COUNTERFACTUAL EXPLANATIONS ON ROBUST PERCEPTUAL GEODESICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent-space optimization methods for counterfactual explanations—framed as minimal semantic perturbations that change model predictions—inherit the ambiguity of Wachter et al.'s objective: the choice of distance metric dictates whether perturbations are meaningful or adversarial. Existing approaches adopt flat or misaligned geometries, leading to off-manifold artifacts, semantic drift, or adversarial collapse. We introduce Perceptual Counterfactual Geodesics (PCG), a method that constructs counterfactuals by tracing geodesics under a perceptually Riemannian metric induced from robust vision features. This geometry aligns with human perception and penalizes brittle directions, enabling smooth, on-manifold, semantically valid transitions. Experiments on three vision datasets show that PCG outperforms baselines and reveals failure modes hidden under standard metrics.

1 INTRODUCTION

As deep learning models grow in scale and impact, interpretability becomes paramount as it offers a crucial lens into their internal reasoning. Traditional saliency-based methods, which highlight influential input features (Simonyan et al., 2014; Sundararajan et al., 2017; Smilkov et al., 2017; Kapishnikov et al., 2021; Ribeiro et al., 2016; Selvaraju et al., 2016; Lundberg & Lee, 2017), have been widely adopted for vision models but produce static, often noisy attributions that lack guidance on how predictions could be altered. **Counterfactual explanation (CE)** methods have emerged as a complementary paradigm grounded in the fundamental human capacity to contemplate *"what if?"* scenarios (Wachter et al., 2017; Ustun et al., 2019; Joshi et al., 2019; Artelt & Hammer, 2019). Rather than merely highlighting salient regions, CEs specify which semantic features should be modified—and how—to produce a different prediction. Wachter et al. (Wachter et al., 2017) formalized this notion as a solution to an optimization problem:

$$\min_{x} \quad \underbrace{r(x^\star, x)}_{\text{Similarity Distance}} + \lambda \ \underbrace{\ell(f(x), y')}_{\text{Classification Loss}} , \tag{1}$$

where $x^\star$ is the original input, y' the desired class, f the classifier, ℓ a loss function (e.g., cross-entropy), r a distance metric, and λ a hyperparameter balancing classification and similarity.

Considerable debate has emerged around whether a CE is fundamentally distinct from an adversarial example (AE), as both arise from the same optimization problem (Wachter et al., 2017; Browne & Swift, 2020; Pawelczyk et al., 2022; Freiesleben, 2022). The choice of distance metric r plays a central role: while it may support meaningful CEs, it can also encourage AEs if it favors imperceptibly small, distributed perturbations. Wachter et al. Wachter et al. (2017) acknowledged this ambiguity, noting that "AEs are counterfactuals by another name," proposing distinction on two grounds: (i) a misalignment of the distance metric with meaningful feature changes—since metrics typically used for AEs favor such dispersed modifications, thereby diminishing their explanatory value, and (ii) adversarial perturbations are non-semantic signals that displace inputs out of the possible world—i.e., off-manifold regions that do not correspond to valid examples under the data distribution.

Rather than directly solving eq. (1), some approaches leverage generative models to produce visual CEs by exploiting low-dimensional semantic representations (Augustin et al., 2022; Mertes et al., 2022; Looveren et al., 2021; Singla et al., 2020; Lang et al., 2021; Khorram & Fuxin, 2022). For instance, Singla et al. (2020) trained a conditional GAN to produce exaggerated CEs, while Lang et al. (2021) used a conditional STYLEGAN2-based approach to generate sparse visual CEs along

disentangled classifier-relevant style-space directions. Khorram & Fuxin (2022) used cycle-consistent losses to train transformations between factual and counterfactual distributions in generative latent spaces. Though visually compelling, these methods rely on exhaustive techniques that depart from the direct optimization formulation and ignore the geometry of the data manifold.

Other research adopt eq. (1) in the latent space of generative models (Joshi et al., 2019; Duong et al., 2023; Dombrowski et al., 2024; Pegios et al., 2024), but either assume flat Euclidean geometry (Joshi et al., 2019; Dombrowski et al., 2024), failing to capture the manifold's intrinsic curvature, or use geometrically informed yet adversarially vulnerable distance metrics (Pegios et al., 2024). For example, REVISE (Joshi et al., 2019) solves the objective in eq. (1) in a VAE latent space under Euclidean assumptions, using explicit ℓ_1/ℓ_2 distance terms. Dombrowski et al. (2024) discard explicit similarity terms and employ Stochastic Gradient Descent (SGD) assuming flat geometry misaligned with the underlying data manifold.

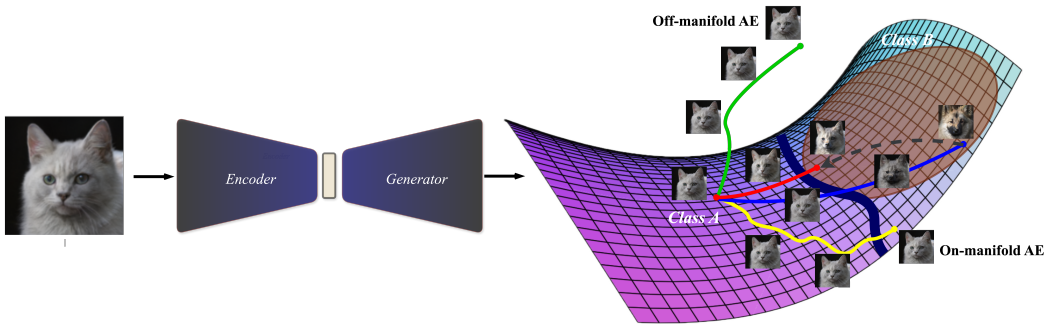

Figure 1: **Schematic of PCG.** An input is mapped through an encoder-generator pair. A linear latent path to a perceptually plausible target-class sample (Class B, brown region) is refined in Phase 1 into the blue geodesic by minimizing robust perceptual energy. In Phase 2, the endpoint and intermediate points are jointly optimized under classification loss and robust energy, resulting in the red counterfactual geodesic. The green trajectory (REVISE, VSGD) ignores manifold geometry, strays off-manifold and produces off-manifold AEs. The yellow trajectory (RSGD/-C) conforms to a fragile geometry, getting stuck in on-manifold adversarial regions (Class B, outside brown region).

This misalignment often causes perturbations to stray off-manifold, leading to implausible or off-manifold AEs. Pegios et al. (2024) proposed equipping the latent space with a Riemannian metric induced by the generator and optimizing with Riemannian SGD (RSGD) to account for the geometry of the data manifold. However, their induced metric is typically derived by pulling back either the pixel-space ℓ_2 or a standard classifier's feature space metric. Both are problematic in the vision domain: the ℓ_2 norm is a poor proxy for human perception (Sinha & Russell, 2011; Jordan et al., 2019; Rybkin, 2022), while a standard feature-based metric is semantically brittle as it inherits the adversarial vulnerabilities of non-robust deep vision models (Sjögren et al., 2022; Ghazanfari et al., 2024).

Such methods, while mostly proposed for tabular data settings, acutely fail in the high-dimensional vision domain, where the counterfactual optimization process can't distinguish between CEs and AEs. Browne & Swift (2020) proposed the notion of a *semantic divide*—a distinction between perturbations that affect human-understandable semantic features or low-level, uninterpretable features. Perturbations with rich semantic content fall on the explanatory side; pixel-level or low-level ones fall on the adversarial side. Browne & Swift (2020) argue that neither distance metrics nor appeals to "possible worlds" fully resolve this distinction; instead, semantic relevance only determines whether a result is a valid CE or an AE.

We agree with Browne & Swift (2020) that the second criterion proposed by Wachter et al.–displacement to off-manifold regions–fails to adequately differentiate AEs from CEs. Several studies have shown that on-manifold AEs exist (Ilyas et al., 2019; Garcia et al., 2023; Song et al., 2018), and can be generated via generative models (Stutz et al., 2019; Zhao et al., 2018), representing a subclass of AEs that reside within Wachter et al.'s "possible worlds". However, we challenge the assertion that distance metrics are inherently incapable of making the distinction. We show that if the

data manifold is endowed with a semantically robust Riemannian metric, solving the counterfactual optimization–when guided appropriately–can cross the semantic divide and produce valid CEs.

Failure Modes of Previous Approaches. We attribute the failure of previous latent-space counterfactual optimization methods in the high-dimensional image data regime to three core limitations:

- **(i) Off-manifold Traversal.** Optimization in latent space often disregards the geometry of the data manifold, leading to off-manifold AEs or semantically implausible counterfactuals (Pegios et al., 2024).
- **(ii) Local Gradient Optimization.** Without global structural guidance, single-point geometry-aware gradient methods operate locally and overlook the global manifold structure, including the existence of on-manifold adversarial regions. As a result, they often converge to either semantically distant counterfactuals or on-manifold AEs.
- **(iii) Versatility of Generators.** Even when accounting for manifold geometry, high-capacity generators can exploit non-robust or misaligned distance metrics to produce on-manifold AEs (Stutz et al., 2019; Zhao et al., 2018; Gilmer et al., 2018), fooling the metric rather than producing semantically meaningful perturbations that genuinely cross the semantic barrier.

Contributions. Motivated by findings in adversarial robustness that show robust models exhibit perceptually aligned gradients (Ganz et al., 2023; Srinivas et al.; Shah et al., 2021; Kaur et al., 2019), robust saliency maps (Etmann et al., 2019; Zhang & Zhu, 2019; Tsipras et al., 2019), and meaningful CEs (Boreiko et al., 2022; Santurkar et al., 2019; Augustin et al., 2020), we introduce a semantically grounded, data-manifold-based approach for perceptually progressive CEs. We emphasize that our focus lies not in interpreting robust classifiers themselves, but in generating explanations for standard models, positioning our work orthogonally to efforts aimed at explaining robust models (Boreiko et al., 2022; Santurkar et al., 2019; Augustin et al., 2020). Our key contributions are as follows:

- **(i) Counterfactual Generation:** We introduce **Perceptual Counterfactual Geodesics (PCG)**, which leverages a robust Riemannian metric on the latent space of a StyleGAN2/3 generator (Karras et al., 2020b; 2021). This metric is induced from feature spaces of robust vision models. PCG optimizes counterfactual trajectories along geodesic paths, ensuring that counterfactual evolution adheres to robust perceptual perturbations that cross the semantic barrier, avoiding off- or on-manifold adversarial regions.
- **(ii) Perceptual Geodesic Interpolation:** We show that the robust latent geometry underlying PCG enables smooth and semantically robust interpolations between samples. Our experiments demonstrate that trajectories aligned with the robust Riemannian metric preserve class coherence and perceptual structure. In contrast, other metrics collapse into visually ambiguous or brittle transitions due to geometric misalignment.

