/2}w_{\mathcal{A}}$. Put $u = \widehat{\Gamma}^{-1}\hat{b}$ and $u^0 = (\widehat{\Gamma}_{\mathcal{A}\mathcal{A}}^{-1}\hat{b}_{\mathcal{A}},\, 0)$ (zero on $\mathcal{I}$). Then $w^\star - w_{\mathcal{A}} = \widehat{J}^\top(u - u^0)$ and, using $\widehat{\Gamma}_{\mathcal{I}\mathcal{A}} = E_{\mathcal{I}\mathcal{A}}$,

$$\widehat{\Gamma}(u - u^0) = (0,\ \hat{r}_{\mathcal{I}}), \qquad \hat{r}_{\mathcal{I}} := \hat{b}_{\mathcal{I}} - E_{\mathcal{I}\mathcal{A}}\widehat{\Gamma}_{\mathcal{A}\mathcal{A}}^{-1}\hat{b}_{\mathcal{A}}. \tag{16}$$

Since $(\widehat{\Gamma}^{-1})_{\mathcal{I}\mathcal{I}}$ is a principal block of $\widehat{\Gamma}^{-1}$, its largest eigenvalue is at most $\|\widehat{\Gamma}^{-1}\|_2 \le C_0$, and $\|E_{\mathcal{I}\mathcal{A}}\|_2 \le \|E\|_2 \le \rho_0$ and $\|\hat{b}_{\mathcal{I}}\|_2 \le \|\hat{b}_{\mathcal{I}}\|_1$, we obtain

$$\|\delta z^\star - \delta z_{\mathcal{A}}^\star\|_G^2 = \|w^\star - w_{\mathcal{A}}\|_2^2 = (u - u^0)^\top \widehat{\Gamma}(u - u^0) \tag{17}$$

$$= \hat{r}_{\mathcal{I}}^\top (\widehat{\Gamma}^{-1})_{\mathcal{I}\mathcal{I}}\, \hat{r}_{\mathcal{I}} \le C_0 \|\hat{r}_{\mathcal{I}}\|_2^2 \tag{18}$$

$$\le C_0 \left(\sum_{j \in \mathcal{I}} \frac{|b_j|}{\|\tilde{g}_j\|_2} + \rho_0 C_0 \|\hat{b}_{\mathcal{A}}\|_2 \right)^2, \tag{19}$$

which proves the claimed bound. If $E_{\mathcal{I}\mathcal{A}} = 0$ or $b_{\mathcal{A}} = 0$, then $\hat{r}_{\mathcal{I}} = \hat{b}_{\mathcal{I}}$ and the second term vanishes. □

Proposition 2 (Spectral Truncation Error)**.** Let $JG^{-1/2} = U\Sigma V^\top$ have rank $r \le \min(q, m)$, and let $\Pi_k = U_k U_k^\top$ project onto the top $k < r$ left singular vectors. For feasible $b \in \operatorname{Im} J$, the rank-$k$ truncated update $\delta z_k^\star = G^{-1/2} V_k \Sigma_k^{-1} U_k^\top b$ satisfies $J\delta z_k^\star = \Pi_k b$, with bounded metric norm $\|\delta z_k^\star\|_G \le \|\Pi_k b\|_2/\sigma_k$ and exact error:

$$\|\delta z^\star - \delta z_k^\star\|_G^2 = \sum_{i=k+1}^{r} \frac{|u_i^\top b|^2}{\sigma_i^2}. \tag{20}$$

Proof of Proposition 2. Write $\widetilde{J} := JG^{-1/2}$. Under the change of variables $w := G^{1/2}\delta z$, the metric norm satisfies $\|\delta z\|_G^2 = \|w\|_2^2$, while the output constraint becomes $J\delta z = \widetilde{J}w$. Let $\widetilde{J} = U\Sigma V^\top = \sum_{i=1}^r \sigma_i u_i v_i^\top$ with $\sigma_1 \ge \cdots \ge \sigma_r > 0$, where $U \in \mathbb{R}^{q\times r}$ and $V \in \mathbb{R}^{m\times r}$ have orthonormal columns $u_i \in \mathbb{R}^q$ and $v_i \in \mathbb{R}^m$, respectively. The exact minimum-norm solution is

$$w^\star = \widetilde{J}^\dagger b = \sum_{i=1}^{r} \frac{u_i^\top b}{\sigma_i} v_i, \tag{21}$$

and $\delta z^\star = G^{-1/2}w^\star$. For the rank-$k$ truncation, $V_k \in \mathbb{R}^{m\times k}$, $U_k \in \mathbb{R}^{q\times k}$, and $\Sigma_k = \operatorname{diag}(\sigma_1, \ldots, \sigma_k)$. Define $w_k^\star := V_k\Sigma_k^{-1}U_k^\top b = \sum_{i=1}^k \frac{u_i^\top b}{\sigma_i} v_i$, so that $\delta z_k^\star = G^{-1/2}w_k^\star$.

