GRACE: POPULATION-LEVEL ADVERSARIALLY ROBUST COUNTERFACTUAL EXPLANATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Counterfactual explanations (CFs) suggest changes to an input to flip a machine learning model's prediction toward a desired outcome. Most existing CF generators are not robust because they assume a static model and data distribution. However, models deployed in practice are retrained periodically, so a CF that is valid today may not be so tomorrow. Institutions typically issue a population of CFs, making robustness a population-level property, rather than an individual one. Existing robustness methods, however, optimize each CF against its own worst-case perturbation rather than the shared population-level risk induced by retraining, which may cost more proximity than necessary. We introduce Geometrically Restricted Adversarially robust Counterfactual Explanations (GRACE), a distributionally robust wrapper that treats the population of CFs produced by any CF generator as the robustness target. GRACE controls a surrogate for the population's worst-case invalidity under future retraining while keeping the refined CFs close to both the original CFs and the observed distribution. Across tabular and image datasets, GRACE substantially improves the future validity of the base CFs with only a small increase in proximity, matching or exceeding the future validity of recent state-of-the-art robust CF baselines on tabular benchmarks.

1 INTRODUCTION

Counterfactual explanations (CFs) help users identify changes that would alter a machine learning model's prediction, thereby providing local insights into the model for particular instances. Most CF methods assume a fixed model and data distribution, but deployment is typically dynamic (Quionero-Candela et al., 2009): evolving data distributions and policy changes trigger retraining, and a CF computed for the deployed model can lose its validity once that model changes (Rawal et al., 2020). A recommended action is only useful if it still delivers the desired outcome under whichever model version that evaluates the outcome when the individual acts on it. For example, an applicant told to raise their income by $3,000 to secure loan approval may find that recourse has expired by the time they can act on it, undermining the trustworthiness and utility of CFs in dynamic environments.

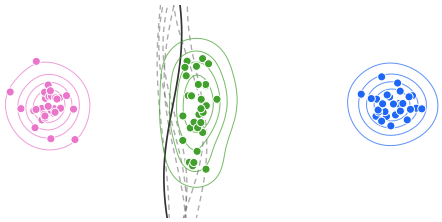


Figure 1: **Illustrative example of decision boundary shift.** A population of CFs (green) sits near a decision boundary (solid curve) separating two class distributions (pink and blue). Dashed curves illustrate plausible boundary shifts induced by retraining. Because the same retrained model evaluates the entire CF population, robustness is naturally a population-level property.

This robustness concern is crucial at the population level (Figure 1). A bank, for instance, does not issue a single CF but a population of them, at least one per rejected applicant. When the model is retrained, a substantial fraction of that population can become invalid simultaneously even when the data distribution changes slightly. From an institutional perspective, what needs to be controlled is not only the validity of any individual CF, but also the *distributional risk of invalidity* across the population as a whole. While instance-by-instance certification can upper-bound the latter conservatively, it does not model it directly: each CF may use a different worst-case model and may guard against scenarios that never jointly arise, potentially moving CFs further than necessary.

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.

Current robustness strategies for CFs, however, primarily formulate robustness at the level of individual CFs rather than at the level of the CF population itself. Some methods improve robustness without explicitly modeling the update mechanism for model retraining, using stability- or confidence-based criteria, data augmentation, or local surrogates (Dutta et al., 2022; Ferrario & Loi, 2022; Bui et al., 2025), while others represent uncertainty directly in model or predictor space through bounded shifts, distributions, or collections of possible updated models (Hamman et al., 2023; Upadhyay et al., 2021; Kayastha et al., 2024; Kyaw et al., 2025; Nguyen et al., 2023; Marzari et al., 2024; Stepka et al., 2025; Duell & Fan, 2026). Closer to our setting, Nguyen et al. (2022) perturb the training distribution directly but certify recourse one input at a time, while Bui et al. (2022) consider multiple CFs for a single instance. None of these methods treats the population of CFs generated across instances as the object of robustness itself.

We address this gap with **G**eometrically **R**estricted **A**dversarially robust **C**ounterfactual **E**xplanations (GRACE), a post-hoc distributional refinement that treats the population of CFs produced by any CF generator as a distribution to be moved toward a region of lower worst-case future risk, subject to a bounded transport detour that balances proximity to the generator's output and the favorable class. Unlike the explicit, per-instance approaches above, our worst case couples the entire CF population through a single retrained classifier induced by a plausible shift in the training distribution, rather than assigning each CF its own worst-case model. As we demonstrate in this work, our new population-level approach substantially improves the future validity of the base CFs with relatively little additional proximity cost, while matching or exceeding the future validity of existing robust methods.

Our main contributions are:

- We formalize CF-population robustness as worst-case risk under a Wasserstein ball of plausible training-distribution shifts, where each shift induces a shared retrained classifier for the entire CF population, and introduce GRACE, which refines the population within a bounded transport-detour region.
- We show that the population robust risk degrades at most linearly with increasing training-distribution uncertainty, and realize the refinement through an update-aware Wasserstein sensitivity objective that propagates training-distribution perturbations through the prescribed model-update procedure.
- Across the settings we test, GRACE improves Future Validity with only modest additional proximity cost, matching or exceeding future validity of robust CF baselines.

2 RELATED WORK

We organize prior robust CF methods primarily by *where* uncertainty about future model changes is represented, while also distinguishing *whether* robustness is defined instance-wise or at the population level.

Update-mechanism-agnostic robustness. Some methods improve CF robustness without explicitly modeling the mechanism that produces the future predictor. They optimize or refine CFs using stability- or confidence-based criteria (Dutta et al., 2022), augment training data with generated CFs (Ferrario & Loi, 2022), or perturb the covariance of data sampled near the decision boundary to train a linear surrogate (Bui et al., 2025). These approaches target properties associated with robustness to future updates, but do not explicitly link a shift in the training distribution to the resulting updated predictor.

Model-space uncertainty. Another line of work represents future model changes directly in parameter or predictor space. Hamman et al. (2023) characterize naturally occurring model changes through constraints on predictive behavior and derive probabilistic robustness guarantees for CFs; Xu & Wu (2024) generalize this model-change condition to a broader subgaussian setting. Upadhyay et al. (2021) introduce ROAR, a minimax objective that generates CFs robust to plausible parameter shifts under a linear surrogate model; Kayastha et al. (2024) give a provably optimal solver for generalized linear models, and Kyaw et al. (2025) extend this formulation to general bounded neighborhoods. Nguyen et al. (2023) similarly hedge against ambiguity in future model-parameter

distributions through a min–max objective based on Gelbrich distance. Jiang et al. (2023; 2024a;c) study robustness under bounded model changes, including interval-based propagation of weights, biases, and hidden-unit activations. APΔS (Marzari et al., 2024; 2026) samples plausible model shifts directly and shows that exact robustness decisions are NP-hard.

A probabilistic line of work instead reasons over distributions or collections of possible updated models. Stepka et al. (2025) move a base CF into a region satisfying a specified robustness level, while Duell & Fan (2026) take a Bayesian approach, optimizing for high confidence and low variance in the posterior predictive distribution of a CF and deriving a KL-based graceful-degradation guarantee. Related work on predictive or model multiplicity studies the existence of multiple near-optimal predictors that can exhibit different individual-level behavior (Breiman, 2001; Marx et al., 2020; Black et al., 2022; Rudin, 2019; Coston et al., 2021). Its implications for CFs and recourse have been studied by Pawelczyk et al. (2020) and Turbal et al. (2025). These approaches differ in how the set of plausible models is constructed, but they represent uncertainty primarily in model or predictor space rather than explicitly linking training-distribution shifts to the resulting retrained predictor. See Jiang et al. (2024b) for a broader survey of robust CF methods.

Data-space uncertainty and population-level robustness. A smaller line of work perturbs the training distribution itself, which is closest to our setting. Nguyen et al. (2022) model training-data uncertainty through a Gaussian-mixture ambiguity set prescribed using Wasserstein distance, but certify robustness one input at a time. Bui et al. (2022) instead consider a counterfactual plan consisting of multiple candidate modifications for a single instance and study their joint validity under model uncertainty. Our proposed method GRACE is closest to Nguyen et al. (2022) in placing uncertainty on the training distribution, but differs by treating the full population of generated CFs across instances, rather than individual inputs, as the object of robustness.

Wasserstein robust optimization (WRO). WRO minimizes worst-case expected loss over distributions within a Wasserstein ambiguity set around a reference distribution (Mohajerin Esfahani & Kuhn, 2018; Blanchet et al., 2019; Gao & Kleywegt, 2023). Its tractability commonly relies on strong duality, which reformulates the inner worst-case expectation as a finite- or semi-infinite-dimensional optimization problem (Mohajerin Esfahani & Kuhn, 2018; Gao & Kleywegt, 2023). In our setting, however, the uncertain future training distribution first induces an updated predictor *through retraining*, which is then evaluated on the CF population. Thus, training-distribution uncertainty induces model uncertainty, preventing direct application of the standard Wasserstein dual reformulation. We therefore use an update-aware first-order Wasserstein sensitivity approximation that propagates perturbations of future training distribution through the prescribed model-update procedure (Section 3.4).

3 DISTRIBUTIONALLY ROBUST COUNTERFACTUAL REFINEMENT

We propose GRACE, a post-hoc refinement of a base CF generator's output that minimizes worst-case future risk under retraining-induced model updates. We treat the generated CF population as the decision variable and measure its geometric displacement with Wasserstein distance. Unrestricted robust optimization, however, can move the population arbitrarily far from both the base CFs, discarding their recourse structure, and the favorable class, leaving the observed geometry of favorable outcomes. We therefore constrain it relative to both anchors, yielding a two-anchor Wasserstein geometry. We approximate the resulting robust objective via update-aware Wasserstein sensitivity and optimize the population's support locations under this constraint. We first formulate the objective and its geometric restriction, then characterize its graceful degradation under increasing shift, and finally describe the numerical realization and deployment.

3.1 PROBLEM SETTING AND POPULATION ROBUSTNESS

Let $f_0 : \mathcal{X} \to [0, 1]$ be the current classifier and $\tau \in (0, 1)$ the decision threshold. An input x is favorable if $f_0(x) \geq \tau$. For unfavorable instances $X = \{x_i\}_{i=1}^n$, a base generator G produces CFs $z_i = G(x_i)$, defining $\mathbb{P}_{\mathrm{CF}} = n^{-1} \sum_{i=1}^n \delta_{z_i}$. We treat CF populations as elements of $\mathcal{P}_2(\mathcal{X})$, the distributions on $\mathcal{X}$ with finite second moment. Let $\mathbb{P}_1$ denote the empirical distribution of training

samples that are favorable under f_0. These two anchors capture where the generator places CFs and where favorable outcomes are observed.

Since model updates are driven by changes in the training data, we place uncertainty on the training distribution rather than directly on model parameters, whose perturbation sets are hard to calibrate and need not correspond to a realizable retraining. Let $\widehat{\mathbb{P}}_{\mathrm{tr}}$ be the empirical training distribution and W_{data} a Wasserstein metric on training distributions, which captures shifts that move samples in feature space rather than merely reweighting them. For a Wasserstein uncertainty radius $r \geq 0$, define the data-side uncertainty set

$$\mathcal{U}_r = \{\mathbb{P} : W_{\mathrm{data}}(\mathbb{P}, \widehat{\mathbb{P}}_{\mathrm{tr}}) \leq r\},$$

where each $\mathbb{P} \in \mathcal{U}_r$ induces a future classifier $f_{\mathbb{P}} = \mathrm{Update}(f_0, \mathbb{P})$ via the prescribed model-update procedure (Section 3.4).

Validity depends on a threshold, $f(z) \geq \tau$, whose indicator is discontinuous and provides no gradient signal. We therefore measure the invalidity of a CF z under a classifier f by the margin loss $\ell_f(z) = [\tau - f(z)]_+$, which is zero when z is valid and grows with its short fall. Averaging over the population and taking the worst case over plausible retraining gives the population robust risk

$$\Phi_r(\mathbb{Q}) := \sup_{\mathbb{P} \in \mathcal{U}_r} \mathbb{E}_{z \sim \mathbb{Q}}\big[\ell_{f_{\mathbb{P}}}(z)\big], \qquad \mathbb{Q} \in \mathcal{P}_2(\mathcal{X}). \tag{1}$$

The expectation measures invalidity across all issued CFs, and the supremum, taken outside it, selects a single updated model $f_{\mathbb{P}}$ against which the *entire* CF population $\mathbb{Q}$ is scored, rather than allowing each CF its own worst-case model, matching deployment where one retrained model evaluates every CF. We minimize this population-level risk over $\mathbb{Q}$ while keeping the base generator G fixed. Although Φ_r is not the invalidity rate itself, it controls population-level margin violations. Appendix A.1 formalizes this connection and describes the smooth approximation used for optimization.

3.2 Two-Anchor Geometry for Robust Refinement

Our goal is to move the base CFs into a region of low robust risk without losing their value as recourse. Unconstrained minimization of Φ_r fails, since the risk is lowest deep inside the favorable region, far from the factual inputs. Anchoring only to $\mathbb{P}_{\mathrm{CF}}$ gives no guidance on where to move, while anchoring only to $\mathbb{P}_1$ ignores the base generator and collapses CFs toward typical favorable points. We therefore constrain the refinement relative to both anchors, measuring displacement with optimal transport, which remains well defined between empirical distributions with disjoint supports. Let W_{CF} be the 2-Wasserstein metric on $\mathcal{P}_2(\mathcal{X})$, unlike W_{data}, it measures displacement of CF populations rather than training data shifts.

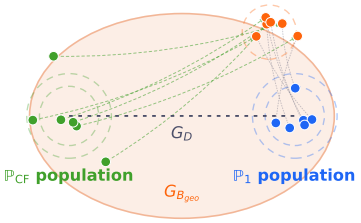


Figure 2: **Bounded-detour region.** $\mathbb{P}_{\mathrm{CF}}$ (green) and $\mathbb{P}_1$ (blue) are connected via the zero-excess path $\mathcal{G}_D$ through the bounded-detour region admitted by B_{geo}.