2 Background

2.1 Differential Geometry of Deep Generative Models

Deep generative models, such as VAEs and GANs, offer a powerful framework for learning high-dimensional data distributions through low-dimensional latent representations (Kingma & Welling, 2022; Higgins et al., 2016; Goodfellow et al., 2014; Karras et al., 2018). These models define a generative function $g : Z \to X$, where $Z \subset \mathbb{R}^d$ is a latent space and $X \subset \mathbb{R}^D$ is a high-dimensional data space, typically $d \ll D$. The image of Z under g, denoted $\mathcal{M} = g(Z) \subset X$, forms a subset of the data space, often referred to as the *data manifold*. Under mild regularity conditions—such as smoothness of g with a full-rank Jacobian mapping $J_g \triangleq \partial g/\partial z : Z \to \mathbb{R}^{D\times d}$—this image is a smooth, d-dimensional immersed submanifold of X (Shao et al., 2017; Arvanitidis et al., 2017). This construction supports the manifold hypothesis, which posits that real-world high-dimensional data concentrates near such a low-dimensional manifold (Brahma et al., 2016; Fefferman et al., 2013; Tenenbaum et al., 2000).

However, while Z is typically treated as Euclidean, this assumption misaligns with the geometry induced by g, as the nonlinear generator significantly distorts its structure. As a result, distances and directions in Z do not reflect the true relationships of the data manifold. This motivates equipping the latent space with a geometry that faithfully reflects the structure of the image manifold $\mathcal{M}$.

2.2 Pullback Metrics and the Geometry of Generators

A smooth manifold $\mathcal{M} \subset X$ inherits a tangent space $T_x\mathcal{M}$ at each point $x \in \mathcal{M}$, consisting of directions along which one can move locally. To measure lengths and angles, we define a smoothly varying inner product $\langle \cdot, \cdot \rangle_x$ on each tangent space. This defines a Riemannian metric $G(x)$, and the pair $(\mathcal{M}, G)$ forms a Riemannian manifold.

Given a smooth generator $g : Z \to X$, we equip the latent space Z with a Riemannian metric via pullback from the ambient space X, assumed to have a metric $G_X(x) \in \mathbb{R}^{D\times D}$. For any $u, v \in T_z Z \cong \mathbb{R}^d$, we define:

$$\langle u, v \rangle_z := \langle J_g(z)u, J_g(z)v \rangle_{G_X(g(z))} = u^\top J_g(z)^\top G_X(g(z)) J_g(z) v,$$

where $J_g(z)$ is the Jacobian of g at z. If $J_g(z)$ has full column rank, this defines the pullback metric as $G_Z(z) = J_g(z)^\top G_X(g(z)) J_g(z)$.

While mathematically well-defined, this construction inherits the limitations of the ambient metric. When $G_X(x) = I$, the geometry is induced from the canonical pixel-wise ℓ_2 metric. In high-dimensional vision tasks, such distances misalign with human perception and are highly sensitive to small, imperceptible perturbations. This issue is not limited to Euclidean metrics; it also applies to other ambient geometries that lack robust semantic grounding. For example, Pegios et al. (2024) pulls back a feature-based metric from a standard classifier, which operates in feature space but still inherits the adversarial vulnerabilities of non-robust models. As a result, the induced latent geometry reflects local structure relative to a brittle and semantically misaligned notion of similarity, often leading to adversarial trajectories (Browne & Swift, 2020).

2.3 Latent Space Counterfactual Optimization

We summarize several methods that solve variations of eq. (1) in the latent space of generative models.

REVISE. Joshi et al. (2019) introduced an approach based on VAEs for tabular data, where the latent code z of an input $x^\star$ is updated via SGD on the objective $\mathcal{L} = d(x^\star, g(z)) + \lambda\, \ell(f(g(z)), y')$. This method relies on two assumptions: that pixel-wise Euclidean distances in ambient space provide meaningful similarity, and that Euclidean SGD updates in latent space correspond to smooth semantic transitions. Both assumptions fail in high-dimensional vision domains, where distances are misaligned with perception and SGD updates stray off-manifold.

Vanilla SGD (VSGD). To adapt to vision settings, Dombrowski et al. (2024) proposed eliminating the distance term in REVISE and directly applying vanilla SGD to the classification loss:

$$z \;\leftarrow\; z \;-\; \eta\, \nabla_z \Big[\, \ell\big(f(g(z)),\, y'\big) \Big].$$

While sidestepping metric misalignment in X, it still assumes a flat Euclidean geometry in Z, ignoring the curvature induced by g. Since g is highly nonlinear in expressive models, such updates often stray off the manifold and lead to off-manifold AEs or perceptually implausible counterfactuals.

Riemannian SGD (RSGD). Pegios et al. (2024) proposed RSGD to account for the curvature of the data manifold by replacing Euclidean gradients with Riemannian ones derived from a pullback metric on the latent space. Given a stochastic VAE generator $g_\varepsilon(z) = \mu(z) + \sigma(z) \odot \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, I)$, the latent metric is defined as the expected pullback of the ambient ℓ_2 metric:

$$\hat{G}_Z(z) \approx J_\mu(z)^\top J_\mu(z) + J_\sigma(z)^\top J_\sigma(z),$$

and optimization proceeds via: $\quad z \leftarrow z - \eta\, \dfrac{r}{\|r\|_2}, \quad$ where $\quad r = \hat{G}_Z(z)^{-1} \nabla_z \ell(f(\mathbb{E}_\varepsilon[g_\varepsilon(z)]), y')$.

A variant, RSGD-C, replaces the ambient metric with the pullback of a classifier-based feature metric, using the final-layer representation of a standard classifier. This introduces task-awareness by aligning updates with decision-relevant directions.

Both methods remain limited by their underlying metrics. Pixel-wise ℓ_2 distances are fragile and misaligned with perception, and standard classifier-based features inherit adversarial vulnerabilities. RSGD/-C does not enforce geodesic paths and has been applied only in low-dimensional domains where adversariality is less evident.

3 METHODOLOGY

Prior approaches fail in the vision domain due to three tightly coupled issues: the use of perceptually misaligned metrics (e.g., ℓ_2 in pixel space or fragile classifier-based metrics), reliance on local gradient updates that ignore global manifold structure, and the expressive power of high-capacity generators that exploit these misalignments to produce adversarial perturbations.

Our method, **PCG**, addresses these limitations by casting counterfactual generation as a global curvature-aware optimization over latent trajectories on a Riemannian manifold, where the generator induces a latent geometry aligned with human perception. To define this geometry, we construct a perceptually robust ambient metric. Unlike standard classifiers, robust models learn representations that are resistant to adversarial perturbations and aligned with human perceptual similarity. These robust intermediate activation spaces exhibit linearly separable structure and encode grounded, semantically meaningful features. As a result, the Euclidean metric becomes a more reliable proxy for perceptual similarity in these robust semantic spaces, unlike its failure in pixel or fragile semantic spaces. We leverage this structure to define a composite ambient metric by aggregating pullbacks of the Euclidean metric from robust feature spaces into the input space, capturing hierarchical, perceptually coherent variations. Formally, we define the robust perceptual metric as:

$$G_R(x) = \sum_{k=1}^{K} w_k \, J_{h_k}(x)^\top J_{h_k}(x), \quad w_k = \frac{1}{N_k},$$

where K is the number of selected intermediate layers of a pretrained robust vision model, $h_k(x)$ denotes the activation of the k-th layer with dimensionality $d_k \gg D$, $J_{h_k}(x) \in \mathbb{R}^{d_k \times D}$ is its Jacobian with respect to the input $x \in \mathbb{R}^D$, and N_k denotes the total size (number of elements) of the activation $h_k(x)$, which normalizes each layer so that no single feature space dominates due to its size. Pulling back G_R through the generator $g : Z \to X$ defines the latent-space metric

$$G_Z(z) = J_g(z)^\top G_R(g(z)) J_g(z),$$

which induces a latent geometry that penalizes brittle or non-robust directions and favors perturbations that produce perceptually smooth, semantically aligned variations in the image space.

We seek a smooth latent trajectory $\gamma : [0, 1] \to Z$ such that $g(\gamma(t))$ evolves through robust semantic regions. The perceptual length of this trajectory, where $\gamma'(t) = d\gamma/dt$ is the latent-space velocity, evaluated under G_R, is

$$L(g(\gamma)) = \int_0^1 \sqrt{\gamma'(t)^\top G_Z(\gamma(t))\gamma'(t)}dt,$$

and minimizing this length under constant-speed parametrization is equivalent to minimizing the robust perceptual energy (Jost, 2017):

$$E(g(\gamma)) = \frac{1}{2} \int_0^1 \gamma'(t)^\top G_Z(\gamma(t))\gamma'(t)dt. \tag{2}$$

Expanding G_Z using the composite metric shows that the pullback energy is a weighted sum of squared velocities in each robust feature space:

$$\gamma'(t)^\top G_Z(\gamma(t))\gamma'(t) = \sum_{k=1}^{K} w_k \left\| \frac{d}{dt} h_k\big(g(\gamma(t))\big) \right\|_2^2.$$

Minimizing $E(g(\gamma))$ thus amounts to finding a geodesic whose generator outputs move smoothly and consistently across all robust semantic layers. To do this, we discretize γ into $T + 1$ points $\{z_0, \ldots, z_T\}$, where z_0 is the latent encoding of the input $x^\star$, and z_T is initialized as the latent encoding of an arbitrary target-class sample from the dataset. This initialization is critical: unlike previous methods that perform iterative updates from a single starting point—which often converge to on-manifold adversarial endpoints—we initialize between two manifold-conforming points to guide global transitions across semantically valid regions under the robust metric. Using forward finite differences as in Shao et al. (2017), we approximate the robust feature-space velocity at t_i

as $dh_k(g(\gamma(t)))/dt \mid_{t=t_i} \approx (h_k(g(z_{i+1})) - h_k(g(z_i)))/\delta t$. This gives the discrete robust energy equivalent of eq. (2):

$$E_{\text{robust}}(\mathbf{z}) = \frac{1}{2} \sum_{i=0}^{T-1} \sum_{k=1}^{K} \frac{w_k}{\delta t} \left\| h_k\big(g(z_{i+1})\big) - h_k\big(g(z_i)\big) \right\|_2^2, \quad \text{where } \mathbf{z} \triangleq [z_0, \ldots, z_T] \text{ and } \delta t = 1/T.$$

Optimization proceeds in two stages. In Phase 1, we fix z_0 and z_T and minimize $E_{\text{robust}}(\mathbf{z})$ with respect to the intermediate points to obtain a geodesic consistent with the robust semantic geometry induced by the generator. In Phase 2, we release z_T and jointly optimize the energy and a classification loss to ensure the endpoint maintains the desired prediction under f. The combined loss is

$$\mathcal{L}(\mathbf{z}) = E_{\text{robust}}(\mathbf{z}) + \lambda \cdot \ell(f(g(z_T)), y'),$$

Minimizing the combined loss traces a robust counterfactual path: E_{robust}, built on adversarially robust features, supplies perceptually aligned, manifold-conforming gradients (Ganz et al., 2023; Zhang & Zhu, 2019; Tsipras et al., 2019; Ilyas et al., 2019; Stutz et al., 2019) that guide on-manifold updates and temper the endpoint loss, keeping the trajectory on a robust geodesic. The overall structure of our two-stage optimization and the contrast with prior methods is illustrated in Figure 1; full algorithm, and optimization details are provided in Appendix A.1.