By orthonormality of the right-singular vectors ($V^\top V_k = \begin{bmatrix} I_k \\ 0 \end{bmatrix}$),

$$J\delta z_k^\star = \widetilde{J}w_k^\star = U\Sigma V^\top V_k \Sigma_k^{-1} U_k^\top b = U_k \Sigma_k \Sigma_k^{-1} U_k^\top b = U_k U_k^\top b = \Pi_k b. \tag{22}$$

Thus truncation exactly realizes the projection of b onto the retained left-singular subspace, and the residual is $b - J\delta z_k^\star = (I - \Pi_k)b$. Because $\Pi_k b \in \operatorname{Im} \widetilde{J}$, the minimum-norm solution of $\widetilde{J}w = \Pi_k b$ is $\widetilde{J}^\dagger \Pi_k b = V\Sigma^{-1}U^\top U_k U_k^\top b = w_k^\star$; since $w = G^{1/2}\delta z$ maps Euclidean norm to G-norm, $\delta z_k^\star$ is the minimum-G-norm perturbation satisfying Equation 22. Its norm satisfies

$$\|\delta z_k^\star\|_G^2 = \|w_k^\star\|_2^2 = \sum_{i=1}^{k} \frac{|u_i^\top b|^2}{\sigma_i^2} \leq \frac{1}{\sigma_k^2} \sum_{i=1}^{k} |u_i^\top b|^2 = \frac{\|\Pi_k b\|_2^2}{\sigma_k^2}, \tag{23}$$

so $\|\delta z_k^\star\|_G \leq \|\Pi_k b\|_2 / \sigma_k$. Finally, the discarded component in the whitened coordinate space is $w^\star - w_k^\star = \sum_{i=k+1}^{r} \frac{u_i^\top b}{\sigma_i} v_i$. Since $v_1, \ldots, v_r \in \mathbb{R}^m$ are orthonormal right singular vectors,

$$\|\delta z^\star - \delta z_k^\star\|_G^2 = \left\| G^{1/2}(\delta z^\star - \delta z_k^\star) \right\|_2^2 = \|w^\star - w_k^\star\|_2^2 = \sum_{i=k+1}^{r} \frac{|u_i^\top b|^2}{\sigma_i^2}. \tag{24}$$

□

B Algorithm 1 details

B.1 Activation Pull-Back

Our formulation requires a target displacement vector $b \in \mathbb{R}^q$. In unconstrained Euclidean regression, this reduces trivially to the residual $b = y^\star - y$. However, when the output layer incorporates a nonlinear activation η (e.g., a softmax for classification), the preimage of a target output $y^\star$ is an underdetermined level set, rendering the choice of b non-unique rather than nonexistent. To resolve this ambiguity constructively, we define b via the minimum-norm projection of the current pre-activation logits $h(z)$ onto the target level set $\mathcal{L}(y^\star) = \{\varepsilon \in \mathbb{R}^q \mid \eta(\varepsilon) = y^\star\}$:

$$b = \left(\underset{\substack{\varepsilon \in \mathbb{R}^q \\ \eta(\varepsilon) = y^\star}}{\arg\min} \|h(z) - \varepsilon\|_2 \right) - h(z). \tag{25}$$

B.2 Task-Adaptive Target Selection

To instantiate the active subspace, our framework relies on a task-dependent selector that determines the active output support $\Omega \subseteq \{1, \ldots, q\}$, isolating the coordinates along which the directional displacement $b \in \mathbb{R}^d$ is enforced. The choice of selection operator depends on the downstream task geometry:

- **Classification Regimes:** In multi-class classification, target selection is governed by logit separation. The active support Ω can simply be defined by isolating the target class index or retaining the top quantile of task-relevant logits corresponding to competing decision boundaries.
- **Magnitude-Based Regression Filtering:** For continuous regression where output dimensions share comparable physical units or normalized scales (as in our semiconductor metrology experiment; Appendix C.3), we construct Ω via absolute magnitude thresholding. Specifically, given target residual vector b, output channels are sorted by magnitude $|b_i|$, and the top p-quantile of coordinates is retained ($p \in (0, 1]$):

  $$\Omega = \left\{ i \in \{1, \ldots, q\} \;\middle|\; |b_i| \geq \text{Quantile}\left(|b|, 1 - p\right) \right\}. \tag{26}$$

  Restricting the pullback strictly to Ω ensures that small, noise-dominated residuals do not degrade condition numbers or destabilize the local right inverse.
- **Dynamic Sensitivity and Gradient Momentum:** Alternatively, when target coordinate scales diverge or b is unknown a priori, Ω can be determined dynamically along the trajectory. One robust variant isolates coordinates exhibiting maximal step-to-step velocity, $|y_t - y_{t-1}|$, prioritizing channels with the highest active response. When the logit displacement b is explicitly available via activation pullback (Eq. 25), scaling b by the inverse feature variances prior to quantile truncation yields a scale-invariant active set.