From a two-anchor lens to a single geometry budget. Intersections of Wasserstein balls provide a natural way to constrain a distribution relative to two reference measures (Zhao et al., 2026). In our setting, the two anchors provide complementary geometric controls on the refinement: $\mathbb{P}_{\mathrm{CF}}$ limits departure from the recourse proposed by the base CF generator, while $\mathbb{P}_1$ anchors the refinement to the observed geometry of favorable outcomes. For radii $\delta, \epsilon \geq 0$, the two anchors define

$$\mathcal{L}_{\delta,\epsilon} = \{\mathbb{Q} \in \mathcal{P}_2(\mathcal{X}) : W_{\mathrm{CF}}(\mathbb{P}_{\mathrm{CF}}, \mathbb{Q}) \leq \delta,\ W_{\mathrm{CF}}(\mathbb{Q}, \mathbb{P}_1) \leq \epsilon\}\,.$$

Fixing δ and ϵ separately prescribes in advance how the available geometric freedom is allocated between the two anchors. For a geometry budget $B_{\mathrm{geo}} \geq 0$, we instead constrain only the sum:

$$\mathcal{G}_{B_{\mathrm{geo}}} = \bigcup_{\delta+\epsilon \leq B_{\mathrm{geo}}} \mathcal{L}_{\delta,\epsilon} = \{\mathbb{Q} \in \mathcal{P}_2(\mathcal{X}) : W_{\mathrm{CF}}(\mathbb{P}_{\mathrm{CF}}, \mathbb{Q}) + W_{\mathrm{CF}}(\mathbb{Q}, \mathbb{P}_1) \leq B_{\mathrm{geo}}\}\,.$$

This replaces two parameters with one and lets the optimizer choose the allocation based on where low-risk regions lie. Let $D := W_{\mathrm{CF}}(\mathbb{P}_{\mathrm{CF}}, \mathbb{P}_1)$. By the triangle inequality, $\mathcal{G}_{B_{\mathrm{geo}}}$ is nonempty only if $B_{\mathrm{geo}} \geq D$, so we set $B_{\mathrm{geo}} = \beta_{\mathrm{geo}} D$ with $\beta_{\mathrm{geo}} \geq 1$. Here $\beta_{\mathrm{geo}} = 1$ forces the population onto a shortest path between the anchors, while $\beta_{\mathrm{geo}} > 1$ permits bounded detours (Figure 2).

Geometry-constrained robust refinement. Our central refinement problem is

$$\boxed{\inf_{\mathbb{Q}\in\mathcal{G}_{B_{\mathrm{geo}}}} \Phi_r(\mathbb{Q}) = \inf_{\mathbb{Q}\in\mathcal{G}_{B_{\mathrm{geo}}}} \sup_{\mathbb{P}\in\mathcal{U}_r} \mathbb{E}_{z\sim\mathbb{Q}}[\ell_{f_{\mathbb{P}}}(z)]} \tag{2}$$

Here r controls uncertainty in future training data, whereas B_{geo} controls geometric freedom on the CF side. Unlike per-instance methods, which hedge each CF against its own worst-case model, Eq. (2) is a single problem with one shared adversary and one budget, so robustness is concentrated where the population is jointly most fragile.

3.3 Stability under Future Model Updates

A deployed CF population stays fixed even if future shifts exceed those anticipated, so we study how its robust risk $\Phi_r(\mathbb{Q})$ changes as the uncertainty radius r increases. To obtain a stable, differentiable data-to-model map, we instantiate the update operator as a prescribed T-step gradient update from the deployed model (Franceschi et al., 2017; Shaban et al., 2019), which supports both the analysis below and the sensitivity computation in Section 3.4.

For a differentiable parametric predictor f_θ with $f_0 = f_{\theta_0}$, define $L_{\mathbb{P}}(\theta) := \mathbb{E}_{\xi\sim\mathbb{P}}[\ell_{\mathrm{tr}}(\theta;\xi)]$, $\theta^{\mathbb{P}}_{t+1} = \theta^{\mathbb{P}}_t - \eta_t \nabla_\theta L_{\mathbb{P}}(\theta^{\mathbb{P}}_t)$, $\theta^{\mathbb{P}}_0 = \theta_0$, for $t = 0,\dots,T-1$, with prescribed step sizes $\eta_t \geq 0$, and set $f_{\mathbb{P}} := f_{\theta^{\mathbb{P}}_T}$. The following result shows that stability of this update map with respect to the training distribution yields graceful degradation of the population-level robust risk as the uncertainty radius increases.

Proposition 1 (Graceful degradation under increasing data shift)**.** *Assume the data-side Wasserstein space is geodesic. Suppose that, over the parameter region traversed by the T-step update trajectories, there exist finite constants $L_\theta, L_{\mathrm{data}}, L_f$ such that*

$$\|\nabla L_{\mathbb{P}}(\theta) - \nabla L_{\mathbb{P}}(\theta')\| \leq L_\theta \|\theta-\theta'\|, \quad \|\nabla L_{\mathbb{P}}(\theta) - \nabla L_{\mathbb{P}'}(\theta)\| \leq L_{\mathrm{data}} W_{\mathrm{data}}(\mathbb{P},\mathbb{P}'),$$
$$\sup_{z\in\mathcal{X}} |f_\theta(z) - f_{\theta'}(z)| \leq L_f \|\theta-\theta'\|.$$

Then

$$\sup_{z\in\mathcal{X}} |f_{\mathbb{P}}(z) - f_{\mathbb{P}'}(z)| \leq L_f C_T W_{\mathrm{data}}(\mathbb{P},\mathbb{P}'),$$

where $C_T := L_{\mathrm{data}} \sum_{t=0}^{T-1} \eta_t \prod_{s=t+1}^{T-1}(1+\eta_s L_\theta)$.

Consequently, for any fixed $\mathbb{Q}\in\mathcal{P}_2(\mathcal{X})$ and $0 \leq r_1 \leq r_2$,

$$\boxed{\Phi_{r_2}(\mathbb{Q}) \leq \Phi_{r_1}(\mathbb{Q}) + L_f C_T (r_2 - r_1)}. \tag{3}$$

Thus, for a CF population fixed at deployment, the worst-case robust risk degrades gracefully, growing at most linearly as data-side uncertainty increases. The proof is given in Appendix A.2.

3.4 Update-Aware Numerical Realization

Minimizing Φ_r directly is intractable: its inner supremum ranges over perturbed training distributions, each affecting the model only through the full T-step update. A common approach in robust CF methods is to perturb the model parameters directly. However, this does not explicitly preserve the link between training-distribution shifts and the resulting shared retrained model. We thus seek a surrogate that (i) is tractable to optimize over the CF population, (ii) preserves this shared-model coupling Eq. (1), and (iii) provably approximates Φ_r for small r. Rather than perturbing model parameters directly, we construct an update-aware population sensitivity: we differentiate the shared CF-population loss with respect to training-distribution perturbations propagated through update. The resulting surrogate is the nominal population loss plus a worst-case local sensitivity penalty. We derive the two terms below.

Local training perturbations. Let $\widehat{\mathbb{P}}_{\mathrm{tr}} = N^{-1}\sum_{i=1}^N \delta_{\xi_i}$ be the empirical distribution of the N training samples, where $\xi_i = (x_i, y_i)$, $x_i \in \mathbb{R}^d$, and y_i is the label. We represent local perturbations by $h \in \mathcal{H} := L^2(\widehat{\mathbb{P}}_{\mathrm{tr}};\mathbb{R}^d)$, with $\|h\|^2_{\mathcal{H}} := \mathbb{E}_{\widehat{\mathbb{P}}_{\mathrm{tr}}}\|h(\xi)\|_2^2$. For $h\in\mathcal{H}$, define $T_h(x,y) := (x + h(x,y), y)$, $\mathbb{P}_h := (T_h)_{\#}\widehat{\mathbb{P}}_{\mathrm{tr}}$, where $\#$ denotes pushforward. Let $\theta_T(h)$ denote the parameters after the prescribed T-step update on $\mathbb{P}_h$, and let $\theta^{\mathrm{nom}}_T := \theta_T(0)$ denote the corresponding parameters under the unperturbed training distribution. In general, $\theta^{\mathrm{nom}}_T \neq \theta_0$, since the former is obtained after applying the T update steps starting from θ_0.

Update-aware sensitivity. For a CF population $\mathbb{Q}$, define $F_{\mathbb{Q}}(\theta) := \mathbb{E}_{z\sim\mathbb{Q}}[\ell_{f_\theta}(z)]$. Let D_h denote differentiation with respect to the perturbation h. We define the first-order population sensitivity as $s_{\mathbb{Q}} := D_h F_{\mathbb{Q}}(\theta_T(h))\big|_{h=0}$. By the chain rule, this sensitivity propagates the aggregate population-loss gradient through the shared model-update map:

$$s_{\mathbb{Q}} = J_T^* \nabla_\theta F_{\mathbb{Q}}(\theta_T^{\mathrm{nom}}), \qquad J_T := D_h \theta_T(h)|_{h=0}, \tag{4}$$

where J_T^* denotes the corresponding adjoint of J_T (see Appendix A.3). Thus, the population loss is aggregated before sensitivity to the shared future update is evaluated. We measure the worst-case first-order magnitude by the dual norm $\|s_{\mathbb{Q}}\|_* := \sup_{\|h\|_{\mathcal{H}}\le 1} s_{\mathbb{Q}}[h]$. This quantity captures the largest first-order change in population loss over unit training-distribution perturbations. Crucially, this derivative is taken after aggregating the loss over $\mathbb{Q}$, so all CFs respond to the same perturbed model update rather than receiving independent worst-case perturbations.

First-order robust surrogate. The following proposition motivates the surrogate used in our numerical realization.

Proposition 2 (First-order update-aware approximation)**.** *Under the Wasserstein differentiability and regularity conditions stated in Appendix A.3, for any fixed feasible* $\mathbb{Q}$*,*

$$\Phi_r(\mathbb{Q}) = \widetilde{\Phi}_r(\mathbb{Q}) + o(r), \qquad r \downarrow 0.$$

Motivated by Proposition 2, we approximate the robust risk by

$$\widetilde{\Phi}_r(\mathbb{Q}) := F_{\mathbb{Q}}(\theta_T^{\mathrm{nom}}) + r\|s_{\mathbb{Q}}\|_*. \tag{5}$$

Thus, Eq. (5) is the first-order small-radius approximation of the full population robust objective, rather than an independent parameter-sensitivity regularizer. *The nominal loss drives refinement toward validity under the unperturbed model update, while the sensitivity penalty encourages the refined CF population to move toward regions whose validity is less sensitive to training-distribution-induced model updates.* We can compute $s_{\mathbb{Q}}$ in Eq. (4) efficiently by reverse-mode differentiation through the T-step update trajectory (see Appendix A.4).

Finite-support control population. For computation, we represent the decision distribution as $\mathbb{Q} = \frac{1}{n}\sum_{j=1}^n \delta_{w_j}$, and optimize the support locations $w_j \in \mathcal{X}$. We approximately minimize $\widetilde{\Phi}_r(\mathbb{Q})$ subject to $\mathbb{Q} \in \mathcal{G}_{B_{\mathrm{geo}}}$ using an augmented-Lagrangian primal–dual scheme, and denote the returned population by $\widetilde{\mathbb{Q}}_{r,B_{\mathrm{geo}}}$. Algorithm 1 in Appendix A.6 gives the full pseudocode.

Deployment. The operating parameters r and β_{geo} are selected on the calibration split and frozen before final evaluation. The resulting refined population specifies the optimized CF support points but does not by itself assign them to individual factual instances. For deployment, we therefore compute a minimum-cost one-to-one assignment between the base CFs $\{z_i\}_{i=1}^n$ and the refined supports $\{w_j\}_{j=1}^n$, using the ground cost underlying W_{CF}. Each factual instance x_i is then assigned the refined support matched to its original CF z_i. This preserves the optimized empirical population while recovering one deployable CF per instance.

4 Experiments

We evaluate GRACE as a post-hoc distributional robustness wrapper for counterfactual generators. Code is available at `https://anonymous.4open.science/r/robust_cf-6E18`.

Datasets. We evaluate GRACE across three modalities. For tabular data, we use four classification benchmarks: **HELOC** (Oliabev), **German Credit** (Hofmann, 1994), **FICO** (Fair Isaac Corporation (FICO), 2018), and **Diabetes** (Smith et al., 1988). For images, we use **MNIST** (LeCun et al., 1998), **Galaxy10 DECaLS** (Dey et al., 2019; Willett et al., 2013), and **FFHQ** (Karras et al., 2019). For text, we consider a rewriting task on **ParaDetox** (Logacheva et al., 2022) (results in Appendix B.1). Further experiment details, including our treatment of categorical features, are provided in Appendix B.1.

Baselines. On tabular data, we compare against four representative robust CF methods (see Appendix B.2): T-Rex (Hamman et al., 2023), ROAR (Upadhyay et al., 2021), PROPLACE (Jiang

Table 1: Comparison with robust CF baselines on tabular datasets. We report Future Validity (FV, %, ↑) and recourse Proximity (↓), ± denotes the standard deviation. The *DiCE (Base)* and *Wachter (Base)* rows show the unrefined base CFs and are not included in the robust-baseline comparison. Bold marks the best FV, and the lowest proximity among methods with FV $\geq 99\%$.