4 Experiments

We evaluate PCG against prior latent-space optimization methods. In section 4.1, we first show the failure mode of interpolation methods inherent in their geometrical assumptions, and demonstrate the effect of our proposed robust Riemannian metric in generating perceptually smooth geodesics that underpins PCG. In section 4.2, we compare PCG with other approaches in terms of the perceptual plausibility of the generated counterfactuals. Finally, we quantitatively evaluate PCG under both typical and geometry-aware distance measures. Code for our experiments is available here.

Datasets. We evaluate our method on three high-dimensional real-image datasets: (1) AFHQ (Choi et al., 2020), with high-resolution images of cats, dogs, and wild animals; (2) FFHQ (Karras et al., 2019), containing 70,000 diverse human face images; and (3) PlantVillage (Hughes & Salathé, 2015), with labeled images of healthy and diseased plant leaves across species.

Models. We train StyleGAN2 generators from scratch on AFHQ and PlantVillage (≈140 NVIDIA H100 GPU-hours per model) (Karras et al., 2020a). For AFHQ, we also use a pretrained StyleGAN3 generator (Karras et al., 2021). For FFHQ, we use pretrained StyleGAN2 and StyleGAN3. Post hoc, we train image-to-latent encoders (used for all counterfactual optimization in z-space) and then briefly fine-tune the encoder–generator pair jointly. For classifiers, we train binary models based on the VGG-19 backbone (Simonyan & Zisserman, 2014): one per AFHQ class pair and a healthy–vs–unhealthy classifier for PlantVillage. Because FFHQ lacks labels, we train attribute classifiers on CelebA (Liu et al., 2015) and apply them to FFHQ. Architectural and training details appear in Appendix A.3.

Baselines. We compare PCG against the following latent-space based approaches:

- **REVISE** (Joshi et al., 2019). Latent-space equivalent of Wachter et al.'s objective based on SGD.
- **VSGD** (Dombrowski et al., 2024). It performs distance-free vanilla SGD in the latent space.
- **RSGD/-C** (Pegios et al., 2024). In these variants, a Riemannian metric is used to guide SGD. The metrics are pull-back from either the Euclidean metric in the ambient space or in the final layer of the classifier under explanation.

4.1 Effect of Latent Geometry on Interpolation

In Figure 2, we illustrate how latent-space geometry shapes interpolation. The top row linearly interpolates in latent space Z under a Euclidean assumption, which ignores the nonlinear distortion induced by the generator and produces mid-path off-manifold artifacts such as class ambiguity, unnatural warping, and deformed textures. The second row minimizes pixel-space MSE in X, which induces a latent-space geometry by pulling back the Euclidean metric from X to Z; transitions remain brittle and semantically incoherent, with midway blends of disparate attributes that expose

the fragility and misalignment of pixel-wise distances. The third row uses the pullback of a feature metric from a standard ResNet-50 (He et al., 2016) (see appendix A.4); semantics improve, yet fading, illumination shifts, and class discontinuities persist. These instabilities reflect the vulnerability of non-robust models to adversarial perturbations and reliance on brittle features, with similar failure modes reported in Laine (2018) using VGG-19. In contrast, the fourth row applies our robust perceptual metric derived from a robust ResNet-50, producing smooth, on-manifold trajectories with consistent semantics and coherent evolution. This confirms our hypothesis that robust Riemannian geometry enables smooth, semantically valid on-manifold interpolations while avoiding adversarial collapse.

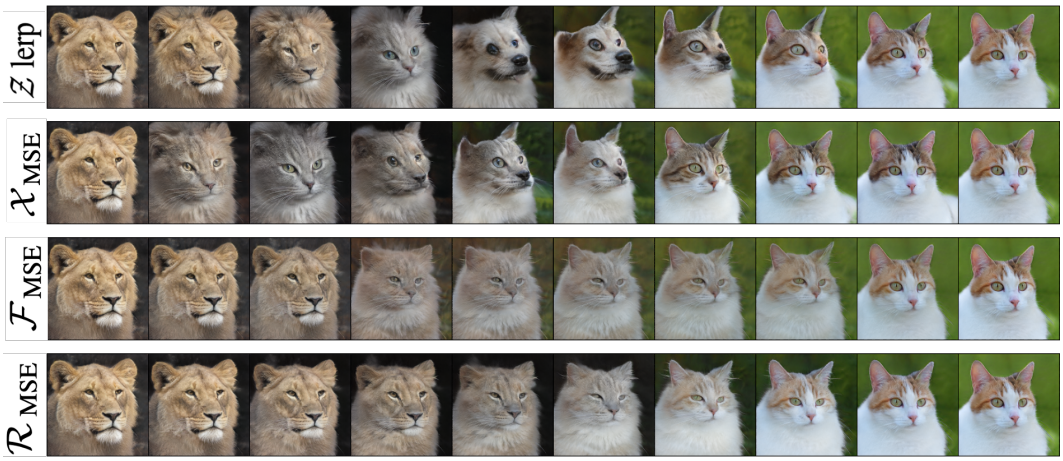

Figure 2: Interpolation paths under four latent geometries based on StyleGAN2 (top→bottom). (a) Z-linear (Euclidean): flat latent metric; off-manifold artifacts. (b) Pixel MSE pullback: Euclidean metric pulled back to Z; brittle, incoherent paths. (c) Standard feature pullback: non-robust ResNet-50; better semantics but still fading and discontinuities. (d) Robust perceptual pullback (ours): robust ResNet-50; smooth, consistent, on-manifold trajectories. See Appendix B.1 for StyleGAN3 results.

4.2 Perceptual Counterfactual Geodesics

Having established smooth perceptual geodesics under our proposed metric, we now demonstrate their refinement into plausible CEs. Figure 3 showcases the two-stage nature of our approach. In Phase 1 (rows 1 and 3), we generate an initial perceptual geodesic between the input and an arbitrary target-class sample, such as a dog image for a cat input, or a non-blonde face for a blonde input. Although the target is semantically distant, the path remains coherent, illustrating the alignment of our metric with perceptual structure. In Phase 2 (rows 2 and 4), we release the endpoint and jointly optimize it with the path under the classification loss, allowing the counterfactual to move closer to the input while maintaining geodesicity. The resulting counterfactual geodesics trace robust regions of the data manifold and maintain consistent semantics throughout the trajectory, retaining the semantic continuity and avoiding adversarial shortcuts or abrupt transitions. This step ensures the whole path travels through perceptually robust regions on the manifold as shown in Fig 1. We show that different choices of the target-class exemplar lead optimization to converge within a small neighborhood of the input, producing diverse yet faithful counterfactual explanations; see Appendix B.3

Comparison with Baselines. We now evaluate the final counterfactuals produced by PCG against existing latent-space optimization methods. As shown in Figure 4, our method consistently produces semantically valid CEs that remain close to the input while effecting the desired class transition. In contrast, RSGD and RSGD-C, despite accounting for local curvature, rely on fragile metrics (e.g., pixel-space ℓ_2 or non-robust classifier features) that remain vulnerable to adversarial manipulation. Many of the generated counterfactuals collapse into on-manifold AEs—as seen in rows 1, 2, 4, 5, and 6. Like the outputs of other baselines, they fall on the adversarial side of the semantic divide. Even when RSGD variants converge (e.g., row 3), the output is visibly distant from the input in pose and structure, reflecting the lack of geodesic constraint and a tendency to traverse longer manifold paths. VSGD, which assumes flat Euclidean geometry, produces off-manifold perturbations that are either perceptually implausible, or adversarial. In row 2, the generated counterfactual exhibits class

ambiguity and disoriented eye alignment; in row 3, the face is unnaturally elongated with distortions under the chin; in row 6, the leaf counterfactual contains an unnatural cusp-like protrusion that breaks the expected symmetry, fullness, and surface continuity of leaves. These artifacts arise from ignoring the data manifold altogether. REVISE exhibits similar failure modes: the strong pixel-wise distance penalty constrains outputs to remain close in ℓ_2 norm, but adversarial. All REVISE outputs in the figure represent off-manifold AEs, driven by the optimization pressure to minimize distance rather than induce meaningful semantic change. In contrast, PCG navigates robust regions of the manifold along perceptual geodesics, producing minimal, semantically faithful changes.