B.3 Stochastic Hutchinson Pullback Metric for RSGD.

For the Riemannian baseline (*RSGD*), traversing the manifold under the Riemannian metric induced by the generator $G : \mathcal{W}^+ \to \mathcal{X}$ requires evaluating the pullback metric tensor $G(w) = J_G(w)^\top J_G(w) \in \mathbb{R}^{d\times d}$, where $J_G(w) = \partial G(w)/\partial w$. Because computing and inverting J_G explicitly at each step is intractable for high-dimensional synthesis outputs ($1024 \times 1024 \times 3$), we implemented a regularized stochastic Hutchinson estimator.

Specifically, at each iteration, we drew $K = 16$ Rademacher random vector projections $v_i \in \{-1, +1\}^{C\times H\times W}$ $(i = 1, \ldots, K)$ and computed the vector-Jacobian products via reverse-mode automatic differentiation:

$$g_i = \nabla_w \big\langle G(w),\, v_i \big\rangle = J_G(w)^\top v_i. \tag{27}$$

Stacking these scaled rows yields the low-rank projection matrix $\tilde{J} = \frac{1}{\sqrt{K}}[g_1, \ldots, g_K]^\top \in \mathbb{R}^{K\times d}$. The regularized pullback metric is approximated as:

$$G(w) \approx \tilde{J}^\top \tilde{J} + \lambda_{\text{damp}} I_d, \tag{28}$$

with $K = 16$ projections and damping parameter $\lambda_{\text{damp}} = 10^{-3}$. To stably compute the inverse metric $G(w)^{-1}$ in the presence of near-zero singular directions, we performed an eigendecomposition $G(w) = Q\Lambda Q^\top$, where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d)$ with eigenvalues sorted in ascending order ($\lambda_d = \lambda_{\max}$). A truncated spectral filter was applied with relative tolerance $\tau_{\text{spec}} = 0.02$:

$$\tilde{\Lambda}^{-1}_{jj} = \begin{cases} \frac{1}{\lambda_j}, & \text{if } \lambda_j \geq \tau_{\text{spec}} \cdot \lambda_{\max}, \\ 0, & \text{otherwise,} \end{cases} \tag{29}$$

yielding the regularized inverse metric $G(w)^{-1} = Q\tilde{\Lambda}^{-1}Q^\top$ used to precondition the Riemannian gradient descent step.

B.4 Pixel-Level Saliency Maps

To visually isolate semantic edits from spurious background drift, pixel-level saliency maps were generated by taking the channel-averaged absolute difference between the recourse image at the decision boundary $x_{\text{boundary}} = G(z_{\text{boundary}})$ and the factual origin $x_0 = G(z_0)$:

$$\Delta x(u, v) = \frac{1}{3}\sum_{c=1}^{3} \big|x_{\text{boundary}}(c, u, v) - x_0(c, u, v)\big|. \tag{30}$$

To ensure fair and rigorous visual comparison across all evaluated solvers without arbitrary local normalization, heatmaps were scaled against a shared global upper bound:

$$v_{\max} = \max_{m\in\mathcal{M}} \text{Quantile}_{0.99}\left(\Delta x^{(m)}\right), \tag{31}$$

where $\mathcal{M}$ denotes the set of evaluated methods. Intensities were rendered on a fixed scale $[0, v_{\max}]$ using a custom white-zero colormap, suppressing near-zero noise while faithfully exposing localized facial distortions.

B.5 Black-Box Approximation of the Jacobian

The update in Eq. 7 requires the local predictive Jacobian $J_h(z)$, which may be unavailable when the predictive model f is accessed only through queries. We therefore consider a setting in which the decoder ψ is differentiable, while f is available only through real-valued forward evaluations. The decoder-induced metric $G(z)$ remains directly computable, whereas $J_h(z)$ must be estimated locally.

Local Jacobian Estimation. At a latent point $z \in \mathcal{Z}$, draw N local perturbations $\varepsilon_1, \ldots, \varepsilon_N$ satisfying $\|\varepsilon_i\|_2 \leq \rho$, and define $z_i = z + \varepsilon_i$. Collect the latent and predictive displacements as

$$\Delta Z = [\varepsilon_1, \ldots, \varepsilon_N], \qquad \Delta H = [h(z_1) - h(z), \ldots, h(z_N) - h(z)]. \tag{32}$$

By Eq. 4,

$$\Delta H = J_h(z)\Delta Z + R_\rho, \tag{33}$$

where R_ρ contains higher-order terms. A local least-squares estimate of the predictive Jacobian is therefore

$$\widehat{J}_h(z) = \Delta H\, \Delta Z^\dagger. \tag{34}$$

If the sampled perturbations span the latent tangent space, this approximates the full local differential; otherwise, it approximates its restriction to the sampled subspace. The neighborhood radius ρ controls the trade-off between local linearization accuracy and numerical conditioning.