	HELOC		Diabetes		German Credit		FICO	
Method	FV ↑	Prox. ↓	FV ↑	Prox. ↓	FV ↑	Prox. ↓	FV ↑	Prox. ↓
T-Rex	*93.0 ± 1.4*	*3.6 ± 0.7*	*100.0 ± 1.6*	*3.2 ± 0.3*	*100.0 ± 1.8*	*9.1 ± 0.8*	96.3 ± 1.2	4.7 ± 0.5
ROAR	87.9 ± 1.8	3.6 ± 0.9	81.7 ± 1.2	1.3 ± 0.6	79.8 ± 2.0	2.5 ± 0.2	*84.9 ± 1.5*	3.6 ± 0.7
PROPLACE	72.4 ± 1.3	4.1 ± 0.4	66.8 ± 2.1	2.7 ± 0.8	64.2 ± 1.5	4.3 ± 0.6	69.6 ± 1.1	2.9 ± 0.3
ElliCE	**100.0 ± 2.0**	2.5 ± 0.5	**100.0 ± 1.4**	2.5 ± 0.9	**100.0 ± 2.4**	10.8 ± 0.7	**100.0 ± 1.6**	6.2 ± 0.4
DiCE (Base)	92.3 ± 1.7	4.3 ± 0.8	95.0 ± 1.3	2.8 ± 0.5	88.3 ± 1.9	3.9 ± 1.0	94.9 ± 1.4	5.7 ± 0.6
GRACE–DiCE	**100.0 ± 1.1**	4.3 ± 0.3	99.9 ± 0.9	3.0 ± 0.7	99.9 ± 1.2	4.0 ± 0.4	**100.0 ± 0.8**	5.7 ± 0.9
Wachter (Base)	77.9 ± 1.5	1.0 ± 0.2	73.3 ± 1.2	1.1 ± 0.5	58.7 ± 1.8	2.0 ± 0.9	79.3 ± 1.3	1.1 ± 0.3
GRACE–Wachter	**100.0 ± 1.2**	**1.2 ± 0.6**	**100.0 ± 1.0**	**1.8 ± 0.4**	**100.0 ± 1.1**	**2.6 ± 0.8**	**100.0 ± 0.9**	**1.5 ± 0.5**

et al., 2024a), and ElliCE (Turbal et al., 2025). These methods are designed for tabular features and, to the best of our knowledge, have no implementations that transfer directly to image or text inputs, so we include them only on tabular benchmarks. A runtime comparison is reported in Appendix B.3.

Base CF generators. Unlike the baselines, GRACE wraps existing CF generators: we use the Wachter (Wachter et al., 2018) and DiCE (Mothilal et al., 2020) for tabular data and LeapFactual (Cao et al., 2025) for images.

Future shifts. We obtain 45 future models per dataset by fine-tuning the base classifier under mean, scale, and tail shifts, with three severity levels and five realizations for tabular datasets (See Appendix B.5 for more details for tabular and image datasets). CFs are generated before the shifts and remain fixed during evaluation. These are predefined tests not required to lie inside the Wasserstein uncertainty set used internally by GRACE.

Evaluation metrics. We report *Future Validity*, the fraction of CFs that retain the desired prediction after future model updates, and *Proximity*, measured as the average ℓ_2 distance between each factual instance and its robust CF. Together, they capture the commonly observed robustness-proximity trade-off (Turbal et al., 2025). For MNIST, we additionally report the nominal population loss $F_{\mathbb{Q}}$ and update-aware sensitivity $\|s_{\mathbb{Q}}\|_*$ (Section 3.4) to separate the effects of nominal refinement and the sensitivity penalty in Figure 4.

Experimental setup. For each dataset, we use disjoint splits for base-model training, calibration, and final evaluation. **(1) Base classifier and prescribed updates.** We first train the deployed classifier f_0 and generate base CFs for instances predicted unfavorable. GRACE's internal update model uses T gradient steps with learning rate η (e.g., $T = 10$ for HELOC and $T = 3$ for MNIST, both with $\eta = 10^{-2}$) Appendix B.7 shows that GRACE remains robust when the realized update horizon exceeds T. The main tabular experiments use a multilayer perceptron classifier; Appendix B.7 additionally reports tabular experiments with logistic regression, while the text experiment uses a logistic classifier on frozen sentence embeddings. Further update-protocol and model details are provided in Appendix B.1. **(2) Calibration and population refinement.** Each method's robustness parameters are selected on the calibration split to maximize Future Validity: the original method parameters $\tau_{\text{T-Rex}}$ for T-Rex, δ for ROAR and PROPLACE, ϵ for ElliCE, and (r, β_{geo}) for GRACE. **(3) Held-out future-model evaluation.** Final robustness is evaluated on a disjoint collection of future models obtained by retraining f_0 under held-out training-distribution shifts. All methods are evaluated on the same factual instances and future-model realizations, and their returned CFs fixed.

4.1 GRACE REFINES COUNTERFACTUAL ROBUSTNESS

Table 1 shows that GRACE raises the Future Validity of both base generators to near 100% on all four datasets, matching or exceeding every robust baseline, with only a small increase in proximity over the base CFs, e.g., Wachter on German Credit: FV from 58.7% to 100%, proximity from 2.0 to 2.6. Importantly, the test shifts are not tailored to the uncertainty model of any particular method: future classifiers are induced by heterogeneous data-level shifts that are held out from model selection and

need not lie inside GRACE's Wasserstein uncertainty set. This can expose a mismatch for methods that assume a particular form of future model change, such as bounded parameter- or predictor-space perturbations, whereas GRACE models uncertainty at the training-distribution level and propagates it through the prescribed update procedure.

4.2 ROBUSTNESS UNDER SHIFT SEVERITY AND GEOMETRY BUDGET

Effect of Future Shift Severity. Our theory in Section 3.3 bounds how the population robust risk grows with the data-side uncertainty radius r increases. In evaluation, however, we deliberately avoid identifying test severity with r, since doing so would couple the benchmark to GRACE's own uncertainty model. Instead, we evaluate all methods on the same held-out future-shift scenarios and vary their realized severity directly. Figure 3 shows Future Validity on HELOC as the realized shift severity increases. GRACE maintains equal or higher Future Validity across the full range of severities, and degrades more gradually than the competing methods, e.g., the best competitor ElliCE drops sharply from S3 to S4, losing about 20%, whereas GRACE loses about 8%.

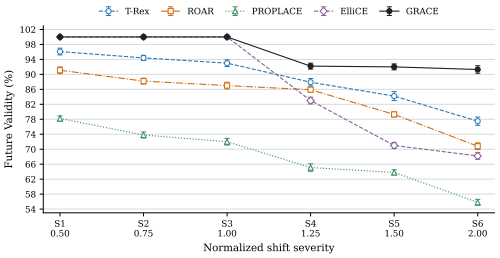


Figure 3: Future Validity under increasing future-shift severity on HELOC.

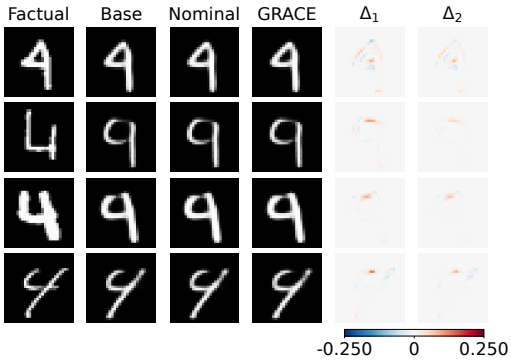
Effect of the Geometry Budget. We next study how the CF-side geometry budget controls the trade-off between robustness and proximity (calibration of β_{geo} and r in Appendix B.6). Increasing β_{geo} enlarges the bounded-detour region, giving GRACE more freedom to move the refined population away from the base CF distribution toward lower-risk regions. This constraint limits individual proximity only indirectly, since it bounds population-level transport rather than each factual-to-CF ℓ_2 distance. On HELOC (Table 2), increasing β_{geo} from 1.00 to 1.10 lowers the surrogate risk $\widetilde{\Phi}_r$ from 0.48 to 0.31 and raises FV from 81.8% to 100%, showing that the shortest geodesic ($\beta_{\text{geo}} = 1$) can be overly restrictive for robustness. This gain costs only a small increase in proximity (0.9 to 1.2). Beyond $\beta_{\text{geo}} = 1.10$, all metrics are unchanged; we leave a finer-grained characterization of this saturation behavior to future work. Additional diagnostics in Appendix B.7 show that removing the geometry constraint provides little additional FV while increasing proximity and transport detour.

Table 2: Effect of the geometry budget on GRACE for HELOC.

β_{geo}	FV ↑	Prox. ↓	$\widetilde{\Phi}_r$ ↓
1.00	81.8	**0.9**	0.48
1.05	83.4	1.0	0.42
1.10	**100.0**	1.2	**0.31**
1.20	**100.0**	1.2	**0.31**

4.3 VISUALIZING ROBUST REFINEMENT

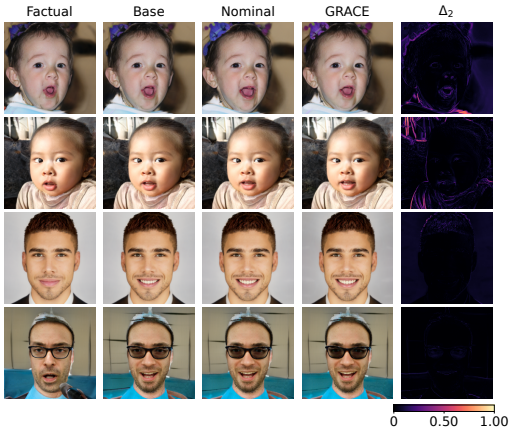
In Figure 4, we examine how GRACE modifies image CFs by comparing the factual input, the base CF generated by LeapFactual, and the robust refinement produced by GRACE. We additionally visualize the incremental changes from the base CF to the refined solution using difference heatmaps. For MNIST, we also report the nominal population loss $F_{\mathbb{Q}}(\theta_T^{\text{nom}})$ and the update-aware sensitivity $\|s_{\mathbb{Q}}\|_*$ to quantify the contributions of nominal refinement and robustness regularization. Across MNIST, Galaxy10 DECaLS, and FFHQ, the refined CFs retain recognizable instance-specific visual structure. Although our theory does not provide a formal guarantee of semantic or identity preservation, this behavior is consistent with the role of the geometric constraint in limiting excessive departure from the Base CF population. On MNIST, the refinement is often concentrated around the upper connecting stroke, a visually important region for distinguishing a "4" from a "9". On FFHQ and Galaxy10, the refinements are smaller and less localized, with FFHQ changes concentrated along facial contours. Understanding why visual domains induce such different refinement patterns is an interesting direction for future work. See Appendix B.8 for additional quantitative results on Galaxy10 and FFHQ, as well as enlarged qualitative examples that more clearly visualize the refinements.



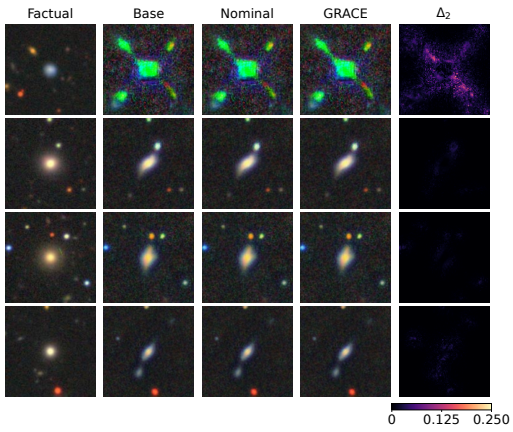
(a) MNIST ($4 \to 9$)

	FV ↑	$F_{\mathbb{Q}}$ ↓	$\|s_{\mathbb{Q}}\|_*$ ↓	Prox. ↓
Base	77.4	4.32×10^{-2}	3.63	5.35
$r = 0$	87.9	1.23×10^{-4}	1.61	5.41
$r = .05$	92.9	3.53×10^{-5}	0.0022	5.43
$r = .10$	99.8	2.43×10^{-5}	0.0014	5.44
$r = .25$	100.0	1.29×10^{-5}	0.0006	5.47

(b) MNIST analysis



(c) FFHQ (Face → Smiling Face)



(d) Galaxy10 (Round → Cigar)

Figure 4: Qualitative and quantitative analysis of GRACE on image datasets. in (a), heatmaps show signed pixel differences $\Delta_1 = \text{Nominal} - \text{Base}$ and $\Delta_2 = \text{GRACE} - \text{Nominal}$, where Nominal corresponds to $r = 0$; (c-d) show the magnitude $|\Delta_2|$. Color scales differ across panels. **(a)** On MNIST, GRACE focuses refinement on the upper connecting stroke that distinguishes "4" from "9". **(b)** The decrease in both nominal loss and update-aware sensitivity suggests that GRACE further moves the CF population along directions that are not only favorable under the nominal updated model but also less sensitive, with only a small increase in pixel-space proximity cost, measured by the factual-to-CF ℓ_2 distance. **(c–d)** On FFHQ and Galaxy10, refinements are small and retain recognizable instance-specific structure; on FFHQ, changes concentrate along facial contours. The artifact in the first Galaxy10 row originates in the base CF.

5 Discussion, Limitations, and Conclusion

We introduced GRACE, a post-hoc framework that makes a population of CFs robust to future model updates induced by training-distribution shift. Its population robust risk degrades at most linearly as data-side uncertainty grows, under the stated regularity conditions, and made the refinement tractable through an update-aware first-order sensitivity objective that propagates training-distribution perturbations through the prescribed retraining procedure. Across our experiments, GRACE improves Future Validity while maintaining competitive recourse proximity, and remains effective under held-out future shifts that are not tailored to its internal uncertainty model. Together, these results support population-level, data-induced model uncertainty as a useful perspective on robust CFs. GRACE does not enforce further CF requirements such as actionability or domain-specific feasibility. These could potentially enter the Wasserstein ground cost, at some expense of robustness and proximity. Our text study is also currently limited to a single detoxification task, calling for broader validation across language domains in the future.