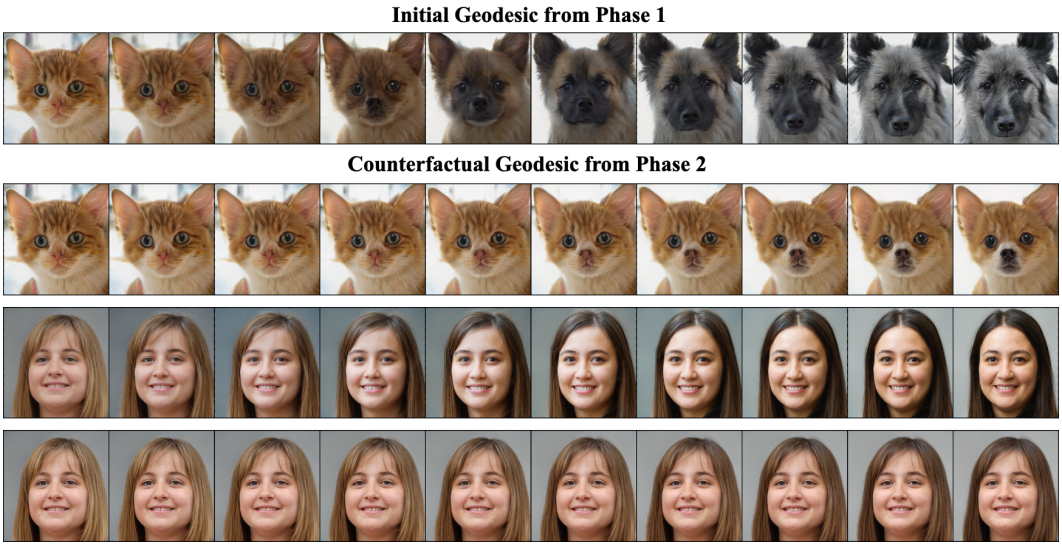

Figure 3: Perceptual Counterfactual Geodesics. Rows 1 and 3: initial geodesics from Phase 1 between an input and a target-class sample. Rows 2 and 4: counterfactual geodesics after Phase 2, where the endpoint is optimized with the path. Trajectories from Phase 2 stay in robust regions of the manifold and preserve semantic continuity. Results from StyleGAN2 (see Appendix B.2 for StyleGAN3)

Table 1: Quantitative comparison across datasets for StyleGAN2 (see Appendix B.4 for StyleGAN3 and Appendix B.5 for runtime complexity). Columns report $\mathcal{L}_1$ (pixel ℓ_1), $\mathcal{L}_2$ (pixel ℓ_2), $\mathcal{L}_\mathcal{F}$ (pullback from standard VGG-16), and $\mathcal{L}_\mathcal{R}$ (pullback from robust Inception-V3). Lower is better.

Method	AFHQ				FFHQ				PlantVillage			
	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_\mathcal{F}$	$\mathcal{L}_\mathcal{R}$	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_\mathcal{F}$	$\mathcal{L}_\mathcal{R}$	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_\mathcal{F}$	$\mathcal{L}_\mathcal{R}$
REVISE	1.20±0.12	**0.73**±0.18	1.08±0.10	2.70±0.05	0.82±0.08	**0.32**±0.13	0.82±0.08	2.78±0.06	0.50±0.13	**0.38**±0.15	0.96±0.06	2.87±0.07
VSGD	1.31±0.11	1.49±0.15	1.60±0.09	2.90±0.08	0.79±0.11	0.96±0.10	1.50±0.12	2.86±0.07	0.83±0.13	0.94±0.17	1.18±0.07	3.01±0.09
RSGD	0.85±0.08	1.32±0.09	0.70±0.07	1.85±0.05	0.61±0.05	0.84±0.07	0.61±0.04	2.41±0.05	0.78±0.08	0.82±0.11	0.54±0.05	2.28±0.04
RSGD-C	0.93±0.10	1.45±0.17	0.65±0.08	1.75±0.06	0.68±0.06	0.93±0.09	0.48±0.04	2.11±0.04	0.80±0.10	0.86±0.13	0.45±0.05	2.03±0.06
PCG (ours)	**0.79**±0.07	1.14±0.10	**0.53**±0.06	**0.31**±0.02	**0.42**±0.03	0.72±0.09	**0.39**±0.05	**0.22**±0.06	**0.36**±0.03	0.56±0.05	**0.34**±0.04	**0.20**±0.05

Quantitative Evaluation. We assess counterfactual proximity using four distance metrics: $\mathcal{L}_1$ (pixel-wise ℓ_1), $\mathcal{L}_2$ (pixel-wise ℓ_2), $\mathcal{L}_\mathcal{F}$ (distance induced by the pullback from standard ResNet-50 features), and $\mathcal{L}_\mathcal{R}$ (pullback from robust ResNet-50 features). Each induced metric is computed between the input and the final counterfactual in image space using the local quadratic form $\mathcal{L}_G(z_0, z_T) = \sqrt{(g(z_T) - g(z_0))^\top G(g(z_0))(g(z_T) - g(z_0))}$, where $G \in \{G_\mathcal{F}, G_\mathcal{R}\}$ is the respective ambient metric. This approximates perceptual distance in the feature space around the input. To avoid entanglement between optimization and evaluation, we compute $\mathcal{L}_\mathcal{F}$ using an independent VGG-16 model that was never involved in training or counterfactual optimization, and we compute $\mathcal{L}_\mathcal{R}$ using a robustly trained Inception-V3 model (Alfarra et al., 2022) separate from the robust ResNet-50 that defines our metric. As shown in Table 1, our method achieves the lowest distances across all geometry-aware metrics and also under $\mathcal{L}_1$, indicating sparse, perceptually meaningful changes. The margin is largest under $\mathcal{L}_\mathcal{R}$, and extends to $\mathcal{L}_\mathcal{F}$, since our robust geodesics stay closer even under weaker perceptual proxies. REVISE and VSGD often stray off-manifold, producing AEs that appear close under $\mathcal{L}_2$ (unsurprisingly, as REVISE directly minimizes this metric) but deviate sharply in

all perceptual geometries. RSGD and RSGD-C operate under their metrics, but lack geodescity and remain vulnerable to on-manifold AEs—perturbations smooth under ℓ_2 and $\mathcal{L}_{\mathcal{F}}$ yet semantically fragile. These cases highlight that our proposed $\mathcal{L}_{\mathcal{R}}$ serves as a more faithful evaluation metric, exposing failure modes that remain hidden under non-robust distances. Low scores in $\mathcal{L}_1, \mathcal{L}_2$, or $\mathcal{L}_{\mathcal{F}}$ do not guarantee meaningful proximity and can coincide with adversarial behavior.

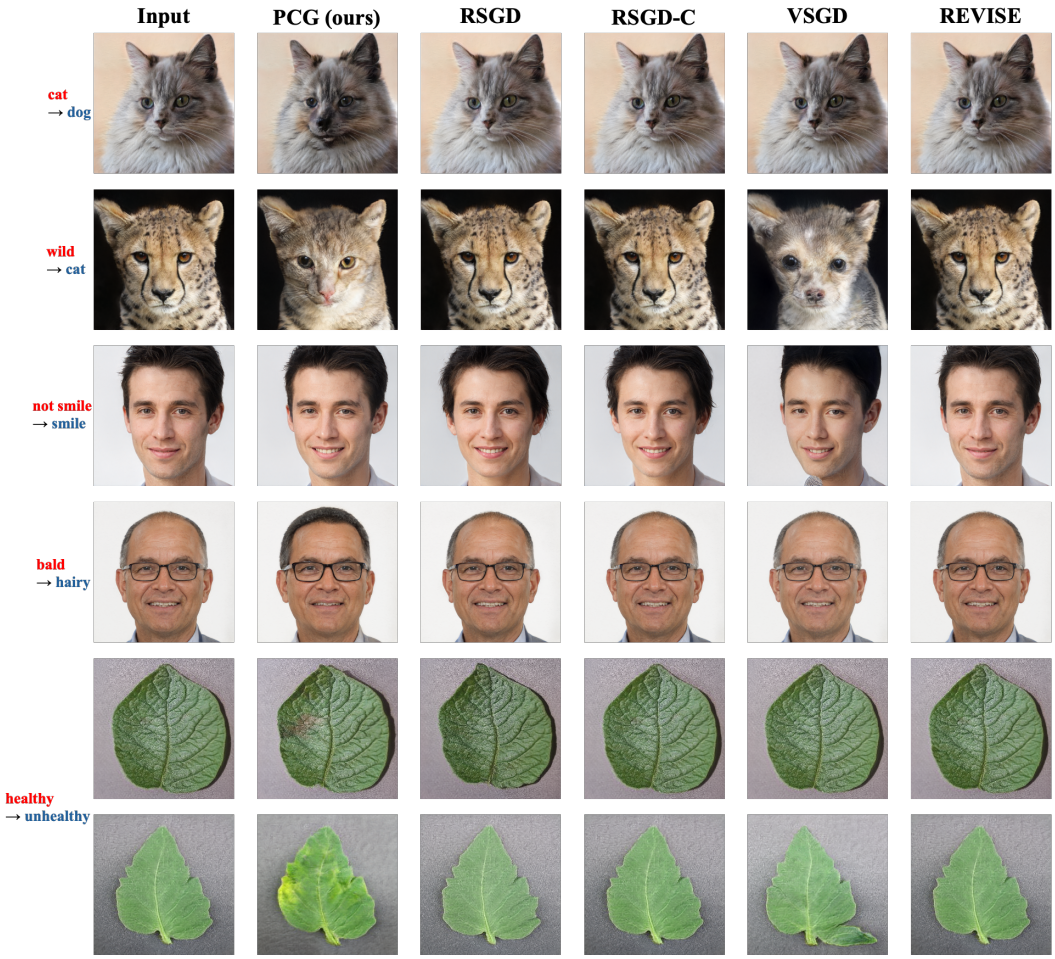

Figure 4: Qualitative comparison of counterfactuals across methods with StyleGAN2. Columns show input images followed by counterfactuals from PCG (ours), RSGD, RSGD-C, VSGD, and REVISE. Rows indicate input and target attribute/class. PCG produces minimal, semantically faithful changes along robust geodesics, while baselines often show off-manifold artifacts, semantic drift, or adversarial collapse. Optimization details for baselines are presented in Appendix A.2.

5 Conclusion

We introduced Perceptual Counterfactual Geodesics (PCG), a method for generating semantically faithful counterfactuals by optimizing smooth trajectories on a latent Riemannian manifold equipped with a robust perceptual metric. Our two-phase framework operationalizes established ideas from pullback geometry and robust perception into a practical algorithm. Empirically, PCG outperforms latent-space baselines and avoids their common failure modes (off- and on-manifold adversarial collapse, semantic drift). In addition, the robust geometry-aware evaluation $\mathcal{L}_{\mathcal{R}}$ exposes errors that remain hidden under standard distances, providing a more reliable yardstick for counterfactual quality. Conceptually, the contribution is algorithmic: we show that when the latent space is endowed with a robust, perceptually aligned geometry and optimized globally along paths, counterfactuals become smooth, diverse, and faithful. Limitations and future work are discussed in Appendix C

6 Ethics Statement

All authors have read and will adhere to the ICLR Code of Ethics. Our experiments use publicly available vision datasets under their licenses; no new human-subject data were collected, and we do not perform re-identification or demographic inference. Any released code is intended for research use and will include guidance discouraging harmful or deceptive applications. *LLM usage disclosure:* in line with ICLR policy, we used a large language model only for light copy-editing (grammar, typos, minor phrasing/formatting); it did not contribute to research ideation, analysis, or claims.