Black-Box Counterfactual Update. In the black-box setting, we apply the construction of Proposition 2 with $J_h(z)$ replaced by the estimate $\widehat{J}_h(z)$, retaining k singular directions to obtain the update $\widehat{\delta z}_k$. The counterfactual path is then updated as

$$z_{t+1} = z_t + \eta_t \widehat{\delta z}_{t,k}, \qquad x_{t+1} = \psi(z_{t+1}), \tag{35}$$

with the local Jacobian estimate, metric, and spectral decomposition recomputed along the trajectory. The truncation rank k controls the trade-off between predictive coverage and stability. In practice, we retain either a fixed number of directions or those satisfying $\widehat{\sigma}_i \geq \tau \widehat{\sigma}_1$ for a relative threshold $\tau > 0$. Since truncation realizes the projected change $\widehat{P}_k b$ rather than necessarily the full displacement b, any remaining predictive residual is addressed by subsequent local updates. Since optimization is performed in latent space, on-manifold feasibility is enforced by construction. Provided that $z_{t+1} \in \mathcal{Z}$, decoding yields $x_{t+1} = \psi(z_{t+1}) \in \psi(\mathcal{Z}) = \mathcal{M}$. Thus, unlike ambient-space updates, the proposed update does not require an additional projection onto the data manifold.

C Methods for Numerical Experiments

C.1 Qualitative Experiment methods

Latent Initialization and Rejection Sampling. To ensure an unambiguous factual baseline for qualitative evaluation, candidate face portraits were generated using the official pre-trained StyleGAN2 model trained on FFHQ at 1024×1024 resolution (`stylegan2-ffhq-config-f.pkl`) within the extended latent space $\mathcal{W}^+ \subset \mathbb{R}^{18\times 512}$. We performed rejection sampling across candidate random seeds using a frozen CLIP (ViT-B/32) visual encoder via the official OpenAI package. Specifically, synthesized candidates were evaluated against the source text prompt:

1. `"a face with a neutral or frowning expression"`
2. `"a face with a smiling and happy expression"`

Candidates were accepted as valid factual origins w_0 only when the baseline classifier assigned dominant confidence to the neutral prompt, ensuring the base subject exhibited no pre-existing target emotion. This sampling pipeline isolated the representative neutral portrait at seed 1000 shown in Figure 2.

Zero-Shot CLIP Classification and Steering Objective. To guide recourse trajectories through $\mathcal{W}^+$, we constructed a multi-class zero-shot classifier by composing the pre-trained StyleGAN2 generator $G(w)$ directly with the frozen CLIP (ViT-B/32) visual encoder. Denoting normalized image embeddings by $f_{\text{img}}(G(w))$ and normalized text embeddings for candidate prompts by $f_{\text{text}}(t_k)$, logit predictions across the emotional spectrum were computed via scaled cosine similarity:

$$h_k(w) = \tau \cdot \langle f_{\text{img}}(G(w)), f_{\text{text}}(t_k) \rangle, \quad k \in \{0, \ldots, 5\}, \tag{36}$$

where $\tau = 100$ denotes the default CLIP logit scaling factor. The candidate emotion prompts were specified as:

[0] `"a face with a neutral, blank expression"` (*Neutral*, Source anchor)

[1] `"a face with a happy, wide smiling expression"` (*Smiling*)

[2] `"a face with an angry, frowning expression"` (*Angry*)

[3] `"a face with a crying, tearful expression"` (*Crying*)

[4] `"a face with a surprised, shocked expression"` (*Surprised*)

[5] `"a face with a disgusted, contemptuous expression"` (*Disgusted*)

Passing the complete logit vector $h(w) \in \mathbb{R}^6$ through a softmax yields class posterior probabilities $p(y = k \mid w) = \mathrm{softmax}(h(w))_k$. Steering toward a specified target emotion $y_{\text{target}} \in \{1, \ldots, 5\}$ was executed until reaching the decision boundary threshold $p(y_{\text{target}} \mid w) \geq 0.50$ (or exhausting the iteration budget).