AI USE STATEMENT

We did not use generative AI tools for any of the tasks requiring mandatory disclosure. However, we did use them to polish and edit the text of the paper for readability and to help identify related work we might have missed, both of which fall under recommended disclosure categories. All suggested edits were reviewed and approved to ensure that they reflect our intended meaning, and all suggested related works were manually verified before inclusion in the paper. We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, we provide an anonymous code repository, including data preprocessing pipeline and experiment scripts at `https://anonymous.4open.science/r/robust_cf-6E18`. Details of our experiment setup, including parameter choices or compute resources are provided in the Appendix B.

REFERENCES

Emily Black, Klas Leino, and Matt Fredrikson. Selective ensembles for consistent predictions. In *International Conference on Learning Representations (ICLR)*, 2022. URL `https://openreview.net/forum?id=HfUyCRBeQc`.

Jose Blanchet, Jiayi Cheng, Hao Liu, and Yang Liu. Bayesian distributionally robust merton problem with nonlinear wasserstein projections. *arXiv preprint arXiv:2512.01408*, 2025.

Jose H. Blanchet, Yang Kang, and Karthyek Rajhaa A. M. Robust wasserstein profile inference and applications to machine learning. *J. Appl. Probab.*, 56(3):830–857, 2019. doi: 10.1017/JPR.2019.49. URL `https://doi.org/10.1017/jpr.2019.49`.

Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001. doi: 10.1214/ss/1009213726.

Ngoc Bui, Duy Nguyen, and Viet Anh Nguyen. Counterfactual plans under distributional ambiguity. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=noaG7SrPVK0`.

Ngoc Bui, Duy Nguyen, Man-Chung Yue, and Viet Anh Nguyen. Coverage-validity-aware algorithmic recourse. *Oper. Res.*, 73(6):3294–3310, 2025. doi: 10.1287/OPRE.2023.0629. URL `https://doi.org/10.1287/opre.2023.0629`.

Zhuo Cao, Xuan Zhao, Lena Krieger, Hanno Scharr, and Ira Assent. Leapfactual: Reliable visual counterfactual explanation using conditional flow matching. In *NeurIPS*, 2025.

Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. Characterizing fairness over the set of good models under selective labels. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2144–2155, 2021.

Arjun Dey, David J Schlegel, Dustin Lang, Robert Blum, Kaylan Burleigh, Xiaohui Fan, Joseph R Findlay, Doug Finkbeiner, David Herrera, Stéphanie Juneau, et al. Overview of the DESI Legacy Imaging Surveys. *The Astronomical Journal*, 157(5):168, 2019.

Luke Ditria and Tom Drummond. Long-term prediction of natural video sequences with robust video predictors. *CoRR*, abs/2308.11079, 2023.

Jamie Duell and Xiuyi Fan. Provably robust bayesian counterfactual explanations under model changes. *arXiv preprint arXiv:2601.16659*, 2026.

Sanghamitra Dutta, Jason Long, Saumitra Mishra, Cecilia Tilli, and Daniele Magazzeni. Robust counterfactual explanations for tree-based ensembles. In *International conference on machine learning*, pp. 5742–5756. PMLR, 2022.

Fair Isaac Corporation (FICO). FICO Explainable Machine Learning Challenge: Home Equity Line of Credit (HELOC) Dataset. FICO Community, 2018.

Andrea Ferrario and Michele Loi. The robustness of counterfactual explanations over time. *IEEE access*, 10:82736–82750, 2022.

Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1165–1173. PMLR, 2017.

Rui Gao and Anton J. Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *Math. Oper. Res.*, 48(2):603–655, 2023. doi: 10.1287/MOOR.2022.1275. URL https://doi.org/10.1287/moor.2022.1275.

Faisal Hamman, Erfaun Noorani, Saumitra Mishra, Daniele Magazzeni, and Sanghamitra Dutta. Robust counterfactual explanations for neural networks with probabilistic guarantees. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12351–12367. PMLR, 2023.

Hans Hofmann. Statlog (german credit data). UCI Machine Learning Repository, 1994.

Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Formalising the robustness of counterfactual explanations for neural networks. In *AAAI*, pp. 14901–14909. AAAI Press, 2023.

Junqi Jiang, Jianglin Lan, Francesco Leofante, Antonio Rago, and Francesca Toni. Provably robust and plausible counterfactual explanations for neural networks via robust optimisation. In Berrin Yanikoglu and Wray L. Buntine (eds.), *Asian Conference on Machine Learning, ACML 2023, 11-14 November 2023, Istanbul, Turkey*, volume 222 of *Proceedings of Machine Learning Research*, pp. 582–597. PMLR, 2024a. URL https://proceedings.mlr.press/v222/jiang24a.html.

Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Robust counterfactual explanations in machine learning: A survey. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 8086–8094. International Joint Conferences on Artificial Intelligence Organization, 8 2024b. doi: 10.24963/ijcai.2024/894. URL https://doi.org/10.24963/ijcai.2024/894. Survey Track.

Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Interval abstractions for robust counterfactual explanations. *Artificial Intelligence*, 336:104218, 2024c.

Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Kshitij Kayastha, Vasilis Gkatzelis, and Shahin Jabbari. Learning-augmented robust algorithmic recourse. *arXiv preprint arXiv:2410.01580*, 2024.

Phone Kyaw, Kshitij Kayastha, and Shahin Jabbari. Optimal robust recourse with l^p-bounded model change. *arXiv preprint arXiv:2509.21293*, 2025.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. Paradetox: Detoxification with parallel data. In *ACL (1)*, pp. 6804–6818. Association for Computational Linguistics, 2022.

Charles T. Marx, Flavio P. Calmon, and Berk Ustun. Predictive multiplicity in classification. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 6765–6774, 2020.

Luca Marzari, Francesco Leofante, Ferdinando Cicalese, and Alessandro Farinelli. Rigorous probabilistic guarantees for robust counterfactual explanations. In *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI)*, pp. 1059–1066, 2024.

Luca Marzari, Francesco Leofante, Ferdinando Cicalese, and Alessandro Farinelli. Probabilistically robust counterfactual explanations under model changes. *Artif. Intell.*, 351:104459, 2026.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *FAT**, pp. 607–617. ACM, 2020.

Duy Nguyen, Ngoc Bui, and Viet Anh Nguyen. Distributionally robust recourse action. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=E3ip6qBLF7.

Tuan-Duy H Nguyen, Ngoc Bui, Duy Nguyen, Man-Chung Yue, and Viet Anh Nguyen. Robust bayesian recourse. In *Uncertainty in Artificial Intelligence*, pp. 1498–1508. PMLR, 2022.

Averkiy Oliabev. Home equity line of credit (heloc). Kaggle Dataset. URL https://www.kaggle.com/datasets/averkiyoliabev/home-equity-line-of-creditheloc. Accessed: 2026-09-25.

Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. On counterfactual explanations under predictive multiplicity. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pp. 809–818, 2020.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2009.

Kaivalya Rawal, Ece Kamar, and Himabindu Lakkaraju. Algorithmic recourse in the wild: Understanding the impact of data and model shifts. *arXiv preprint arXiv:2012.11788*, 2020.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 2019. doi: 10.1038/s42256-019-0048-x.

Amirreza Shaban, Ching-An Cheng, Nathan Hatch, and Byron Boots. Truncated back-propagation for bilevel optimization. In *AISTATS*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1723–1732. PMLR, 2019.

Jack W. Smith, James E. Everhart, William C. Dickson, William C. Knowler, and Richard S. Johannes. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pp. 261–265. American Medical Informatics Association, 1988.

Ignacy Stepka, Jerzy Stefanowski, and Mateusz Lango. Counterfactual explanations with probabilistic guarantees on their robustness to model change. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 1277–1288, 2025.

Bohdan Turbal, Iryna Voitsitska, and Lesia Semenova. Ellice: Efficient and provably robust algorithmic recourse via the rashomon sets. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4128.

Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. Towards robust and reliable algorithmic recourse. In *NeurIPS*, pp. 16926–16937, 2021.

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.

Kyle W Willett, Chris J Lintott, Steven P Bamford, Karen L Masters, Brooke D Simmons, Kevin RV Casteels, Edward M Edmondson, Lucy F Fortson, Sugata Kaviraj, William C Keel, et al. Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 435(4):2835–2860, 2013.

Ao Xu and Tieru Wu. Generally-occurring model change for robust counterfactual explanations. In *International conference on artificial neural networks*, pp. 215–229. Springer, 2024.

Xuan Zhao, Lena Krieger, Zhuo Cao, Arya Bangun, Hanno Scharr, and Ira Assent. RECAST: Model reconstruction via counterfactual-aware wasserstein geometry under limited data. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026.

A Proofs for Geometry-Constrained Refinement

A.1 Relation between margin risk and threshold validity

For any $\gamma > 0$ and any classifier f,

$$\mathbf{1}\{f(z) \le \tau - \gamma\} \le \frac{[\tau - f(z)]_+}{\gamma}.$$

Therefore, for any counterfactual population $\mathbb{Q}$,

$$\sup_{\mathbb{P}\in\mathcal{U}_r} \Pr_{z\sim\mathbb{Q}}\left(f_{\mathbb{P}}(z) \le \tau - \gamma\right) \le \frac{\Phi_r(\mathbb{Q})}{\gamma}.$$

Hence minimizing Φ_r controls the fraction of CFs that violate the decision threshold by at least margin γ, while the exact threshold-based Future Validity is reported separately in our experiments. The pointwise inequality follows because $f(z) \le \tau - \gamma$ implies $[\tau - f(z)]_+ \ge \gamma$. Taking expectations and then the supremum over $\mathbb{P} \in \mathcal{U}_r$ gives the result. For numerical optimization, we replace the hinge loss by the smooth upper bound $\ell_f^{\mathrm{sm}}(z) = \kappa \log(1 + \exp((\tau - f(z))/\kappa))$, which provides the differentiability required by the sensitivity computation while preserving the margin-risk control above.

A.2 Proof of Proposition 1

Recall that, for a training distribution $\mathbb{P}$, the training objective is

$$L_{\mathbb{P}}(\theta) := \mathbb{E}_{\xi\sim\mathbb{P}}[\ell_{\mathrm{tr}}(\theta;\xi)],$$

and the prescribed T-step update is

$$\theta_{t+1}^{\mathbb{P}} = \theta_t^{\mathbb{P}} - \eta_t \nabla_\theta L_{\mathbb{P}}(\theta_t^{\mathbb{P}}), \qquad \theta_0^{\mathbb{P}} = \theta_0,$$

with $f_{\mathbb{P}} = f_{\theta_T^{\mathbb{P}}}$.

We first establish stability of this finite-horizon update with respect to the training distribution. Fix two admissible distributions $\mathbb{P}$ and $\mathbb{P}'$ and define

$$\Delta_t := \|\theta_t^{\mathbb{P}} - \theta_t^{\mathbb{P}'}\|.$$

Since both trajectories start from the same deployed model, $\Delta_0 = 0$. From the update rule and the triangle inequality,

$$\begin{aligned}\Delta_{t+1} &= \left\|\theta_t^{\mathbb{P}} - \eta_t \nabla L_{\mathbb{P}}(\theta_t^{\mathbb{P}}) - \theta_t^{\mathbb{P}'} + \eta_t \nabla L_{\mathbb{P}'}(\theta_t^{\mathbb{P}'})\right\| \\ &\le \Delta_t + \eta_t \left\|\nabla L_{\mathbb{P}}(\theta_t^{\mathbb{P}}) - \nabla L_{\mathbb{P}'}(\theta_t^{\mathbb{P}'})\right\|.\end{aligned}$$

Adding and subtracting $\nabla L_{\mathbb{P}}(\theta_t^{\mathbb{P}'})$ and applying the two regularity assumptions gives

$$\begin{aligned}\Delta_{t+1} &\le \Delta_t + \eta_t \left\|\nabla L_{\mathbb{P}}(\theta_t^{\mathbb{P}}) - \nabla L_{\mathbb{P}}(\theta_t^{\mathbb{P}'})\right\| \\ &\quad + \eta_t \left\|\nabla L_{\mathbb{P}}(\theta_t^{\mathbb{P}'}) - \nabla L_{\mathbb{P}'}(\theta_t^{\mathbb{P}'})\right\| \\ &\le (1 + \eta_t L_\theta)\Delta_t + \eta_t L_{\mathrm{data}} W_{\mathrm{data}}(\mathbb{P}, \mathbb{P}').\end{aligned}$$

Unrolling this recursion from $\Delta_0 = 0$ yields

$$\|\theta_T^{\mathbb{P}} - \theta_T^{\mathbb{P}'}\| \le C_T W_{\mathrm{data}}(\mathbb{P}, \mathbb{P}'),$$

where

$$C_T := L_{\mathrm{data}} \sum_{t=0}^{T-1} \eta_t \prod_{s=t+1}^{T-1} (1 + \eta_s L_\theta).$$

By the assumed Lipschitz continuity of the predictor with respect to its parameters,

$$\sup_{z \in \mathcal{X}} |f_{\mathbb{P}}(z) - f_{\mathbb{P}'}(z)| = \sup_{z \in \mathcal{X}} |f_{\theta_T^{\mathbb{P}}}(z) - f_{\theta_T^{\mathbb{P}'}}(z)| \le L_f C_T W_{\mathrm{data}}(\mathbb{P}, \mathbb{P}'). \tag{A.1}$$

We now establish graceful degradation. Fix $0 \le r_1 \le r_2$ and an arbitrary $\mathbb{P} \in \mathcal{U}_{r_2}$. Since the data-side Wasserstein space is geodesic, choose $\mathbb{P}'$ on a shortest Wasserstein geodesic from $\widehat{\mathbb{P}}_{\mathrm{tr}}$ to $\mathbb{P}$ such that

$$W_{\mathrm{data}}(\widehat{\mathbb{P}}_{\mathrm{tr}}, \mathbb{P}') = \min\{r_1, W_{\mathrm{data}}(\widehat{\mathbb{P}}_{\mathrm{tr}}, \mathbb{P})\}.$$