7 Reproducibility Statement

All methodological details, derivations, and hyperparameter settings required to reproduce our experiments are described in the main text (Section 3) and in Appendix A.1, where we also provide pseudocode for our two-stage optimization procedure. Architectural specifications, training protocols for generators, encoders, and classifiers, and additional results (including sensitivity to initialization) are included in Appendices A and B. Anonymized source code implementing PCG and all evaluation metrics is provided in Section 4 to enable full replication of our experiments.

References

Motasem Alfarra, Juan C. Pérez, Anna Frühstück, Philip H. S. Torr, Peter Wonka, and Bernard Ghanem. On the robustness of quality measures for gans, 2022. URL https://arxiv.org/abs/2201.13019.

André Artelt and Barbara Hammer. On the computation of counterfactual explanations – A survey, November 2019.

Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models. *arXiv: Machine Learning*, 2017.

Maximilian Augustin, Alexander Meinke, and Matthias Hein. Adversarial robustness on in- and out-distribution improves explainability, 2020. URL https://arxiv.org/abs/2003.09461.

Maximilian Augustin, Valentyn Boreiko, Francesco Croce, and Matthias Hein. Diffusion Visual Counterfactual Explanations, October 2022.

Valentyn Boreiko, Maximilian Augustin, Francesco Croce, Philipp Berens, and Matthias Hein. *Sparse Visual Counterfactual Explanations in Image Space*, pp. 133–148. Springer International Publishing, 2022. ISBN 9783031167881. doi: 10.1007/978-3-031-16788-1_9. URL http://dx.doi.org/10.1007/978-3-031-16788-1_9.

Pratik Prabhanjan Brahma, Dapeng Wu, and Yiyuan She. Why deep learning works: A manifold disentanglement perspective. *IEEE Transactions on Neural Networks and Learning Systems*, 27 (10):1997–2008, 2016. doi: 10.1109/TNNLS.2015.2496947.

Kieran Browne and Ben Swift. Semantics and explanation: Why counterfactual explanations produce adversarial examples in deep neural networks, December 2020.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8188–8197, 2020.

Ann-Kathrin Dombrowski, Jan E. Gerken, Klaus-Robert Müller, and Pan Kessel. Diffeomorphic Counterfactuals With Generative Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3257–3274, May 2024. ISSN 1939-3539. doi: 10.1109/TPAMI.2023.3339980.

Tri Dung Duong, Qian Li, and Guandong Xu. CeFlow: A Robust and Efficient Counterfactual Explanation Framework for Tabular Data using Normalizing Flows, March 2023.

Christian Etmann, Sebastian Lunz, Peter Maass, and Carola Schoenlieb. On the connection between adversarial robustness and saliency map interpretability. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1823–1832. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/etmann19a.html.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis, 2013. URL https://arxiv.org/abs/1310.0425.

Timo Freiesleben. The Intriguing Relation Between Counterfactual Explanations and Adversarial Examples. *Minds and Machines*, 32(1):77–109, March 2022. ISSN 0924-6495, 1572-8641. doi: 10.1007/s11023-021-09580-9.

Roy Ganz, Bahjat Kawar, and Michael Elad. Do perceptually aligned gradients imply robustness? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10628–10648. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/ganz23a.html.

Washington Garcia, Pin-Yu Chen, Hamilton Scott Clouse, Somesh Jha, and Kevin R.B. Butler. Less is more: Dimension reduction finds on-manifold adversarial examples in hard-label attacks. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 254–270, 2023. doi: 10.1109/SaTML54575.2023.00025.

Sara Ghazanfari, Alexandre Araujo, Prashanth Krishnamurthy, Farshad Khorrami, and Siddharth Garg. LipSim: A Provably Robust Perceptual Similarity Metric, March 2024.

Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S. Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres, 2018.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL https://arxiv.org/abs/1512.03385.

Magnus R. Hestenes and Eduard Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49(6):409–436, 1952. doi: 10.6028/jres.049.044.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. November 2016.

David P Hughes and Marcel Salathé. An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060*, 2015. URL https://arxiv.org/abs/1511.08060.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial Examples Are Not Bugs, They Are Features, August 2019.

Matt Jordan, Naren Manoj, Surbhi Goel, and Alexandros G. Dimakis. Quantifying Perceptual Distortion of Adversarial Examples, February 2019.

Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems, July 2019.

Jürgen Jost. *Riemannian geometry and geometric analysis*. Springer, 7 edition, 2017.

Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Ben Wedin, Michael Terry, and Tolga Bolukbasi. Guided Integrated Gradients: An Adaptive Path Method for Removing Noise, June 2021.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2018.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019. URL https://arxiv.org/abs/1812.04948.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proc. NeurIPS*, 2020a.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan, 2020b.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks, 2021. URL https://arxiv.org/abs/2106.12423.

Simran Kaur, Jeremy Cohen, and Zachary C. Lipton. Are perceptually-aligned gradients a general property of robust classifiers?, 2019. URL https://arxiv.org/abs/1910.08640.

Saeed Khorram and Li Fuxin. Cycle-consistent counterfactuals by latent transformations, 2022. URL https://arxiv.org/abs/2203.15064.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

Samuli Laine. FEATURE-BASED METRICS FOR EXPLORING THE LATENT SPACE OF GENERATIVE MODELS. 2018.

Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T. Freeman, Phillip Isola, Amir Globerson, Michal Irani, and Inbar Mosseri. Explaining in Style: Training a GAN to explain a classifier in StyleSpace, September 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Arnaud Van Looveren, Janis Klaise, Giovanni Vacanti, and Oliver Cobb. Conditional Generative Models for Counterfactual Explanations, January 2021.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017. URL https://arxiv.org/abs/1705.07874.

Silvan Mertes, Tobias Huber, Katharina Weitz, Alexander Heimerl, and Elisabeth André. GANterfactual—Counterfactual Explanations for Medical Non-experts Using Generative Adversarial Learning. *Frontiers in Artificial Intelligence*, 5, April 2022. ISSN 2624-8212. doi: 10.3389/frai.2022.825565.

Martin Pawelczyk, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. Exploring Counterfactual Explanations Through the Lens of Adversarial Examples: A Theoretical and Empirical Analysis. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pp. 4574–4594. PMLR, May 2022.

Paraskevas Pegios, Aasa Feragen, Andreas Abildtrup Hansen, and Georgios Arvanitidis. Counterfactual Explanations via Riemannian Latent Space Traversal, November 2024.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, San Francisco California USA, August 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939778.

Oleg Rybkin. The reasonable ineffectiveness of mse pixel loss for future prediction (and what to do about it), 2022.

Shibani Santurkar, Dimitris Tsipras, Brandon Tran, Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Image synthesis with a single (robust) classifier, 2019.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. 2016.

Harshay Shah, Prateek Jain, and Praneeth Netrapalli. Do input gradients highlight discriminative features?, 2021. URL https://arxiv.org/abs/2102.12781.

Hang Shao, Abhishek Kumar, and P. Thomas Fletcher. The riemannian geometry of deep generative models, 2017. URL https://arxiv.org/abs/1711.08014.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

Sumedha Singla, Brian Pollack, Junxiang Chen, and Kayhan Batmanghelich. Explanation by Progressive Exaggeration, February 2020.

Pawan Sinha and Richard Russell. A Perceptually Based Comparison of Image Similarity Metrics. *Perception*, 40(11):1269–1281, November 2011. ISSN 0301-0066, 1468-4233. doi: 10.1068/p7063.

Oskar Sjögren, Gustav Grund Pihlgren, Fredrik Sandin, and Marcus Liwicki. Identifying and Mitigating Flaws of Deep Perceptual Similarity Metrics, July 2022.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: Removing noise by adding noise, June 2017.

Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models, 2018. URL https://arxiv.org/abs/1805.07894.

Suraj Srinivas, Sebastian Bordt, and Himabindu Lakkaraju. Which Models have Perceptually-Aligned Gradients? An Explanation via Off-Manifold Robustness.

David Stutz, Matthias Hein, and Bernt Schiele. Disentangling adversarial robustness and generalization, 2019. URL https://arxiv.org/abs/1812.00740.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks, June 2017.

Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. doi: 10.1126/science.290.5500.2319.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy, 2019.

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, January 2019. doi: 10.1145/3287560.3287566.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal*, 2017. ISSN 1556-5068. doi: 10.2139/ssrn.3063289.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Tianyuan Zhang and Zhanxing Zhu. Interpreting adversarially trained convolutional neural networks, 2019. URL https://arxiv.org/abs/1905.09797.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples, 2018. URL https://arxiv.org/abs/1710.11342.

A FURTHUR DETAILS ON THE PCG ALGORITHM, BASELINES, MODELS, AND METRICS.

A.1 PCG OPTIMIZATION

Our objective minimizes the discrete robust perceptual energy of a latent trajectory under the pullback geometry. Because we differentiate the energy itself (squared feature increments along $h_k \circ g$, backprop through h_k and g automatically inserts the Jacobian factors that define the pullback metric. Two implications follow. First, in Phase 1 (energy-only), standard gradient descent already converges to a manifold-conforming geodesic for the path variables, so a Riemannian correction brings no additional benefit. Second, in Phase 2 we add a classification term that touches only the endpoint z_T; while one could Riemannian-correct that update in isolation, it is unnecessary in our coupled objective: the energy term continues to regularize all latent points (including z_T), steering the entire trajectory to remain a counterfactual geodesic.

PCG proceeds in two phases. The first constructs a smooth geodesic path between the input and a target-class sample, optimized for 200 steps with a fixed learning rate of $1\mathrm{e}{-3}$. The second refines the path into a faithful counterfactual over 300 steps, using the same learning rate and a dynamic λ schedule: starting from $1\mathrm{e}{-4}$ and multiplying by 5 every 50 steps. At each such interval, we apply a re-anchoring strategy: the path endpoint is reassigned to the closest point to the input along the trajectory that is classified as belonging to the target class. We then increase the resolution of the path by inserting midpoints between each pair of consecutive latent codes, restoring the original path length. Optimization resumes to refine the updated path, progressively giving closer counterfactuals. For completeness, the PCG optimization pseudocode is given in Algorithm 1.