Optimization Hyperparameters and Baselines. All optimization algorithms were executed for a maximum of MAX_STEPS $= 300$ with an identical fixed step size of $\eta = 0.01$. This step size was calibrated by determining the largest value ensuring stable, non-divergent convergence for *RSGD* over 300 steps, providing an on-manifold baseline reference that was kept uniform across all compared solvers. Optimization terminated early as soon as the target confidence satisfied $p(y_{\text{target}} \mid z) \geq 0.70$; otherwise, the final iterate at step 300 was selected as the candidate recourse. For *RSGD*, the generator's pullback metric $G(z)^{-1}$ was estimated at each step using the regularized stochastic Hutchinson estimator detailed in Appendix B.3. We intentionally adopt the canonical pullback metric as a representative, foundational Riemannian baseline; while more sophisticated metric structures (e.g., Zaher et al., 2026) could be integrated, our primary objective is to establish the fundamental behavior of our proposed framework against standard geometric recourse. For our metric-free formulation (*Ours*), the update was evaluated directly through the sparse local right inverse without additional identity or discriminator penalties. Specifically, the target displacement vector was set to $b = e_{y_{\text{target}}} - p(y \mid z)$, where $e_{y_{\text{target}}}$ is the one-hot target indicator and $p(y \mid z)$ is the current predicted probability vector. The target selector retained the top 90% of output dimensions sorted by post-softmax classifier confidence, filtering out the lowest-10% probability channels to isolate the task-relevant active subspace. Pixel-level saliency heatmaps comparing localized semantic edits against background drift were generated following the protocol described in Appendix B.4.

C.2 Multi-Target Experiment methods

Dataset, Pretrained Backbones, and Rejection Sampling. For the multi-target experiment, we utilized a pretrained StyleGAN2 generator (Karras et al., 2020) trained on the AFHQ-cat benchmark at 512×512 resolution. As the downstream task model, we employed a frozen, standard ImageNet-pretrained ResNet-50 classifier operating over the full $K = 1000$ label space. Factual seeds were synthesized from the latent prior $z_0 \sim \mathcal{N}(0, I)$ and filtered via rejection sampling, accepting candidates only when the predicted confidence for the source class satisfied $p(y_{\text{tabby}} \mid z_0) \geq 0.50$ (ImageNet index 281). This procedure was repeated until exactly $N = 50$ distinct factual origins were collected.

While the classifier output and probability normalization spanned all 1000 ImageNet categories, we evaluated counterfactual recourse along four semantically distinct feline target trajectories:

$$y_{\text{target}} \in \{\text{Cougar (286), Lynx (287), Persian cat (283), Siamese cat (284)}\}. \tag{37}$$

Consistent with our previous protocol, the displacement vector for our right-inverse formulations was defined over the full 1000-dimensional probability simplex as the instantaneous target residual $b = e_{y_{\text{target}}} - p(y \mid z) \in \mathbb{R}^{1000}$, where $e_{y_{\text{target}}}$ is the 1000-dimensional one-hot target indicator vector and $p(y \mid z)$ represents the complete predicted softmax probability distribution at latent state z.

Specifically, to isolate the task-relevant active subspace over the 1000-dimensional output space, the target selector dynamically retained the indices corresponding to the top 90% of cumulative post-softmax classifier confidence, while explicitly enforcing the inclusion of the source class index (y_{tabby}) alongside the active target y_{target}. All remaining low-probability channels were zeroed out, preventing gradient or right-inverse updates from dispersing effort across irrelevant background categories.

Optimization Hyperparameters and Baselines. All counterfactual searches were conducted for a maximum of MAX_STEPS $= 500$ across all evaluated methods. We compare against standard

gradient-based optimization (*Wachter et al.* and *RSGD*), alongside two formulations of our framework: a metric-free Euclidean variant (*Ours*) and a metric-equipped variant (*Ours (+G)*). For methods operating without an explicit metric tensor, including *Wachter et al.* and *Ours*—updates were executed via gradient descent with a conservative step size of $\eta = 0.08$, matching the threshold calibrated to ensure stable, on-manifold trajectories in *RSGD*. This conservative step size prevents unguided trajectories from departing the data support when the local distortion of the generator is unknown. In contrast, because *Ours (+G)* incorporates the intrinsic pullback metric $G(z)^{-1}$ (Appendix B.3) to account directly for local geometric stretching, it eliminates the need for conservative learning rate tuning; the right-inverse update was applied directly, bounded only by a step-length cap $\|\Delta z\|_2 \leq 1.5$. A key architectural advantage of our framework is that the metric-weighted local right-inverse explicitly scales the update vector via the inverted forward differential $[J_h G^{-1} J_h^\top]^\dagger$, resolving the local step scale analytically without requiring sensitive learning rate tuning.