Then $\mathbb{P}' \in \mathcal{U}_{r_1}$ and

$$W_{\mathrm{data}}(\mathbb{P}, \mathbb{P}') \le r_2 - r_1.$$

Applying (A.1) therefore gives

$$\sup_{z \in \mathcal{X}} |f_{\mathbb{P}}(z) - f_{\mathbb{P}'}(z)| \le L_f C_T (r_2 - r_1).$$

Since the margin loss $\ell_f(z) = [\tau - f(z)]_+$ is 1-Lipschitz in the prediction $f(z)$, for every $z \in \mathcal{X}$,

$$\ell_{f_{\mathbb{P}}}(z) \le \ell_{f_{\mathbb{P}'}}(z) + L_f C_T (r_2 - r_1).$$

Taking expectation over $z \sim \mathbb{Q}$ yields

$$\mathbb{E}_{z \sim \mathbb{Q}}[\ell_{f_{\mathbb{P}}}(z)] \le \mathbb{E}_{z \sim \mathbb{Q}}[\ell_{f_{\mathbb{P}'}}(z)] + L_f C_T (r_2 - r_1).$$

Because $\mathbb{P}' \in \mathcal{U}_{r_1}$,

$$\mathbb{E}_{z \sim \mathbb{Q}}[\ell_{f_{\mathbb{P}'}}(z)] \le \Phi_{r_1}(\mathbb{Q}).$$

Hence, for every $\mathbb{P} \in \mathcal{U}_{r_2}$,

$$\mathbb{E}_{z \sim \mathbb{Q}}[\ell_{f_{\mathbb{P}}}(z)] \le \Phi_{r_1}(\mathbb{Q}) + L_f C_T (r_2 - r_1).$$

Taking the supremum over $\mathbb{P} \in \mathcal{U}_{r_2}$ gives

$$\Phi_{r_2}(\mathbb{Q}) \le \Phi_{r_1}(\mathbb{Q}) + L_f C_T (r_2 - r_1).$$

Since $\mathcal{U}_{r_1} \subseteq \mathcal{U}_{r_2}$, we also have $\Phi_{r_1}(\mathbb{Q}) \le \Phi_{r_2}(\mathbb{Q})$. Therefore,

$$\boxed{0 \le \Phi_{r_2}(\mathbb{Q}) - \Phi_{r_1}(\mathbb{Q}) \le L_f C_T (r_2 - r_1).}$$

This proves the proposition.

A.3 First-order Wasserstein sensitivity

For a fixed CF population $\mathbb{Q}$, define

$$\Psi_{\mathbb{Q}}(\mathbb{P}) := F_{\mathbb{Q}}(\theta_T^{\mathbb{P}}),$$

so that

$$\Phi_r(\mathbb{Q}) = \sup_{W_{\mathrm{data}}(\mathbb{P}, \widehat{\mathbb{P}}_{\mathrm{tr}}) \le r} \Psi_{\mathbb{Q}}(\mathbb{P}).$$

Proposition 3 (Update-aware first-order sensitivity)**.** *Fix $\mathbb{Q}$. Suppose that, in a neighborhood of $\widehat{\mathbb{P}}_{\mathrm{tr}}$, the functional $\Psi_{\mathbb{Q}}$ admits a first variation $\Psi'_{\mathbb{Q},\mathbb{P}}(\xi)$ whose covariate gradient*

$$g_{\mathbb{Q},\mathbb{P}}(\xi) := \nabla_u \Psi'_{\mathbb{Q},\mathbb{P}}(\xi)$$

satisfies the continuity, growth, and integrability conditions for nonlinear Wasserstein sensitivity (Blanchet et al., 2025, Theorem 3). Assume these conditions hold uniformly over feasible $\mathbb{Q}$. Then, as $r \downarrow 0$,

$$\Phi_r(\mathbb{Q}) = F_{\mathbb{Q}}(\theta_T^{\mathrm{nom}}) + r \|g_{\mathbb{Q},\widehat{\mathbb{P}}_{\mathrm{tr}}}\|_{L^2(\widehat{\mathbb{P}}_{\mathrm{tr}})} + o(r).$$

Moreover, an asymptotically worst-case first-order perturbation can be chosen as a deterministic pushforward of $\widehat{\mathbb{P}}_{\mathrm{tr}}$.

For the prescribed differentiable T*-step update, the same first-order functional is represented in transport coordinates by*

$$s_{\mathbb{Q}} = J_T^{\top} \nabla_\theta F_{\mathbb{Q}}(\theta_T^{\mathrm{nom}}),$$

and

$$\|s_{\mathbb{Q}}\|_{\mathcal{H}^*} = \|g_{\mathbb{Q},\widehat{\mathbb{P}}_{\mathrm{tr}}}\|_{L^2(\widehat{\mathbb{P}}_{\mathrm{tr}})}.$$

Hence

$$\Phi_r(\mathbb{Q}) = \widetilde{\Phi}_r(\mathbb{Q}) + o(r).$$

Proof. The first expansion follows by applying the nonlinear-functional Wasserstein sensitivity result of Blanchet et al. (2025, Theorem 3) to $\Psi_{\mathbb{Q}}$. Although the supremum is taken over the full Wasserstein ball, that result also provides an asymptotically optimal Monge-type pushforward, so restricting to deterministic transport directions is sufficient for identifying the first-order term.

It therefore remains to identify this first-order derivative for the prescribed update. For a transport direction h,

$$\theta_T(h) = \theta_T^{\mathrm{nom}} + J_T h + o(\|h\|_{\mathcal{H}}).$$

Differentiability of $F_{\mathbb{Q}}$ gives

$$D_h \left[F_{\mathbb{Q}}(\theta_T(h))\right]_{h=0}[h] = \left\langle J_T^{\top} \nabla_\theta F_{\mathbb{Q}}(\theta_T^{\mathrm{nom}}), h \right\rangle.$$

Thus the Riesz representative of the Wasserstein first variation in the empirical transport space is

$$s_{\mathbb{Q}} = J_T^{\top} \nabla_\theta F_{\mathbb{Q}}(\theta_T^{\mathrm{nom}}),$$

whose dual norm equals the corresponding $L^2(\widehat{\mathbb{P}}_{\mathrm{tr}})$ Wasserstein-gradient norm. Substitution into the full-ball sensitivity expansion proves the claim. □

A.4 Hypergradient Computation

For completeness, let $J_t := D_h \theta_t(h)|_{h=0}$ and $\theta_t^0 := \theta_t(0)$. Differentiating the prescribed update

$$\theta_{t+1}(h) = \theta_t(h) - \eta_t \nabla_\theta L_{\mathbb{P}_h}(\theta_t(h))$$

with respect to h gives

$$J_{t+1} = (I - \eta_t H_t) J_t - \eta_t B_t, \qquad J_0 = 0,$$

where

$$H_t := \nabla^2_{\theta\theta} L_{\widehat{\mathbb{P}}_{\mathrm{tr}}}(\theta_t^0), \qquad B_t := D_h \nabla_\theta L_{\mathbb{P}_h}(\theta_t^0)\big|_{h=0}.$$

Hence Hence

$$s_{\mathbb{Q}} = J_T^{\top} \nabla_\theta F_{\mathbb{Q}}(\theta_T^{\mathrm{nom}}).$$

In implementation we do not form J_T explicitly. Instead, we differentiate $F_{\mathbb{Q}}(\theta_T(h))$ with respect to h by reverse-mode automatic differentiation through the T update steps. This requires vector–Jacobian and Hessian–vector products rather than storing the full Jacobian.

A.5 Mixed Continuous–Categorical Inputs

For mixed tabular inputs, the differentiable numerical realization is applied in a continuous relaxation of the encoded feature space. Continuous coordinates are perturbed additively as above. For a categorical variable with K categories, let $\alpha_i \in \mathbb{R}^K$ denote nominal logits and define the relaxed categorical representation

$$p_i(\delta_i) = \mathrm{softmax}\left(\frac{\alpha_i + \delta_i}{\tau_{\mathrm{cat}}}\right),$$

where $\tau_{\mathrm{cat}} > 0$ is a fixed temperature and δ_i is the categorical perturbation variable. The complete differentiable perturbation is then written $u = (h^{\mathrm{cont}}, \delta^{\mathrm{cat}})$, and the update-aware sensitivity is computed as

$$s_{\mathbb{Q}} = \nabla_u F_{\mathbb{Q}}(\theta_T(u))\big|_{u=0}.$$

The softmax parameterization keeps relaxed categorical coordinates on the probability simplex and allows differentiation through the future-update procedure. It is, however, a relaxation of discrete categorical changes rather than an exact parameterization of hard category transitions. Accordingly, the local smoothness result above applies to the relaxed representation; robustness to hard categorical changes is evaluated separately in the experiments.

A.6 Optimization Algorithm

Algorithm 1 summarizes the numerical optimization and deployment procedure for GRACE.

B Experiment Details

B.1 Dataset and Experiment Details

We provide additional details on the datasets, preprocessing, and modality-specific constructions used in our experiments. Across all modalities, the data used to fit the reference predictor are kept separate from the calibration and final evaluation populations. Preprocessing statistics and representation models are fixed before calibration and test evaluation.

Tabular datasets. We consider four binary-classification benchmarks: HELOC, German Credit, FICO, and Diabetes. Continuous features are normalized using statistics computed on the corresponding training split, and the same normalized representation is used for classifier training, counterfactual optimization, Wasserstein geometry, and recourse-cost computation. Categorical variables, when present, are encoded before model fitting; their differentiable treatment within GRACE follows Appendix A.5. HELOC and FICO provide continuous and categorical variants of the same underlying benchmark, respectively, and we observe comparable robustness gains across both representations.

HELOC. The Home Equity Line of Credit (HELOC) benchmark contains credit-risk records described by credit-history and account-level attributes, with the prediction task distinguishing favorable from unfavorable credit risk (Oliabev). We remove non-predictive bookkeeping variables and normalize the resulting numerical representation using training-set statistics. For the experiments reported in the main text, the classifier and counterfactual procedure operate on this normalized feature space.

German Credit. The German Credit dataset contains 1,000 credit applications described by 20 attributes, including both numerical and categorical variables, with labels corresponding to good or bad credit risk (Hofmann, 1994). We encode categorical attributes and standardize continuous coordinates using training-split statistics before fitting the reference classifier. The same transformed representation is used for counterfactual refinement.

FICO. We use the FICO Explainable Machine Learning Challenge credit-risk benchmark (Fair Isaac Corporation (FICO), 2018). The task predicts whether an individual belongs to the favorable or unfavorable credit-risk class from credit-history variables. The features are converted to a numerical representation and normalized using statistics from the training split before reference-model fitting and counterfactual generation. In our experiments, HELOC refers to the continuous numerical representation of this benchmark while FICO refers to the original one with categorical features.

Diabetes. We use the Pima Indians Diabetes benchmark (Smith et al., 1988), consisting of 768 observations with eight numerical predictive attributes. These include measurements related to glucose, blood pressure, body-mass index, insulin, age, pregnancy count, and diabetes history. The binary target indicates diabetes status. All input variables are standardized using the training split before classifier fitting and counterfactual optimization.

For each tabular dataset, we first fit the reference classifier f_0 on the training split. Factual instances predicted as unfavorable are then used to construct the counterfactual population. The Base CFs form the proximal anchor, while favorable examples from the training data define the target anchor. For each candidate (r, β_{geo}), GRACE independently refines the calibration CF population and evaluates the resulting CFs under calibration-only future model updates. The selected operating pa-

rameters are frozen before test evaluation. On the test split, a new Base CF population is generated and GRACE is rerun using the frozen configuration; no calibration CFs are reused at test time.

Image datasets. We study GRACE on MNIST, Galaxy10 DECaLS, and FFHQ. For the image experiments, we build on the representation and counterfactual-generation pipeline of LeapFactual (Cao et al., 2025). LeapFactual first maps an input image to a latent representation using either a VAE encoder or GAN inversion, transports the latent representation toward a target class using a conditional flow-matching model, and finally decodes the transported latent point back to image space. We use the resulting LeapFactual CFs as the Base population for GRACE and subsequently refine this population in the same latent representation. The image encoder/generator and Base-CF generation procedure are fixed before GRACE optimization. We use the standard LeapFactual output after the blending stage, without the subsequent reliability-injection stage, as the Base CF.

MNIST. MNIST consists of 28×28 grayscale images of handwritten digits (LeCun et al., 1998). The image representation is learned with a convolutional VAE whose encoder outputs the parameters of a 64-dimensional latent variable and whose mirrored decoder reconstructs the image. The VAE is trained with a weighted combination of reconstruction MSE and KL regularization. The CNN digit classifier achieves approximately 96.6% test accuracy. We consider the $4 \to 9$ task. Each factual image is encoded into the VAE latent space, LeapFactual performs its blending procedure until the target class is reached, and the decoded output is used as the Base CF. GRACE then refines the resulting Base-CF population directly in the same latent space, with all validity calculations performed after decoding the latent points back to image space. GRACE refines the Base-CF population in the frozen decoder's 64-dimensional latent space, retaining the actual latent codes used to generate Base rather than re-encoding the decoded images. Each population contains $M = n = 64$ equally weighted support points, with 64 favorable reference points defining the second geometric anchor. Classifier predictions and counterfactual losses are computed on decoded images, while Wasserstein distances and the final one-to-one transport assignment are computed in latent space.

The prescribed classifier update consists of $T = 3$ full-batch gradient-descent steps with learning rate $\eta = 0.01$, using binary cross-entropy on a fixed set of 256 training images. Batch-normalization statistics remain frozen, and dropout is disabled. GRACE differentiates through all three update steps to compute sensitivity to perturbations of the training images in pixel space. Thus, the training-distribution perturbations and the optimized counterfactual supports occupy different spaces. The nominal counterfactual loss and sensitivity term both use the smooth probability-margin loss with $\kappa = 0.05$.