A.2 BASELINES OPTIMIZATION

To ensure comparability, all baselines start from the same initialization $z_0 = e(x^\star)$, use the same encoder–generator pair, and are optimized for the same number of steps as PCG (200 + 300). We use Adam (Kingma & Ba, 2015) and select the step size by a small sweep $\eta \in \{1\mathrm{e}{-4},\ 3\mathrm{e}{-4},\ 1\mathrm{e}{-3}, 1\mathrm{e}{-2}\}$ on a held-out split; we report the best setting per method.

VSGD. Vanilla latent descent minimizes only the classification loss $\ell(f(g(z)), y')$ (no similarity term, no λ). We run Adam with the learning-rate sweep above.

REVISE. We optimize $d(x^\star, g(z)) + \lambda\, \ell(f(g(z)), y')$ in latent space. For fairness and due diligence, λ follows the same dynamic schedule used in PCG Phase 2 (start $1\mathrm{e}{-4}$, $\times 5$ every 50 steps). We use the same Adam sweep for η.

RSGD/-C. These variants require the inverse of the induced latent metric. We compute the natural-gradient direction by solving $G_Z(z)\, r = \nabla_z \mathcal{L}$ with Conjugate Gradients (Hestenes & Stiefel, 1952), using Jacobian–vector products via autodiff; this avoids explicit Jacobian assembly and matrix inversion. Since the original code targets VAEs on tabular data and is not public, we implement a deterministic metric compatible with our GAN setting (pixel ℓ_2 pullback for RSGD; classifier-feature pullback for RSGD-C) and apply the same Adam step-size sweep for the outer update.

A.3 GENERATORS, ENCODERS, AND CLASSIFIERS

Style-based generator (image prior). We use the official **StyleGAN2-ADA** (and, where noted, **StyleGAN3**) implementations as our image prior. The generator provides a smooth latent manifold on which we optimize trajectories; we do not introduce architectural modifications beyond standard configuration (resolution/weights).

Image→latent encoder (inversion). To place real images on the generator's latent manifold, we train a lightweight encoder that maps an input image to a single latent vector compatible with the generator's input space. Its role is purely representational: enable mapping for endpoints and faithful reconstructions; exact layer choices are not critical to the method.

Discriminator (training-only). When (re)training a generator, we use the standard discriminator bundled with the official StyleGAN repositories. It is only a training counterpart—*never* used by our optimization or evaluation procedures.

Task classifiers (decision function f). For each dataset/attribute, we use a conventional supervised image classifier (e.g., VGG-19 from TorchVision) as the decision function whose prediction we seek to change. These models are straightforward baselines chosen for familiarity and availability; they are not part of the perceptual metric.

Robust backbones (perceptual geometry & evaluation). To define our robust perceptual metric and for geometry-aware , we rely on *adversarially trained* ImageNet backbones sourced from public robustness libraries. These networks are used *only* to induce a perceptually aligned geometry and to score distances; they are distinct from the task classifier f.

Why these choices. The generator supplies a strong visual prior (manifold parameterization), the encoder puts real data on that manifold, the classifier defines the target decision boundary, and robust backbones define a perceptually grounded geometry. This separation lets us optimize counterfactual *paths* on a high-quality manifold while keeping the decision function and the perceptual metric decoupled.

Requirements for each method. Tables 2 and 3 summarize practical requirements and optimization burden. All methods require a generator g and (for real images) an encoder e; only PCG additionally uses a robust backbone to induce the perceptual geometry. Unlike RSGD variants, PCG does not perform metric inversion (no CG solves), which keeps its runtime *below* RSGD/RSGD-C despite being path-based; qualitatively it is "Medium," while RSGD and RSGD-C are "High" and "Highest," respectively. REVISE and VSGD remain the lightest due to single-point Euclidean updates without metric operations.

Table 2: Component requirements by method. "Yes/Optional" means the encoder is needed for real-image inversion but optional for synthetic latents.

Method	Generator g	Encoder e	Classifier f	Robust backbone
PCG (ours)	Yes	Yes/Optional	Yes	Yes
RSGD-C	Yes	Yes/Optional	Yes	No
RSGD	Yes	Yes/Optional	Yes	No
REVISE	Yes	Yes/Optional	Yes	No
VSGD	Yes	Yes/Optional	Yes	No

Table 3: Optimization and compute summary. "Metric inversion" refers to solving $G_Z(z)\, r = \nabla_z \mathcal{L}$ (e.g., via Conjugate Gradients).

Method	Optimization Style	Metric inversion	Relative compute
PCG (ours)	Path optimization (two-phase: energy then energy+cls)	No	Medium
RSGD-C	Single-point Riemannian descent (feature-space pullback)	Yes (CG)	Highest
RSGD	Single-point Riemannian descent (pixel-space pullback)	Yes (CG)	High
REVISE	Single-point Euclidean descent (distance + cls)	NA	Low
VSGD	Single-point Euclidean descent (cls only)	NA	Lowest

A.4 Metric Composition, Robust Backbones, and Smoothness

Backbone choice. We instantiate the perceptual geometry using *adversarially trained* ImageNet backbones (default: robust ResNet-50). These networks are used only to induce the metric and for geometry-aware evaluation; they are *never* the same model as the task classifier f, and their weights remain frozen.

Composite metric (layer aggregation). Our composite perceptual metric is constructed by pulling back the Euclidean metric from multiple activation layers of the robust backbone. Concretely, we aggregate features from the stem (layer 0) and all four residual stages (layers 1–4) of a robust ResNet-50. Layers are balanced via simple normalization weights so that no single stage dominates. The standard metric is based on the same layers but from a standard ResNet-50.

Algorithm 1 Perceptual Counterfactual Geodesics (PCG)

Require: Input image $x^\star$, target class y', encoder e, generator g, classifier f
Require: Robust feature maps $\{h_k\}_{k=1}^{K}$, path length T, Phase-1 steps S_1, Phase-2 steps S_2
Require: Learning rate η, loss weight schedule $\{\lambda_s\}_{s=1}^{S_2}$, re-anchoring period P

function RobustEnergy($\mathbf{z} = [z_0, \dots, z_T]$)
  $\delta t \leftarrow 1/T, \quad E \leftarrow 0$
  for $i = 0$ to $T-1$ **do**
    for $k = 1$ to K **do**
      $u_{ik} \leftarrow h_k(g(z_{i+1})) - h_k(g(z_i))$
      $E \leftarrow E + \frac{1}{2}\,\frac{1}{\delta t}\,\|u_{ik}\|_2^2$
    end for
  end for
  return E
end function

Initialization:
$z_0 \leftarrow e(x^\star)$
Choose a target-class sample x_{tgt} with $\arg\max f(x_{\text{tgt}}) = y'$
$z_T \leftarrow e(x_{\text{tgt}})$
Initialize $\{z_i\}_{i=1}^{T-1}$ by linear interpolation between z_0 and z_T

Phase 1: Robust geodesic with fixed endpoints
for $s = 1$ to S_1 **do**
  $E \leftarrow$ RobustEnergy($[z_0, \dots, z_T]$)
  Compute $\nabla_{z_1,\dots,z_{T-1}} E$ by backprop
  for $i = 1$ to $T-1$ **do**
    $z_i \leftarrow z_i - \eta\,\nabla_{z_i} E$
  end for
end for

Phase 2: Endpoint-aware refinement under classification constraint
for $s = 1$ to S_2 **do**
  $E \leftarrow$ RobustEnergy($[z_0, \dots, z_T]$)
  $\mathcal{L}_{\text{cls}} \leftarrow \ell(f(g(z_T)), y')$
  $\mathcal{L} \leftarrow E + \lambda_s\,\mathcal{L}_{\text{cls}}$
  Compute $\nabla_{z_1,\dots,z_T} \mathcal{L}$ by backprop
  for $i = 1$ to $T-1$ **do**
    $z_i \leftarrow z_i - \eta\,\nabla_{z_i} \mathcal{L}$
  end for
  $z_T \leftarrow z_T - \eta\,\nabla_{z_T} \mathcal{L}$ ▷ endpoint update
  if $s \bmod P = 0$ **then** ▷ re-anchoring
    Re-anchor z_T to the closest point along the path classified as y'
    Densify path by inserting midpoints and resampling to $T+1$ points
  end if
end for

Return final path $[z_0, \dots, z_T]$ and counterfactual $x_{\text{cf}} = g(z_T)$

Smoothness. To ensure the induced metric varies smoothly, we replace non-smooth ReLU variants in our models with Softplus *post hoc* (after training). In practice this does not materially change behavior, as activations typically operate in smooth regions; it only guarantees that the metric field is differentiable along the paths we optimize.

Rationale. Robust backbones provide activation spaces that better align with human perceptual similarity than standard models, leading to latent geometries that discourage brittle directions and yield smoother, semantically coherent trajectories on the generator's manifold. Keeping the robust

backbone distinct from f avoids biasing optimization toward a particular decision head and cleanly separates geometry from classification.

B More Results & Analysis

B.1 Interpolation Results

Figures 5 and 6 compare straight-line interpolations under four geometries based on StyleGAN3. From top to bottom in each panel: (i) Z-linear interpolation (flat latent space), (ii) pixel-space MSE pullback ($\mathcal{X}_{\text{MSE}}$), (iii) standard feature pullback ($\mathcal{F}_{\text{MSE}}$), and (iv) our robust perceptual pullback ($\mathcal{R}_{\text{MSE}}$). The robust metric produces smooth, on-manifold transitions with consistent semantics (identity/pose for faces; class coherence for animals), while Z-lerp and pixel MSE exhibit mid-trajectory artifacts and blends. The standard feature pullback improves semantics but still suffer from similar failure modes. These visuals mirror the trends discussed in the main text and motivate using a robust geometry for PCG.

Figure 5: Interpolations on FFHQ under four geometries. Rows (top to bottom): Z-lerp, $\mathcal{X}_{\text{MSE}}$ pullback, $\mathcal{F}_{\text{MSE}}$ pullback, and robust $\mathcal{R}_{\text{MSE}}$ pullback. The robust row shows a smooth, semantically consistent evolution (e.g., gradual attribute change without identity drift), whereas the other geometries introduce off-manifold blends and texture/illumination artifacts mid-path.