Evaluation Metrics. To evaluate optimization efficacy, geometric fidelity, and perceptual realism across all $N = 50$ seeds and target classes, we report the following quantitative metrics:

- **Target Probability (%):** The final post-softmax probability assigned by the classifier to the target breed at the terminal step of optimization, $p(y_{\text{target}} \mid z_{\text{cf}}) \times 100$.
- **Flip Rate (%):** The percentage of optimization trajectories that successfully crossed the decision boundary, defined as reaching a predicted target confidence of at least 50%:

$$\text{Flip Rate} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\big(\max_t p(y_{\text{target}} \mid z_t^{(i)}) \geq 0.50 \big) \times 100. \tag{38}$$

- **Steps to 50%:** The earliest iteration index $t \in [1, \text{MAX_STEPS}]$ at which $p(y_{\text{target}} \mid z_t) \geq 0.50$. If a trajectory failed to attain 50% target confidence within the step cap, the value was imputed at the ceiling of MAX_STEPS $= 400$.
- **Latent Drift ($\|\Delta z\|_2$):** The Euclidean distance traversed in the generator's latent space between the terminal recourse z_{cf} and the factual origin z_0:

$$\|\Delta z\|_2 = \|z_{\text{cf}} - z_0\|_2. \tag{39}$$

- **Discriminator Score $\mathcal{D}(x_{\text{cf}})$ and Relative Gain $\Delta\mathcal{D}$:** To quantify sample realism and fidelity to the learned data distribution, synthesized counterfactuals $x_{\text{cf}} = G(z_{\text{cf}})$ were evaluated using the pretrained StyleGAN2 discriminator $\mathcal{D}(\cdot)$ (measured prior to the final logistic activation). In addition to the raw output $\mathcal{D}(x_{\text{cf}})$, we evaluate the shift relative to the factual origin:

$$\Delta\mathcal{D} = \mathcal{D}(x_{\text{cf}}) - \mathcal{D}(x_0), \tag{40}$$

  where positive values denote that the generated recourse remained within or improved upon the generator's realistic manifold support.
- **Perceptual Distance (LPIPS):** The perceptual distortion between the factual image x_0 and the counterfactual x_{cf} measured via Learned Perceptual Image Patch Similarity (Zhang et al., 2018) using a VGG backbone:

$$\text{LPIPS} = \mathcal{L}_{\text{LPIPS}}(x_0, x_{\text{cf}}). \tag{41}$$

- **Wall-Clock Time (s):** The mean optimization execution time in seconds per recourse instance, measured across identical hardware.
- **Distraction Entropy and Excess Entropy (ΔH_{dist}):** To quantify whether recourse trajectories preserve semantic specificity or inadvertently induce classifier confusion across irrelevant classes, we monitor the Shannon entropy of the predictive distribution along the trajectory. At any iteration step t, the predictive entropy over all $K = 1000$ ImageNet categories is defined as:

$$H(t) = -\sum_{k=1}^{K} p(y_k \mid w_t) \ln \big(p(y_k \mid w_t) + \epsilon \big), \tag{42}$$

where $\epsilon = 10^{-12}$ ensures numerical stability. To isolate the entropy directly induced by optimization from the baseline uncertainty of the factual origin, we measure the *Distraction Entropy*:

$$\Delta H_{\text{dist}}(t) = H(t) - H(0), \tag{43}$$

where $H(0)$ denotes the predictive entropy evaluated at the factual seed w_0. While an ideal, focused counterfactual transitions cleanly between the source and target classes (yielding low or negative entropy as target certainty crystallizes), unconstrained or misaligned trajectories disperse probability mass across unrelated classes, reflected by elevated $\Delta H_{\text{dist}}(t) > 0$.

C.3 Metrology Experiment methods

Dataset and Context Set Selection via In-Context Clustering. We evaluate our framework on an inverse metrology benchmark comprising $N = 4{,}000$ simulation points generated via a Finite Element Method (FEM) solver. To calibrate the surrogate forward model, the continuous target variables $Y \in \mathbb{R}^{N \times d_y}$ were zero-centered and normalized to unit variance. We then partitioned the target domain using k-means clustering across dyadic cluster sizes $k \in \{2^2, 2^3, \ldots, 2^6\}$. For each candidate k, the closest simulation points to the respective cluster centroids were selected as an in-context support set $\mathcal{D}_{\text{ctx}} \subset \mathcal{D}$, while the remaining $N - k$ points served as an out-of-context test split.

Evaluation was conducted using a tabular foundation model (*TabFM*). We observed that a context budget of $k = 32$ centroids was sufficient to achieve 90% prediction accuracy over the held-out test split. Consequently, this calibrated 32-point support set was fixed as the global in-context prompt for all subsequent optimization trajectories.

Latent Parameterization and Isometry. Because the physical input parameters of the FEM lookup table represent decoupled, independent operational variables, we construct the latent space $\mathcal{Z}$ by applying a Singular Value Decomposition (SVD) across the standardized input coordinates to whiten and scale the physical dimensions. By construction, this orthogonal transformation forms a Riemannian isometry between the normalized physical parameter space and the latent manifold $\mathcal{Z}$. Consequently, the intrinsic Riemannian metric tensor collapses to the Euclidean identity:

$$G(z) = I. \tag{44}$$

Under this isometric geometry, Riemannian Stochastic Gradient Descent (*RSGD*) simplifies directly to standard gradient updates without pullback distortion corrections.