We optimize the support population for 800 augmented-Lagrangian iterations, using a primal step size of 20 and penalty parameter $\rho = 1$. This support-optimization step size is distinct from the classifier-update learning rate η. We fix $\beta_{\text{geo}} = 1.02$ and consider $r \in \{0, 0.05, 0.10, 0.25\}$; calibration selects $r = 0.25$. The matched $r = 0$ configuration provides a nominal-only refinement control under the same geometry budget and optimization settings. We independently verify the final geometric constraint with relative tolerance 10^{-5} and deploy the refined supports through a one-to-one optimal transport assignment from Base. No post-hoc validity correction is applied. The fixed iteration budget does not imply convergence, and the evaluation concerns the prescribed three-step update procedure.

GRACE can work with multiclass classifiers by defining its loss using the probability of a chosen target class.

Galaxy10 DECaLS. Galaxy10 DECaLS contains 17,736 RGB galaxy images at 256×256 resolution, grouped into ten morphological classes using Galaxy Zoo annotations and imaging from the DESI Legacy Surveys (Dey et al., 2019; Willett et al., 2013). We do the following preprocessing: images are center-cropped to 150×150, and resized to 128×128. We use a VAE with a 64-dimensional latent representation; its architecture is adapted from the VAE of Ditria & Drummond (2023), replacing ELU activations with LeakyReLU and using maximum mean discrepancy (MMD) rather than KL divergence for latent regularization.

The associated image classifier is VGG16. The accuracy of the classifier is approximately 85%. We operate on the 64-dimensional VAE latent representation. For our round-to-cigar task, the factual image is encoded by the VAE. We use the decoded standard LeapFactual counterfactual as the Base CF for GRACE. GRACE subsequently refines these Base supports in the same 64-dimensional latent

space, while classifier evaluation and qualitative visualization are performed on the decoded RGB images.

GRACE refines the Base-CF population in the same frozen 64-dimensional latent space, retaining the original generator codes rather than re-encoding the decoded Base images. Each population contains $M = n = 128$ equally weighted support points. Classifier predictions and counterfactual losses are computed on decoded RGB images, while the Wasserstein geometry is computed using Euclidean distances between latent codes.

The prescribed classifier update consists of $T = 3$ full-batch gradient-descent steps with learning rate $\eta = 10^{-4}$, using multiclass cross-entropy on a fixed class-balanced subset of 40 training images, with four images per class. Only the weights and biases of the final linear classification layer are updated; the preceding VGG16-BN layers and batch-normalization statistics remain frozen. GRACE differentiates through all three update steps. Sensitivity to training-image perturbations is computed in pixel space by pulling gradients back through the frozen feature extractor, rather than treating feature-space perturbations as image perturbations. We retain the original ten-class softmax and use the target-class probability in the smooth margin loss.

We refine the latent supports using 160 augmented-Lagrangian updates, with primal step size 5, penalty parameter $\rho = 1$, and gradient-norm clipping at 0.2. These support-optimization settings are separate from the classifier-update learning rate η. We consider $r \in \{0, 0.25, 1, 4\}$ and $\beta_{\text{geo}} \in \{1.02, 1.10\}$. For Round-to-Cigar, calibration selects $r = 0.25$ and $\beta_{\text{geo}} = 1.02$; the matched $r = 0$ configuration provides a nominal-only control. The geometry constraint bounds the sum of the Wasserstein distances from Base to the refined population and from the refined population to the favorable reference population. We retain the feasible iterate with the lowest surrogate objective, using feasibility tolerance $10^{-5} \max\{1, D\}$, where $D = W_2(\mathbb{P}_{\text{CF}}, \mathbb{P}_1)$. The returned population is independently checked and deployed through a one-to-one optimal transport assignment from Base. No post-hoc validity correction is applied. This experiment evaluates robustness to the prescribed three-step classification-head update, without claiming convergence or robustness to unrestricted retraining.

FFHQ. FFHQ contains 70,000 aligned and cropped face images at 1024×1024 resolution (Karras et al., 2019). A pretrained `stylegan3-r-ffhq-1024x1024` generator provides the image representation and decoder. Latent samples are drawn from the StyleGAN Gaussian input prior, mapped into StyleGAN3 w-space, and decoded to 1024×1024 face images. GRACE uses a pretrained CLIP ViT-B/32 model as the classifier, comparing the scores associated with the prompts "Face" and "Smiling Face"; the larger score defines the predicted class. We use a conditional 1D U-Net flow model on 20,000 random StyleGAN3 latent samples and their corresponding CLIP-predicted labels. For the Face→Smiling Face task, the source representation is transported in w-space toward the smiling class and decoded by StyleGAN3. We use this decoded LeapFactual output as the Base CF and apply GRACE to the corresponding latent population. The StyleGAN3 generator and image representation are kept fixed throughout GRACE refinement, and all visual comparisons are performed after decoding the optimized latent supports back to image space.

GRACE refines the original Base latent codes in the 512-dimensional StyleGAN3 w-space. Each population contains $M = n = 128$ equally weighted support points. Each optimized w vector is shared across the generator's 16 style inputs, and the generator remains frozen throughout refinement. Wasserstein geometry and the final transport assignment are computed in w-space, while classifier predictions are computed on decoded images after CLIP preprocessing. All 512 latent coordinates are optimized jointly, without additional identity or background preservation constraints.

The prescribed classifier update consists of $T = 3$ full-batch gradient-descent steps with learning rate $\eta = 10^{-4}$, using cross-entropy on 32 generated training images balanced across the two original CLIP-predicted classes. These labels remain fixed during adaptation. Only CLIP's 768×512 visual projection matrix is updated; the visual backbone, text encoder, text embeddings, and logit scale remain fixed. GRACE differentiates through all three update steps and computes sensitivity to perturbations of the 224×224 RGB inputs used for CLIP scoring. This sensitivity includes the pixel-space pullback through the frozen visual backbone, rather than measuring uncertainty only in its feature representation. For the nominal loss and sensitivity term, we use the smiling-class probability in smooth margin loss.

We optimize the latent supports for 160 augmented-Lagrangian updates, using primal step size 5, penalty parameter $\rho = 1$, and gradient-norm clipping at 0.2. The support-optimization step size is distinct from the classifier-update learning rate η. We consider $r \in \{0, 0.25, 1, 4\}$ and $\beta_{\text{geo}} \in \{1.02, 1.10\}$. Calibration selects $r = 0.25$ and $\beta_{\text{geo}} = 1.02$ under the recourse-cost tie-break; the $r = 0.25$ configuration at the same geometry budget is retained as a positive-radius comparison in the qualitative analysis. We return the feasible iterate with the lowest surrogate objective, using geometric tolerance $10^{-5} \max\{1, D\}$, where $D = W_2(\mathbb{P}_{\text{CF}}, \mathbb{P}_1)$. After independently checking feasibility, we assign the refined supports to Base through one-to-one optimal transport. No post-hoc validity correction is applied. The evaluation concerns short adaptation of CLIP's visual projection; it does not establish robustness to full-model retraining.

Text dataset. We additionally consider toxic-to-non-toxic rewriting using ParaDetox (Logacheva et al., 2022), an English parallel dataset containing toxic comments and human-written non-toxic rewrites. ParaDetox contains non-toxic paraphrases for more than 10,000 toxic English sentences and was designed specifically for text detoxification.

We fine-tune **FLAN-T5-small** to generate candidate non-toxic rewrites. For each toxic factual sentence, the generator produces a finite collection of natural-language candidates. Sentence representations are obtained using a frozen **MiniLM** encoder, and a logistic classifier trained on these sentence embeddings provides the toxicity prediction used to define validity and future-model updates.

Unlike image latents, arbitrary perturbations of sentence embeddings cannot in general be decoded reliably into grammatical natural-language sentences. We therefore do not optimize unrestricted continuous sentence embeddings. Instead, GRACE is adapted to text by restricting the population supports to the finite pool of actual generated sentences. The Base CF corresponds to a generated rewrite satisfying the target toxicity criterion, and GRACE selects alternative rewrites from the candidate pool according to the population objective and geometry constraint.

The nominal population loss and update-aware sensitivity are evaluated through the frozen sentence-embedding representation and the prescribed toxicity-classifier update. Thus, the population-level coupling of GRACE is unchanged, while the support optimization becomes a discrete selection problem. This experiment should consequently be interpreted as a discrete-text adaptation of GRACE rather than as certified robustness to arbitrary token- or word-level perturbations.

For text evaluation, we report Future Validity under held-out toxicity classifier updates together with semantic preservation and word-edit cost. Candidate-generation settings, the sentence encoder, and the classifier architecture are fixed before final evaluation, while GRACE-specific operating parameters are selected on the calibration split and frozen before evaluation on the held-out test population.

We fine-tune FLAN-T5-small on 2,000 training pairs for 300 AdamW steps to generate candidate non-toxic rewrites. Sentence representations are obtained using the frozen `all-MiniLM-L6-v2` encoder and normalized to unit length, yielding 384-dimensional embeddings. A logistic classifier fitted on these embeddings provides the non-toxic probability used for counterfactual scoring. The generator, sentence encoder, and initial classifier are fixed before counterfactual refinement.

For each factual sentence, we generate eight beam-search candidates and 64 sampled candidates, with 32 samples at each temperature 0.8 and 1.1. Generated alternatives with identical lowercased word sequences are deduplicated, while the original Base and reference anchors are retained. Base is constructed by selecting the nearest classifier-valid rewrite among the first four beam candidates, with the original fallback retained if no valid candidate exists. We do not exclude factuals on the basis of Base-generation success. Human-written neutral counterparts of evaluation sentences are not supplied to candidate generation or support optimization.

Unlike image latents, arbitrary sentence embeddings cannot generally be decoded into grammatical text. We therefore restrict the refined population to a finite codebook of actual sentences, represented by their MiniLM embeddings. Each population contains $M = n = 128$ equally weighted supports. GRACE searches this shared codebook under the two-anchor Wasserstein constraint and deploys the selected supports through a one-to-one optimal transport assignment from Base. The constraint couples the population but does not impose an individual meaning-preservation guarantee.

The prescribed classifier update consists of $T = 30$ full-batch gradient-descent steps with learning rate $\eta = 0.5$ on 128 sentence embeddings, balanced across the two classes. GRACE differentiates

Method	FV ↑	Sens. ↓	Cost ↓
Base	85.42	1.08143	0.3916
GRACE ($r = 0.25$)	99.48	0.02394	0.6307

Table 3: ParaDetox results on eight test factuals and 24 future models. GRACE uses $\beta_{\text{geo}} = 1.5$. FV is reported in percent; Cost denotes mean Euclidean distance between normalized factual and output embeddings.

Example 1: softened criticism with reduced specificity	
Factual	Thomas Jefferson is a pussy.
Base	Thomas Jefferson is a bad person.
GRACE	Thomas Jefferson is not good.
Example 2: removal of abusive wording	
Factual	don't u kid kid me bitch.
Base	don't u kid kid me.
GRACE	Don't kid me.
Example 3: acceptable Base retained	
Factual	And so no, the advice given to young adults is stupid.
Base	And so no, the advice given to young adults is wrong.
GRACE	And so no, the advice given to young adults is wrong.
Example 4: minimal semantic change after refinement	
Factual	If Verizon thinks they can redefine the word "unlimited," they can see me in fucking court.
Base	If Verizon thinks they can redefine the word "unlimited," they can see me in court.
GRACE	If Verizon thinks they can redefine the word "unlimited," then they can see me in court.

Table 4: Illustrative ParaDetox outputs showing both successful refinements and failure cases. GRACE uses $r = 0.25$ and $\beta_{\text{geo}} = 1.5$; examples are selected for qualitative illustration.

through all 30 steps. Sensitivity is measured with respect to continuous perturbations of the training embeddings. It therefore describes embedding-space uncertainty rather than arbitrary token- or word-level changes. Support optimization uses discrete coordinate search with at most ten sweeps, accepting feasible replacements that reduce the population objective by more than 10^{-10}. Geometric feasibility is checked with tolerance 10^{-8}. Termination indicates that no improving single-support replacement was found, rather than global optimality. No post-hoc validity correction or additional semantic gate is applied. We consider $r \in \{0, 0.25, 1, 4\}$ and $\beta_{\text{geo}} \in \{1.02, 1.10, 1.25, 1.50\}$. Calibration selects configurations by Future Validity, breaking ties by embedding cost, then r and β_{geo}.

Final evaluation uses eight factual sentences and held-out classifier updates under training-distribution shifts based on sentence length, classifier error, and topic. We report Future Validity, factual-to-output embedding distance, normalized word-edit distance, and embedding cosine similarity. The latter quantities are automatic proxies for semantic preservation; no human evaluation or independent toxicity evaluator is used. This experiment is a finite-codebook adaptation of GRACE, rather than continuous latent optimization or certified token-level robustness.

B.2 BASELINE DETAILS

We compare GRACE against four representative robust-CF methods: T-Rex (Hamman et al., 2023), ROAR (Upadhyay et al., 2021), PROPLACE (Jiang et al., 2024a), and ElliCE (Turbal et al., 2025). For each method, we follow its original counterfactual construction procedure and vary only the method-specific robustness parameter. Configurations are selected on the calibration split by maximizing Future Validity under calibration-only future-model updates. No future model from the final evaluation set is used for configuration selection.

Baseline hyperparameter grids. For T-Rex, we grid-search the stability threshold over $\{0.2, 0.4, 0.6\}$, while fixing the Gaussian perturbation standard deviation to $\sigma = 0.1$. For ROAR, we search the parameter-perturbation bound $\delta_{\max} \in \{0.005, 0.01, 0.05\}$. For PROPLACE, we jointly

vary the weight and bias perturbation bounds, setting $\delta = \delta_{\text{bias}} \in \{0.001, 0.005, 0.01\}$. For ElliCE, we search the robustness radius $\epsilon \in \{0.005, 0.01, 0.05\}$. Each configuration is selected by maximizing calibration future validity, with ties broken by smaller proximity and then smaller parameter values.