Figure 6: Interpolations on AFHQ under four geometries. Same ordering as Fig. 5. The robust $\mathcal{R}_{\text{MSE}}$ path preserves class coherence and yields clean transitions, while Z-lerp and $\mathcal{X}_{\text{MSE}}$ produce ambiguous hybrids and brittle textures; $\mathcal{F}_{\text{MSE}}$ reduces but does not eliminate these effects.

B.2 PERCEPTUAL COUNTERFACTUAL GEODESICS ACROSS AFHQ AND FFHQ.

Figures 7 (AFHQ, two examples) and 8 (FFHQ, two examples) visualize the two-phase PCG procedure with STYLEGAN3. In each panel, the top row is the initial linear path in Z (straight interpolation between the encoded input and a target exemplar), which often drifts off-manifold or blends semantics mid-trajectory. The middle row is the Phase 1 robust geodesic with fixed endpoints; transitions become smooth and class-consistent. The bottom row is the Phase 2 counterfactual geodesic, where the endpoint is jointly refined with the classification loss; the endpoint moves closer to the input while achieving the target class/attribute, and the entire path remains on-manifold. Qualitatively, AFHQ preserves species structure and textures, while FFHQ preserves identity and pose as attributes change, supporting the claims about semantic fidelity and geometry-aware paths.

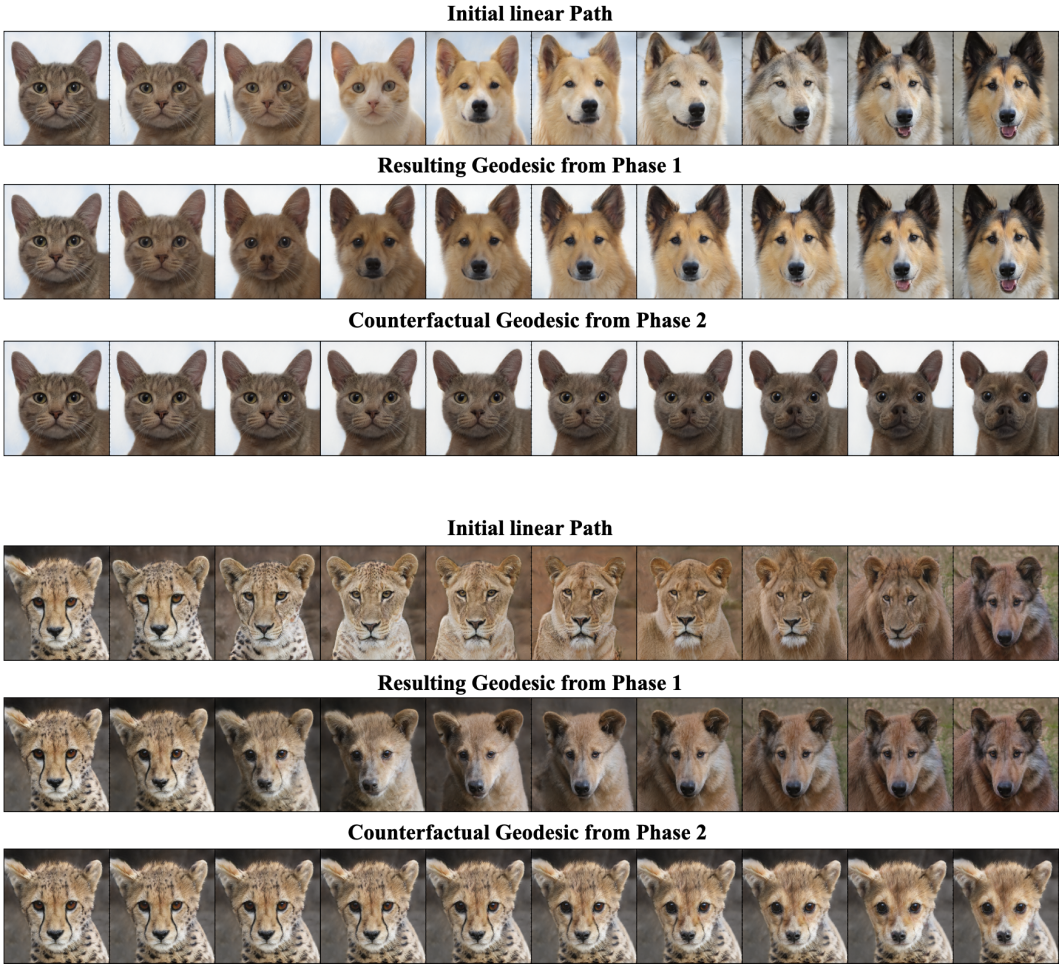

Figure 7: PCG on AFHQ (STYLEGAN3), two examples (Cat → Dog & Wild → Dog). Rows (top to bottom): initial linear path in Z between the encoded input and a target exemplar; Phase 1 robust geodesic (energy-only) with fixed endpoints; Phase 2 counterfactual geodesic after endpoint refinement with classification loss. The geodesic rows remove mid-path blends and keep species-level semantics while reaching the target class.

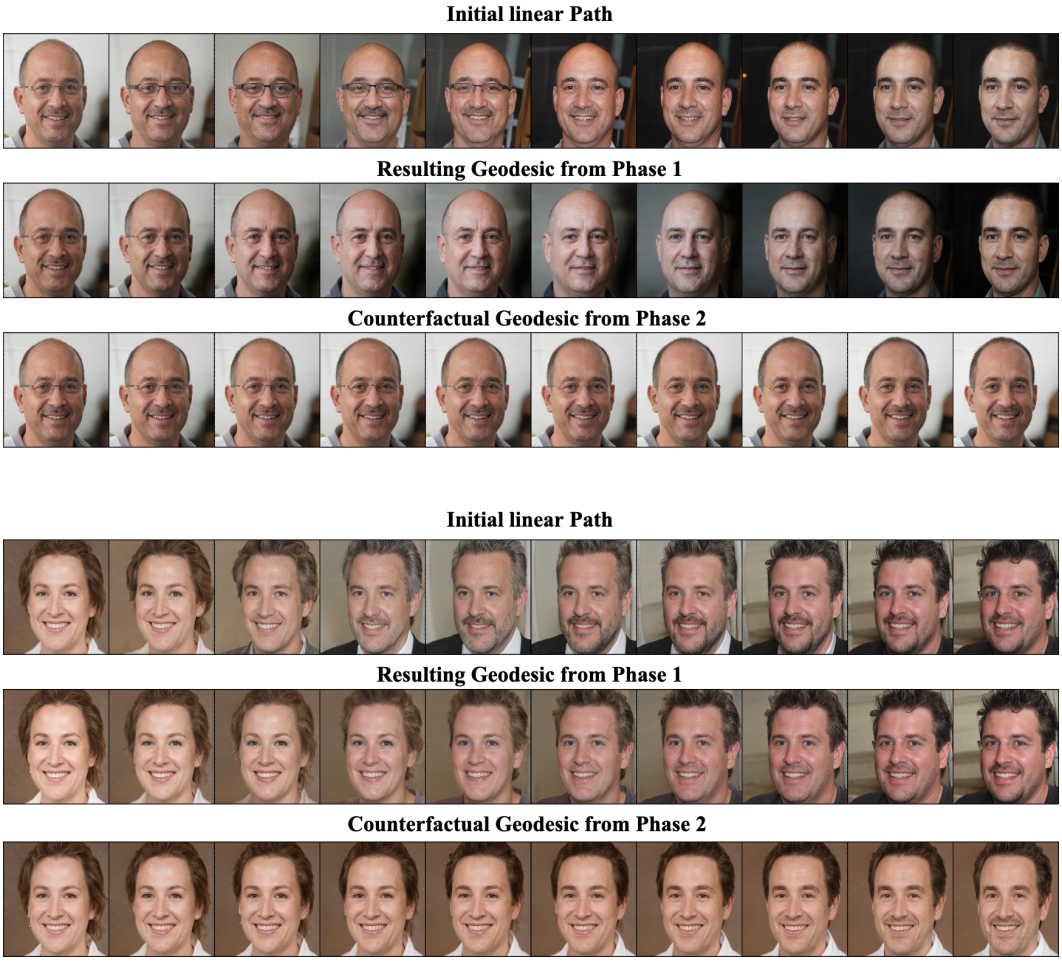

Figure 8: PCG on FFHQ (StyleGAN3), two examples (Glasses → No-glasses & Female → Male). Same layout as Fig. 7. Phase 1 produces smooth, on-manifold transitions; Phase 2 moves the endpoint toward the input while satisfying the target classifier. Identity and pose are largely preserved as the target attribute changes, and intermediate frames remain perceptually coherent.

B.3 Sensitivity to Different Target Class Samples

Figures 9 and 10 test how PCG depends on which target-class exemplar is used to initialize the path. For each input we run PCG twice, once per exemplar. We observe that the Phase 1 geodesic reflects the chosen exemplar (different coarse routes in latent space), but after Phase 2 (endpoint refinement with classification loss) the counterfactual geodesics converge to a tight neighborhood around the input while achieving the target label/attribute. This yields diverse yet faithful counterfactuals and supports the main-text claim about robustness to target initialization.

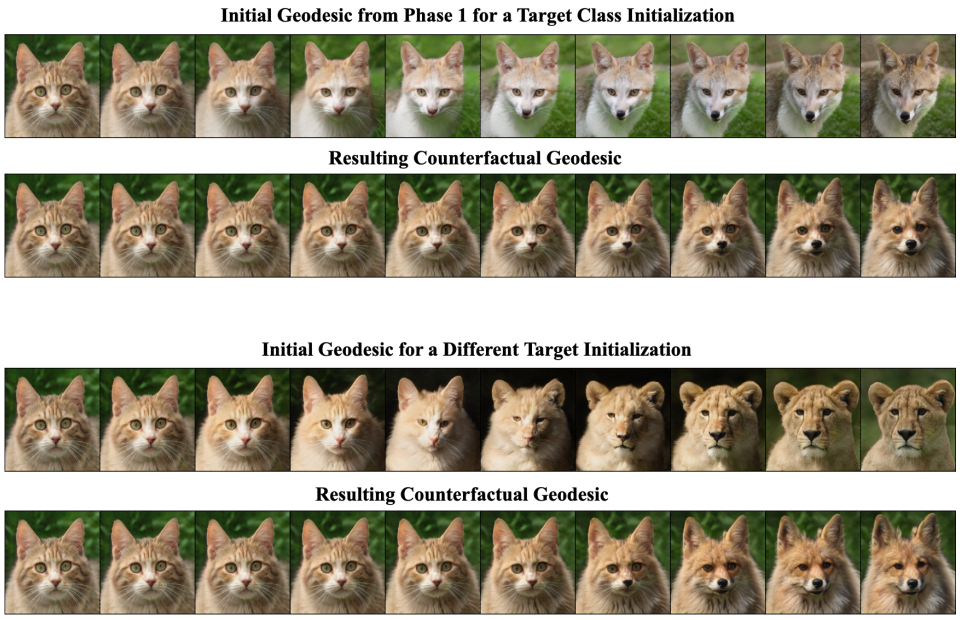

Figure 9: Sensitivity to target exemplar on AFHQ (StyleGAN3) (Cat → Wild). Rows: (1) Phase 1 geodesic initialized with target exemplar A, (2) resulting Phase 2 counterfactual geodesic, (3) Phase 1 geodesic with a different exemplar B, (4) resulting Phase 2 counterfactual geodesic. Although the Phase 1 routes differ, the Phase 2 counterfactuals converge near the input and satisfy the target class, indicating low sensitivity to the exemplar choice and producing diverse but faithful variations.