Inverse Recourse Formulation and Optimization. Unlike classical counterfactual recourse—which penalizes distance to the factual origin z_0 to preserve proximity—inverse metrology treats z_0 merely as an initialization, seeking any physically viable configuration $z^\star \in \mathcal{Z}$ that satisfies target measurement specifications $y^\star$. We therefore evaluate *RSGD* against our sparse local right-inverse formulation (*Ours*) without latent proximity or identity regularizers. Updates were executed directly toward the standardized target displacement $b = y^\star - \hat{y}(z)$ until the predicted surrogate output converged within a tolerance of the target state.

Software, Seeds, and Absorbing Boundary Constraints. All physical inversion benchmarks were executed using *TabFM* (version 1.0.1) over ten evaluation seeds $s \in \{40, 41, \ldots, 49\}$. To prevent optimization trajectories from exploring non-physical geometries, both solvers enforced identical absorbing box constraints across the 6-parameter cross-sectional vector $\theta = [cd, h, swa, t, r_{\text{top}}, r_{\text{bot}}]^\top$. The feasible physical parameter bounds $[\theta_{\min}, \theta_{\max}]$ were defined as:

$$\theta_{\min} = [15.0\,\text{nm},\ 30.0\,\text{nm},\ 80.0^\circ,\ 0.0\,\text{nm},\ 0.1\,\text{nm},\ 0.1\,\text{nm}]^\top, \tag{45}$$

$$\theta_{\max} = [35.0\,\text{nm},\ 60.0\,\text{nm},\ 89.9^\circ,\ 15.0\,\text{nm},\ 15.0\,\text{nm},\ 15.0\,\text{nm}]^\top. \tag{46}$$

At each step, coordinates in the standardized latent space z were mapped to physical units via the inverse standardizer $\mathcal{T}^{-1}$, projected onto the feasible domain via hard box-clipping, and transformed back:

$$\Pi_{\mathcal{C}}(z) = \mathcal{T}\Big(\,\text{clip}\left(\mathcal{T}^{-1}(z),\, \theta_{\min},\, \theta_{\max}\right)\Big). \tag{47}$$

Optimization Dynamics, Hyperparameters, and Stopping Criteria. Both methods estimated the forward operator Jacobian numerically using stochastic finite differences as formulated in Appendix B.5, evaluated with $M = 42$ Gaussian probe samples at perturbation scale $\sigma = 10^{-2}$. Counterfactual searches were run up to MAX_STEPS $= 400$ with an early-stopping convergence criterion of MSE$(y^\star, \hat{y}(z)) \leq 5 \times 10^{-5}$. For the baseline (*RSGD* with $G(z) = I$), updates were executed via gradient descent with learning rate $\eta = 0.25$, calibrated to maintain stability under numerical Jacobian estimation noise. For our method (*Ours*), the right-inverse step was applied directly with a maximum step-norm ceiling of $\|\delta z\|_2 \leq 1.5$, preserving the exact trust-region threshold used in the vision benchmarks.