T-Rex. T-Rex constructs robust counterfactuals using the notion of *naturally occurring model change*, which permits potentially large parameter changes provided that the updated model remains sufficiently close to the original model in its predictions on the data manifold. Its robustness criterion is based on a local stability measure around the candidate counterfactual, yielding probabilistic guarantees against such model changes (Hamman et al., 2023). Following the original method, we use Wachter-style minimum-cost counterfactuals as the starting point and vary the stability threshold τ. Larger values of τ require stronger local stability and generally impose a stricter robustness requirement. We select τ on the calibration split using Future Validity.

ROAR. ROAR formulates robust algorithmic recourse as a minimax problem in which the recourse is optimized against adversarial changes to the predictive model (Upadhyay et al., 2021). In its original construction, model uncertainty is represented through bounded perturbations of model parameters, and the counterfactual is optimized to remain valid under the corresponding worst-case model shift. We vary the perturbation radius δ, which controls the size of this model-uncertainty set, and select it using calibration Future Validity.

PROPLACE. PROPLACE generates counterfactuals that are both plausible and provably robust to bounded changes in neural-network parameters (Jiang et al., 2024a). It formulates counterfactual generation as a robust-optimization problem and explicitly accounts for proximity and plausibility while enforcing validity under the prescribed model-parameter uncertainty set. We vary its robustness radius δ, which controls the magnitude of admissible model changes, and choose the value that maximizes Future Validity on the calibration split.

ElliCE. ElliCE defines robustness with respect to a Rashomon set of near-optimal predictive models (Turbal et al., 2025). Rather than perturbing one trained model directly, it constructs an ellipsoidal approximation of the set of models whose predictive loss remains near that of the reference model and generates recourse intended to remain valid throughout this set. The parameter ϵ controls the allowed increase in predictive loss and therefore the size of the approximated Rashomon set. We tune ϵ on the calibration split using the same Future Validity criterion.

Calibration and evaluation. For T-Rex, ROAR, PROPLACE, and ElliCE, we grid-search τ, δ, δ, and ϵ, respectively. For GRACE, we jointly select the data-side uncertainty radius r and geometry budget β_{geo}. All selected configurations are frozen before final evaluation. Each method is then evaluated on the same factual instances and the same held-out future-model realizations, with its returned counterfactuals fixed before these future models are scored. The exact search grids and optimization settings are reported below.

B.3 COMPUTATIONAL EFFICIENCY

Computational cost. GRACE performs a joint population refinement, whereas the compared baselines construct CFs independently for each factual instance. We therefore compare the wall-clock time required to process the same ten factual instances. For GRACE, we perform a fresh optimization with ten equal-mass supports and report both refinement-only and end-to-end runtime, where the latter additionally includes Wachter Base-CF generation. Seconds per CF is the total runtime divided by ten and should be interpreted as amortized batch cost rather than single-query latency.

Table 5 shows that population-level refinement introduces additional computation but remains practical at the evaluated population size. GRACE processes ten CFs in approximately 20–82 seconds across the four datasets, corresponding to about 2–8 seconds per requested CF. Base-CF generation contributes only a small fraction of the end-to-end runtime, with most computation arising from the update-aware population refinement.

These timings are observed implementation runtimes rather than hardware-matched algorithmic speedups: T-Rex and ElliCE run on a single CPU thread, whereas GRACE refinement uses a GPU.

Table 5: Observed runtime for ten fixed factual instances per dataset. GRACE performs a fresh ten-support population refinement initialized from Wachter CFs; end-to-end runtime additionally includes Base-CF generation. Seconds per CF denotes amortized batch cost over the ten requested factuals. FV is averaged over the held-out future models.

Dataset	Method	Valid / 10	Total (s)	s / CF	FV (%) ↑
HELOC	T-Rex	10/10	90.297	9.030	100.0
	ElliCE	10/10	0.191	0.019	100.0
	GRACE (refinement)	10/10	81.639	8.164	100.0
	GRACE (end-to-end)	10/10	82.050	8.205	100.0
Diabetes	T-Rex	10/10	10.154	1.015	100.0
	ElliCE	10/10	0.207	0.021	100.0
	GRACE (refinement)	10/10	22.780	2.278	100.0
	GRACE (end-to-end)	10/10	23.195	2.319	100.0
German Credit	T-Rex	10/10	42.038	4.204	100.0
	ElliCE	10/10	0.391	0.039	100.0
	GRACE (refinement)	10/10	58.569	5.857	100.0
	GRACE (end-to-end)	10/10	60.516	6.052	100.0
FICO	T-Rex	10/10	88.422	8.842	100.0
	ElliCE	10/10	0.330	0.033	100.0
	GRACE (refinement)	10/10	18.978	1.898	100.0
	GRACE (end-to-end)	10/10	19.701	1.970	100.0

All experiments use an MLP and one fixed experiment seed. Runtimes characterize the evaluated implementations rather than hardware-matched algorithmic speedups. Shared classifier fitting, hyperparameter calibration, model loading, dataset I/O, and final future-model evaluation are excluded.

Importantly, GRACE is a population-specific batch refinement rather than an offline training procedure; a new factual population requires a fresh refinement.

B.4 Different Base CF Generator and Base Classifier

GRACE is designed as a post-hoc robustness wrapper and does not depend on a particular base counterfactual generator. To evaluate this generator-agnostic property, we repeat the refinement procedure using two commonly used base generators: Wachter (Wachter et al., 2018) and DiCE (Mothilal et al., 2020).

Wachter. The Wachter method generates a CF by balancing prediction validity against proximity to the factual instance. This is the default base generator used in our main tabular experiments. The resulting Base CF population defines the proximal anchor, after which GRACE performs population-level robust refinement without modifying the underlying Wachter objective.

DiCE. DiCE generates diverse CF explanations by jointly optimizing for validity, proximity, and diversity. For each factual instance, we use the corresponding DiCE output as the Base CF and construct the population anchor from these outputs. GRACE is then applied unchanged, using the same population-level objective, geometry constraint, calibration procedure, and future-model evaluation protocol as in the Wachter experiments.

Across both generators, GRACE operates only on the resulting Base CF population; no parameters of the base generator are retrained during robust refinement. This experiment therefore isolates whether the gains from GRACE depend on a particular counterfactual construction method.

B.5 Construction of evaluation models.

Each evaluation scenario specifies a perturbation family, a severity s, and a realization seed. We first apply the perturbation to a fixed training set, obtaining a shifted dataset D_k. We then initialize the evaluation classifier from the same original checkpoint θ_0 and apply the full-batch gradient updates.

Scenarios do not continue training from one another. Perturbations are applied to the model-update data, rather than to the CFs being scored. After constructing each evaluation model, we evaluate the same fixed CFs under that model.

Tabular shifts: HELOC. We construct shifts in the inherited normalized feature space. Let A denote the affected training examples, $q = |A|/N$, and J the three features with the largest absolute training-label correlations. Population-wide shifts use all training examples, whereas target-side shifts affect only examples with favorable labels. Mixed shifts use the population-wide perturbation followed by random subsampling. For severity $s \in \{0.5, 1, 2.5\}$, define $a = sb_{\text{ref}}$, where b_{ref} is the saved reference shift scale.

The perturbation operator cycles among the following:

- **Mean shift.** Each selected feature receives a shared random-sign offset:

$$x'_{ij} = x_{ij} + \frac{a}{\sqrt{|J|q}}\,\xi_j, \qquad i \in A,\ j \in J,$$

  where $\xi_j \in \{-1, +1\}$ is sampled independently for each selected feature.

- **Scale shift.** Let μ be the mean of $x_{i,J}$ over $i \in A$ and

$$R = \left(\frac{1}{|A|}\sum_{i\in A}\|x_{i,J} - \mu\|_2^2\right)^{1/2}.$$

  We expand the selected coordinates around their mean:

$$x'_{i,J} = \mu + \left(1 + \frac{a}{\sqrt{q}\max(R, 10^{-6})}\right)(x_{i,J} - \mu).$$

- **Tail resampling.** We divide the affected examples at the median of their most label-correlated feature. Feature rows are sampled with replacement, allocating a fraction $\min(0.5 + 0.18s, 0.95)$ to the lower group and the remainder to the upper group. The implementation retains the labels at their original row positions; consequently, this operator can alter feature–label associations when the affected population contains both classes.

For mixed shifts, we subsequently retain a uniformly sampled subset of

$$N' = \max(96, \lfloor N[1 - \min(0.04s, 0.2)]\rfloor)$$

examples, selecting features and labels together at this stage. Five realizations use the operator sequence mean, scale, tail, mean, scale. Combining three shift families, three severities, and five realizations yields 45 scenarios per split. Each shifted dataset updates all parameters of the original HELOC classifier with learning rate 0.01 and binary cross-entropy.

MNIST image shifts. We perturb a fixed set of 256 class-balanced training images in pixel space, retaining their labels. For $s \in \{0.5, 1, 1.5, 2\}$, each scenario applies one of:

- rotation by independently sampled angles $+15s$ or $-15s$ degrees for each image;
- translation by $(\text{round}(2s), -\text{round}(2s))$ pixels, shared across images;
- additive Gaussian noise, $x' = \text{clip}_{[0,1]}(x + 0.08s\,\epsilon)$, with independent standard-normal pixel noise ϵ.

Each shifted dataset produces an evaluation CNN through full-batch binary-cross-entropy updates with learning rate 0.01. All classifier parameters are updated, with batch-normalization statistics frozen and dropout disabled. Three realizations per family and severity yield 36 scenarios per split. Translation is deterministic at each severity, so its three realizations are repeated distributions.

Table 6: Calibration-selected GRACE configurations. Entries report $(\beta_{\text{geo}}, r/b_{\text{ref}})$. GRACE–DiCE is evaluated with the MLP classifier only.

Dataset	b_{ref}	MLP–Wachter	MLP–DiCE	LogReg–Wachter
HELOC	3.6906	$(1.10, 0.25)$	$(1.05, 3)$	$(1.10, 1)$
Diabetes	1.5143	$(1.20, 3)$	$(1.10, 3)$	$(1.05, 3)$
German Credit	5.5429	$(1.10, 3)$	$(1.05, 1)$	$(1.05, 1)$
FICO	4.7251	$(1.10, 3)$	$(1.05, 1)$	$(1.05, 0.25)$

Galaxy and FFHQ image shifts. Galaxy uses 40 fixed training images, four per class. FFHQ uses 32 balanced generated training images with frozen pseudo-labels from the original classifier. For both datasets, brightness shifts multiply all images in a scenario by

$$c = 1 - 0.2s\,u, \qquad u \sim \mathcal{U}(0.8, 1.2),$$

followed by clipping to $[0, 1]$. Noise shifts use

$$x' = \text{clip}_{[0,1]}(x + 0.04s\,\epsilon).$$

FFHQ additionally includes Gaussian blur with a 9×9 kernel and standard deviation $\sigma = 0.8s\,u$, where one u is sampled per scenario. Each scenario applies a single perturbation family; these image transformations are not composed.

The shifted images are passed through the frozen feature extractor. Starting from the original checkpoint, Galaxy updates only its final classification layer, whereas FFHQ updates only the CLIP visual projection. Both use full-batch cross-entropy updates with learning rate 10^{-4}. Four severities and two realizations produce 16 Galaxy and 24 FFHQ scenarios per split.

B.6 SELECTION OF r AND β_{geo}

For both GRACE–Wachter and GRACE–DiCE, we use the same calibration grid

$$\beta_{\text{geo}} \in \{1.05, 1.10, 1.20\}, \qquad r/b_{\text{ref}} \in \{0, 0.25, 1, 3\},$$

giving 12 configurations per dataset, classifier, and Base generator. Here, β_{geo} controls the geometry budget through

$$W_2(\mathbb{P}_{\text{CF}}, \mathbb{Q}) + W_2(\mathbb{Q}, \mathbb{P}_1) \le \beta_{\text{geo}} W_2(\mathbb{P}_{\text{CF}}, \mathbb{P}_1),$$

while r controls the update-aware sensitivity penalty. We search over the normalized radius r/b_{ref}, with absolute radius $r = (r/b_{\text{ref}})b_{\text{ref}}$.

b_{ref} is a reference number describing how much the dataset changes when we randomly resample it. We use it to scale the robustness setting to each dataset, so the same search values have a comparable meaning across datasets. It is estimated from calibration data, not future test results.

We retain only geometry-feasible calibration solutions and select the configuration with the highest mean robustness. Ties are broken by lower proximity, followed by smaller r and then smaller β_{geo}. The selected configuration is frozen before final test evaluation. Table 6 reports the selected configurations and dataset-specific reference scales.

The selected configurations vary across datasets, classifiers, and Base generators, supporting calibration of (r, β_{geo}) rather than a single universal robustness setting.

B.7 ADDITIONAL EXPERIMENTS

In addition to the main experiments with the original classifier, we further evaluate all methods under a logistic-regression classifier to examine whether the observed trends are robust to the choice of downstream model. Table 7 summarizes the results. Overall, the same conclusion holds: **GRACE–Wachter** and **GRACE–DiCE** consistently achieve higher future validity than their base counterparts, while maintaining competitive or even better proximity. We also note that ROAR and PROPLACE cannot be applied to German Credit and FICO under this classifier, and are therefore marked as "–".

We provide two additional analyses on MNIST to examine the robustness of GRACE to misspecification of the prescribed update horizon and to isolate the role of the CF-side geometry constraint. Unless otherwise stated, the CF population is optimized once and then kept fixed during evaluation.