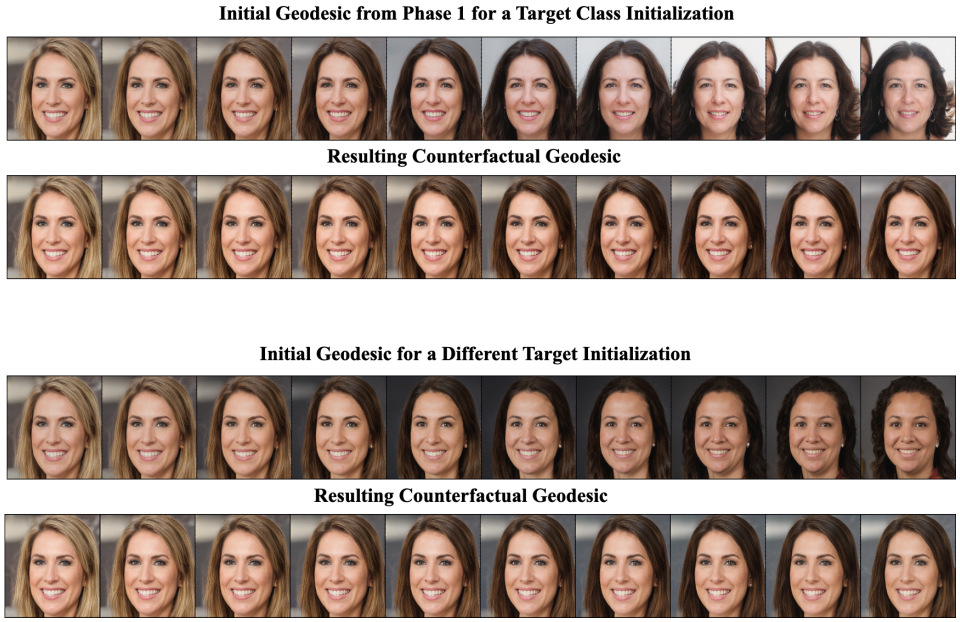

Figure 10: Sensitivity to target exemplar on FFHQ (StyleGAN3) (Blonde → Non-blonde). Same layout as Fig. 9. Two different target exemplars lead to distinct Phase 1 paths, yet the Phase 2 counterfactual geodesics converge to a small neighborhood around the input while achieving the target attribute; identity and pose remain largely preserved.

Quantitative sensitivity (three initializations per input). To quantify low sensitivity to target initialization, we run PCG $M{=}3$ times per input with different target exemplars and measure how

close and consistent the resulting counterfactuals (CFs) are. We use the LPIPS perceptual distance (Zhang et al., 2018) and report two intuitive, scale-aware metrics: (i) **CF dispersion ratio (CDR)** — how tightly CFs cluster compared to typical variation within the target class, and (ii) **CF diameter** — the worst-case dissimilarity among the CFs.

Definitions. Let $C = \{x_{\mathrm{cf}}^{(1)}, x_{\mathrm{cf}}^{(2)}, x_{\mathrm{cf}}^{(3)}\}$ be the CF endpoints for one input $x^\star$. Let $\mathrm{LPIPS}(\cdot,\cdot)$ denote the perceptual distance, and let $\overline{d}_{\mathrm{tgt}}$ be the average LPIPS between random pairs sampled from the *target* class (estimated once per dataset/attribute using 30 random pairs).

(1) CDR:

$$\overline{d}_{\mathrm{CF}} = \frac{2}{M(M-1)} \sum_{m<n} \mathrm{LPIPS}\big(x_{\mathrm{cf}}^{(m)}, x_{\mathrm{cf}}^{(n)}\big), \qquad \mathrm{CDR} = \frac{\overline{d}_{\mathrm{CF}}}{\overline{d}_{\mathrm{tgt}}}.$$

$\mathrm{CDR} \ll 1$ indicates CFs form a cluster much tighter than typical target-class variability.

(2) CF diameter:

$$\mathrm{Diam}_{\mathrm{CF}} = \max_{m<n} \mathrm{LPIPS}\big(x_{\mathrm{cf}}^{(m)}, x_{\mathrm{cf}}^{(n)}\big),$$

so a small value guarantees even the most dissimilar CFs remain close.

Sensitivity summary (AFHQ and FFHQ) based on StyleGAN2. Using the LPIPS-based metrics defined above, Table 4 reports the *CF dispersion ratio* (CDR) and *CF diameter* for three target-initializations per input (mean ± std). CDR is the intra-CF mean LPIPS normalized by a target-class baseline computed from 30 random target pairs; CF diameter is the maximum pairwise LPIPS among the three CFs. For both metrics, *lower is better* (tighter clustering and smaller worst-case gap).

For both datasets, $\mathrm{CDR} \ll 1$ shows that CFs produced from different target exemplars form a tight cluster relative to the target-class spread, consistent with our claim that PCG converges on diverse yet faithful counterfactuals; the small CF diameters confirm this even in the worst case.

Table 4: Sensitivity to target initialization (three runs per input). LPIPS-based tightness across counterfactuals (lower is better). CDR is the intra-CF mean LPIPS normalized by the target-class baseline LPIPS (estimated from 30 random target pairs).

Dataset / Task	CDR (LPIPS)	CF Diameter (LPIPS)
AFHQ: cat → dog	0.19 ± 0.06	0.18 ± 0.05
FFHQ: not-smile → smile	0.21 ± 0.08	0.16 ± 0.04
FFHQ: bald → hairy	0.28 ± 0.05	0.19 ± 0.05

B.4 Quantitative Results based on StyleGAN3

As in the main text, PCG consistently achieves the lowest values under the geometry-aware metrics $\mathcal{L}_{\mathcal{F}}$ and $\mathcal{L}_{\mathcal{R}}$ and remains competitive under pixel metrics. These appendix results, obtained on StyleGAN3, show that the robust geodesic formulation retains its advantage without re-tuning and confirm the stability of PCG's behaviour across model choices.

Table 5: Quantitative comparison across datasets.

Method	AFHQ				FFHQ			
	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_{\mathcal{F}}$	$\mathcal{L}_{\mathcal{R}}$	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_{\mathcal{F}}$	$\mathcal{L}_{\mathcal{R}}$
REVISE	1.18±0.12	**0.72**±0.17	1.05±0.10	2.68±0.04	0.81±0.07	**0.33**±0.12	0.81±0.09	2.75±0.06
VSGD	1.30±0.11	1.48±0.15	1.57±0.09	2.88±0.08	0.78±0.11	0.95±0.10	1.49±0.12	2.83±0.08
RSGD	0.84±0.08	1.30±0.09	0.68±0.07	1.83±0.05	0.60±0.05	0.83±0.07	0.60±0.04	2.39±0.05
RSGD-C	0.92±0.10	1.43±0.16	0.63±0.08	1.73±0.06	0.67±0.06	0.91±0.09	0.47±0.04	2.08±0.05
PCG (ours)	**0.78**±0.07	1.13±0.10	**0.51**±0.06	**0.30**±0.02	**0.41**±0.03	0.71±0.09	**0.38**±0.05	**0.21**±0.05

B.5 Runtime Complexity on AFHQ

On AFHQ, measured on a single NVIDIA H100 GPU, Table 6 reports per-sample wall-clock runtimes and speedups across methods based on StyleGAN2. VSGD is the fastest (1.6 min). PCG runs in 3.4 min per sample despite being path-based (here $T{=}10$): with a GPU, all path nodes and robust-feature evaluations are batched in a single forward/backward, so the extra cost is modest. RSGD is slowest (5.7 min) because each step requires solving $G_Z(z)\, r = \nabla_z \mathcal{L}$ with Conjugate Gradients; the inner CG iterations and repeated Jacobian–vector products through g (and, for RSGD-C, the feature backbone) dominate wall-clock. Absolute times depend on precision and batch sizing, but the relative ordering was consistent across runs.

Table 6: AFHQ per-sample wall-clock runtime (minutes). RSGD serves as a representative for RSGD/RSGD-C; VSGD represents standard Euclidean-gradient methods.

Method	Time (min)	Speedup vs RSGD	Notes
VSGD (rep. Euclidean)	1.6	3.56x	Classification loss only; lowest cost.
PCG (ours)	3.4	1.68x	Path-based with $T{=}10$ nodes; nodes batched on GPU.
RSGD (rep. RSGD/–C)	5.7	1.00x	Natural-gradient via CG; Jacobian–vector products dominate.

C Limitations & Future Work

PCG depends on pretrained generators/encoders, shared by all latent-space approaches, and robust vision backbones; this reliance may limit applicability in domains where such resources are scarce or hard to train. While our robust metric and two-stage path refinement mitigate artifacts from imperfect generators, PCG cannot fully overcome a severely mis-specified latent space. In terms of computation, PCG is path-based, but with short paths and batched evaluation its per-sample cost is moderate and typically below RSGD/RSGD-C, which incur additional expense from metric inversion via conjugate gradients. Our study focuses on images; extending the framework beyond vision (e.g., graphs or language) is non-trivial and left open. Future work includes: (i) multimodal extensions that couple text and image spaces (e.g., CLIP or diffusion backbones) via joint latent geometries and cross-modal robust metrics; (ii) video counterfactuals, incorporating temporal coherence (motion consistency, content persistence) and video backbones to define spatiotemporal perceptual geometry; and (iii) methods for low-resource regimes, such as lightweight robust feature surrogates or few-shot adaptation of perceptual metrics.