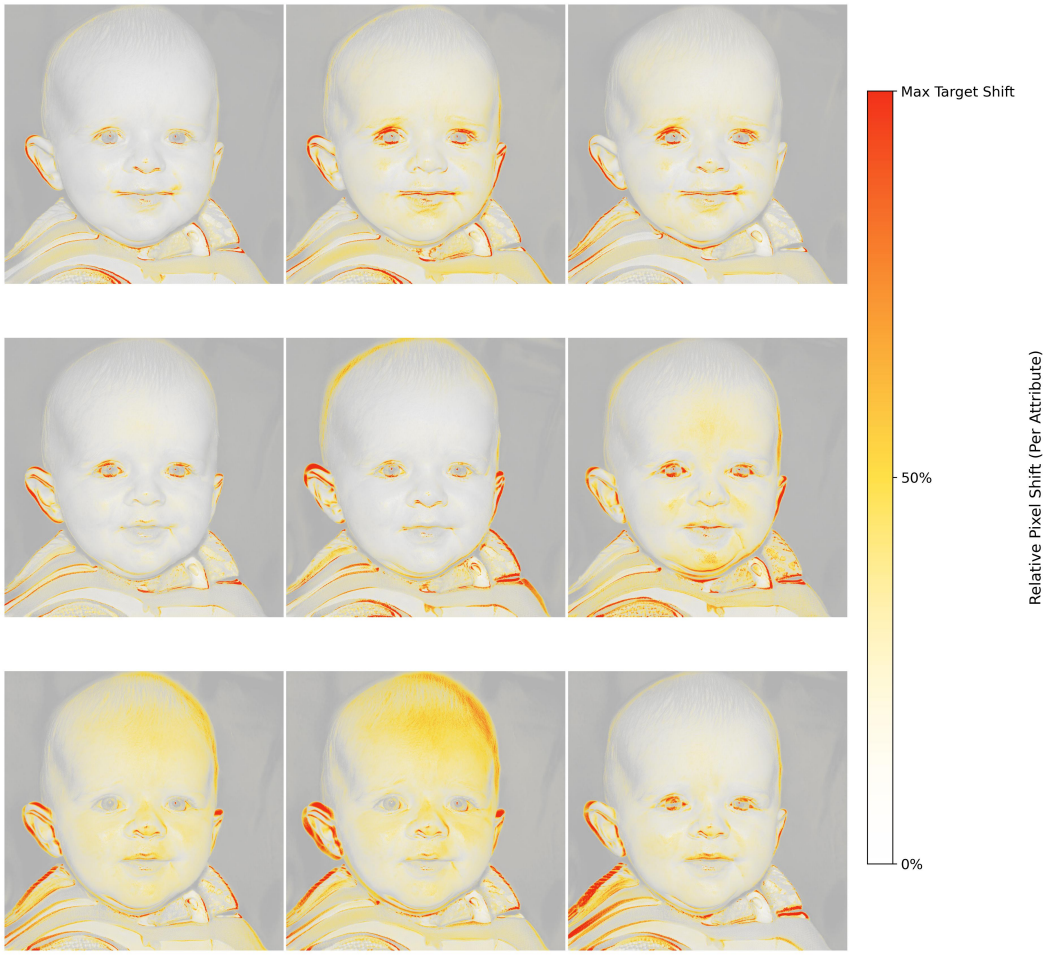

Figure A1: **Pixel attribution saliency maps for Figure 2.** Rows correspond to the target prompts (Smiling, Angry, Crying); columns show the relative per-attribute pixel shift induced by *Wachter et al.*, *RSGD*, and *Ours* respectively, as a percentage of the maximum observed shift (colorbar). Higher values (red) indicate regions the optimizer perturbed most to reach the target decision boundary.

D Metrology Data Generation

Tabular data was generated using the finite element method (FEM) Maxwell solver JCMsuite (JCMwave). The forward model represents an optical scatterometry experiment in which monochromatic ultraviolet light with wavelength $\lambda = 266\,\mathrm{nm}$ is reflected from a periodic silicon line grating with a silicon-dioxide layer. For a geometry vector

$$\mathbf{p} = (cd, t, h, swa, r_{\mathrm{top}}, r_{\mathrm{bot}})^{\mathsf{T}},$$

the forward model returns a 168-dimensional vector of simulated intensities for the reflected zeroth diffraction order,

$$\mathbf{f} : \mathcal{X} \subset \mathbb{R}^6 \to \mathbb{R}^{168}, \qquad \mathbf{y} = \mathbf{f}(\mathbf{p}).$$

The parameters cd, t, h, r_{top}, and r_{bot} are geometric lengths in nanometers. They denote the critical dimension, silicon-dioxide thickness, grating height, top radius, and bottom radius, respectively. The parameter swa is the sidewall angle in degrees. The grating period is fixed at $50\,\mathrm{nm}$.

The 168 channels combine 42 incidence angles, $\theta \in \{5^\circ, 7^\circ, \ldots, 87^\circ\}$, two sample rotations, $\phi \in \{0^\circ, 90^\circ\}$, and S and P polarization.

The forward model was used by Wurm et al. to reconstruct grating parameters Wurm et al. (2011). Subsequent studies used it for Gaussian-process parameter reconstruction Schneider et al. (2019), Bayesian inversion Hammerschmidt et al. (2017), and polynomial-chaos global sensitivity analysis Farchmin et al. (2019). Several JCMwave blog posts also describe the model in detail Schneider (2021b;c;a).

We generated candidate points from a Sobol sequence in the six-dimensional parameter domain. The nominal bounds were $cd \in [15, 35]\,\mathrm{nm}$, $t \in [2.5, 6.5]\,\mathrm{nm}$, $h \in [40, 60]\,\mathrm{nm}$, $swa \in [84, 90]\,^\circ$, $r_{\mathrm{top}} \in [3, 18]\,\mathrm{nm}$, and $r_{\mathrm{bot}} \in [0, 12]\,\mathrm{nm}$. Each candidate was tested against geometric constraints that prevent invalid line-grating profiles. For example, at a sidewall angle of 90°, the top radius cannot exceed $cd/2$. Candidates that violated a constraint were rejected. Sampling continued until 4096 valid samples were accepted. For each sample, the scattering problem was evaluated for every (θ, ϕ) pair and both polarization states. The resulting data set contains one geometry vector and one 168-dimensional simulated intensity vector per sample.