Table 7: Comparison with robust CF baselines on tabular datasets under the logistic-regression classifier. We report Future Validity (FV, %, ↑) and recourse Proximity (↓). Here, the *DiCE (Base)* and *Wachter (Base)* rows are shown only as the unrefined base CFs and are not included in the robust-baseline comparison.

	HELOC		Diabetes		German Credit		FICO	
Method	FV ↑	Prox. ↓	FV ↑	Prox. ↓	FV ↑	Prox. ↓	FV ↑	Prox. ↓
T-Rex	*100.0 ± 1.4*	*4.4 ± 0.7*	*100.0 ± 1.6*	*2.7 ± 0.4*	*100.0 ± 1.8*	*8.8 ± 0.9*	100.0 ± 1.2	4.9 ± 0.6
ROAR	56.3 ± 1.8	2.9 ± 0.9	91.4 ± 1.2	1.9 ± 0.5	–	–	–	–
PROPLACE	99.1 ± 1.3	3.3 ± 0.4	100.0 ± 2.1	2.3 ± 0.8	–	–	–	–
ElliCE	**100.0 ± 2.0**	3.8 ± 0.6	**100.0 ± 1.4**	2.0 ± 0.9	**100.0 ± 2.4**	10.4 ± 0.7	**100.0 ± 1.6**	6.1 ± 0.4
DiCE (Base)	90.2 ± 1.7	4.1 ± 0.8	93.5 ± 1.3	2.6 ± 0.5	86.7 ± 1.9	3.7 ± 1.0	93.8 ± 1.4	5.5 ± 0.6
GRACE–DiCE	**100.0 ± 1.1**	4.2 ± 0.3	99.8 ± 0.9	2.8 ± 0.7	99.7 ± 1.2	3.8 ± 0.4	**100.0 ± 0.8**	5.6 ± 0.9
Wachter (Base)	90.1 ± 1.5	1.6 ± 0.2	85.1 ± 1.2	0.9 ± 0.5	59.6 ± 1.8	1.8 ± 0.9	84.3 ± 1.3	1.8 ± 0.3
GRACE–Wachter	**99.5 ± 1.2**	**2.0 ± 0.6**	**100.0 ± 1.0**	**1.4 ± 0.4**	**99.3 ± 1.1**	**2.1 ± 0.8**	**98.9 ± 0.9**	**2.1 ± 0.5**

Table 8: MNIST update-horizon mismatch. GRACE is optimized with $T_{\text{design}} = 3$, $\eta = 0.01$, $r = 0.05$, and $\beta_{\text{geo}} = 1.02$. The resulting fixed CFs are evaluated across 36 future-update trajectories for each evaluation horizon. Pixel-space proximity remains unchanged across horizons: 5.3584 for Base and 5.4379 for GRACE.

	FV (%) ↑		Training	Parameter
T_{eval}	Base	GRACE	loss	$\|\theta_T - \theta_0\|_2$
3	77.39	99.87	0.05173	0.03599
5	76.95	99.83	0.04097	0.04328
10	75.74	99.74	0.02795	0.05733
20	75.56	99.74	0.01748	0.07728
40	74.61	99.61	0.00993	0.10430

Robustness to update-horizon mismatch. The main formulation assumes a prescribed T-step future model update. We therefore test whether the resulting CF population remains robust when the realized number of update steps differs from the design value. GRACE is optimized assuming $T_{\text{design}} = 3$ and is then evaluated, without regenerating or refining the CFs, under increasingly long update trajectories.

Table 8 shows that GRACE remains highly valid even when the realized update horizon substantially exceeds the design horizon. As T_{eval} increases from 3 to 40, the updated classifier moves progressively farther from θ_0, as reflected by the increasing parameter displacement, while Base Future Validity declines from 77.39% to 74.61%. In contrast, the same fixed GRACE population remains above 99.6% Future Validity throughout. Because the CFs are not changed across evaluation horizons, the proximity cost is constant. This experiment does not establish robustness to arbitrary changes in the update mechanism, but suggests that GRACE can tolerate substantial mismatch in the number of realized update steps.

Role of the geometry constraint. We next remove the CF-side geometry constraint while keeping the remaining GRACE objective unchanged. This comparison isolates whether the transport constraint merely restricts optimization or meaningfully controls how far the refined population moves from the Base population and favorable-class geometry.

As shown in Table 9, removing the geometry constraint does not materially improve Future Validity in this setting: both variants are already near saturation. However, the unconstrained solution incurs substantially larger pixel-space proximity cost and a larger latent-space detour, and it falls outside the original geometry budget. Thus, in this experiment, the geometry constraint acts primarily as a control on population displacement rather than as a mechanism for increasing Future Validity itself. The result also illustrates why geometry and factual-to-CF proximity should not be identified: they are related, but distinct quantities. Taken together, these experiments suggest that the prescribed update horizon need not be matched exactly at evaluation time, while the two-anchor geometry provides useful control over how robustness is obtained.

Table 9: MNIST geometry ablation with $T = 3$, $r = 0.05$, and 800 support updates. The constrained configuration uses $\beta_{\text{geo}} = 1.02$, whereas the unconstrained configuration removes the geometry constraint.

Quantity	Constrained	Unconstrained
FV (%) ↑	99.87	99.91
Pixel-space proximity ↓	5.4379	6.2841
Surrogate $\widetilde{\Phi}_r$ ↓	1.4607×10^{-4}	1.3358×10^{-4}
$W_2(\mathbb{P}_{\text{CF}}, Q)$	1.1789	1.6427
$W_2(Q, \mathbb{P}_1)$	58.3253	59.4861
Total latent-space detour	59.5043	61.1288
Within original budget (59.8108)	Yes	No

Table 10: Held-out image results using calibration-selected GRACE configurations. Current Validity (CV) and Future Validity (FV) are reported in percent. Proximity is the mean factual-to-CF pixel-space ℓ_2 distance. The Base denotes the decoder replay used to initialize the refinement.

Dataset / Task	Method	CV ↑	FV ↑	Prox. ↓
Galaxy10: Merging→Round	Base	75.00	69.92	12.635
	GRACE	**100.00**	**100.00**	12.786
Galaxy10: Round→Cigar	Base	50.00	50.00	14.165
	GRACE	**100.00**	**100.00**	14.193
FFHQ: Face→Smiling Face	Base	87.50	82.29	**38.744**
	GRACE	**100.00**	**100.00**	47.503

B.8 ADDITIONAL IMAGE RESULTS

Table 10 reports held-out results for additional image tasks using the calibration-selected GRACE configurations. Across all three tasks, refinement raises both current and future validity to 100% On the two Galaxy10 tasks, these gains require only a small increase in pixel-space proximity. On FFHQ, the improvement in validity is accompanied by a larger proximity increase.

For Merging→Round and FFHQ, calibration selects $r = 0.25$, whereas for Round→Cigar it selects $r = 0.25$.

Algorithm 1 GRACE update-aware population refinement

Require: base-CF population $\mathbb{P}_{\mathrm{CF}} = \frac{1}{n}\sum_{i=1}^{n}\delta_{z_i}$; favorable anchor $\mathbb{P}_1$; empirical training distribution $\widehat{\mathbb{P}}_{\mathrm{tr}}$; deployed parameters θ_0; calibrated uncertainty radius r; geometry budget B_{geo}; update horizon T and step sizes $\{\eta_t\}_{t=0}^{T-1}$; primal step size γ; augmented-Lagrangian parameter ρ; tolerances $\varepsilon_{\mathrm{feas}}, \varepsilon_{\mathrm{obj}}$; maximum iterations K.

Compute $D \leftarrow W_{\mathrm{CF}}(\mathbb{P}_{\mathrm{CF}}, \mathbb{P}_1)$ and require $B_{\mathrm{geo}} \geq D$.

Initialize

$$w_j^{(0)} \leftarrow z_j, \qquad \mathbb{Q}^{(0)} = \frac{1}{n}\sum_{j=1}^{n}\delta_{w_j^{(0)}}, \qquad \lambda^{(0)} \leftarrow 0.$$

$\triangleright\ \mathbb{Q}^{(0)} = \mathbb{P}_{\mathrm{CF}}$ is feasible.

Construct the prescribed T-step update map $h \mapsto \theta_T(h)$ from Section 3.4 and set $\theta_T^{\mathrm{nom}} \leftarrow \theta_T(0)$.

Set $\widetilde{\Phi}_{\mathrm{prev}} \leftarrow$ NONE.

for $k = 0, \ldots, K-1$ **do**

  Form $\mathbb{Q}^{(k)} = \frac{1}{n}\sum_{j=1}^{n}\delta_{w_j^{(k)}}$.

  Compute the update-aware population sensitivity $s_{\mathbb{Q}^{(k)}}$ using Eq. (4) by reverse-mode differentiation through the T update steps.

  Evaluate the geometry violation

$$g^{(k)} \leftarrow W_{\mathrm{CF}}(\mathbb{P}_{\mathrm{CF}}, \mathbb{Q}^{(k)}) + W_{\mathrm{CF}}(\mathbb{Q}^{(k)}, \mathbb{P}_1) - B_{\mathrm{geo}}.$$

  if $\widetilde{\Phi}_{\mathrm{prev}} \neq$ NONE **then**

    Compute

$$\Delta_{\mathrm{obj}} \leftarrow \frac{|\widetilde{\Phi}_r(\mathbb{Q}^{(k)}) - \widetilde{\Phi}_{\mathrm{prev}}|}{\max\{1, |\widetilde{\Phi}_{\mathrm{prev}}|\}}.$$

    if $g^{(k)} \leq \varepsilon_{\mathrm{feas}}$ **and** $\Delta_{\mathrm{obj}} \leq \varepsilon_{\mathrm{obj}}$ **then**

      break

    end if

  end if

  Form the augmented-Lagrangian objective

$$\mathcal{L}_\rho^{(k)} = \widetilde{\Phi}_r(\mathbb{Q}^{(k)}) + \lambda^{(k)} g^{(k)} + \frac{\rho}{2}[g^{(k)}]_+^2.$$

  Update the support locations

$$w_j^{(k+1)} \leftarrow \Pi_{\mathcal{X}}\left(w_j^{(k)} - \gamma \nabla_{w_j}\mathcal{L}_\rho^{(k)}\right), \qquad j = 1, \ldots, n,$$

  where $\Pi_{\mathcal{X}}$ projects onto the valid feature domain.

  Form $\mathbb{Q}^{(k+1)} = \frac{1}{n}\sum_{j=1}^{n}\delta_{w_j^{(k+1)}}$.

  Update

$$\lambda^{(k+1)} \leftarrow \left[\lambda^{(k)} + \rho\, g(\mathbb{Q}^{(k+1)})\right]_+.$$

  Set $\widetilde{\Phi}_{\mathrm{prev}} \leftarrow \widetilde{\Phi}_r(\mathbb{Q}^{(k)})$.

end for

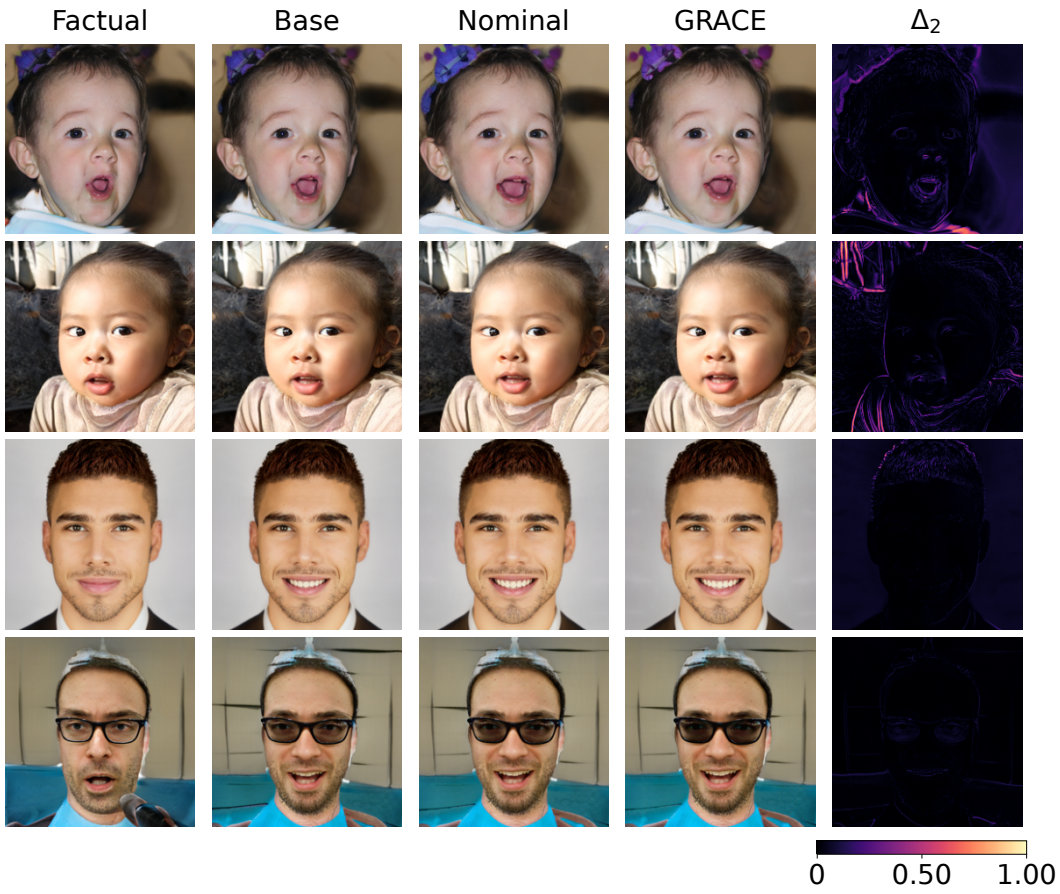
Let $\widetilde{\mathbb{Q}}_{r,B_{\mathrm{geo}}}$ be the final iterate.

Verify the geometry constraint using the final unregularized discrete OT computation:

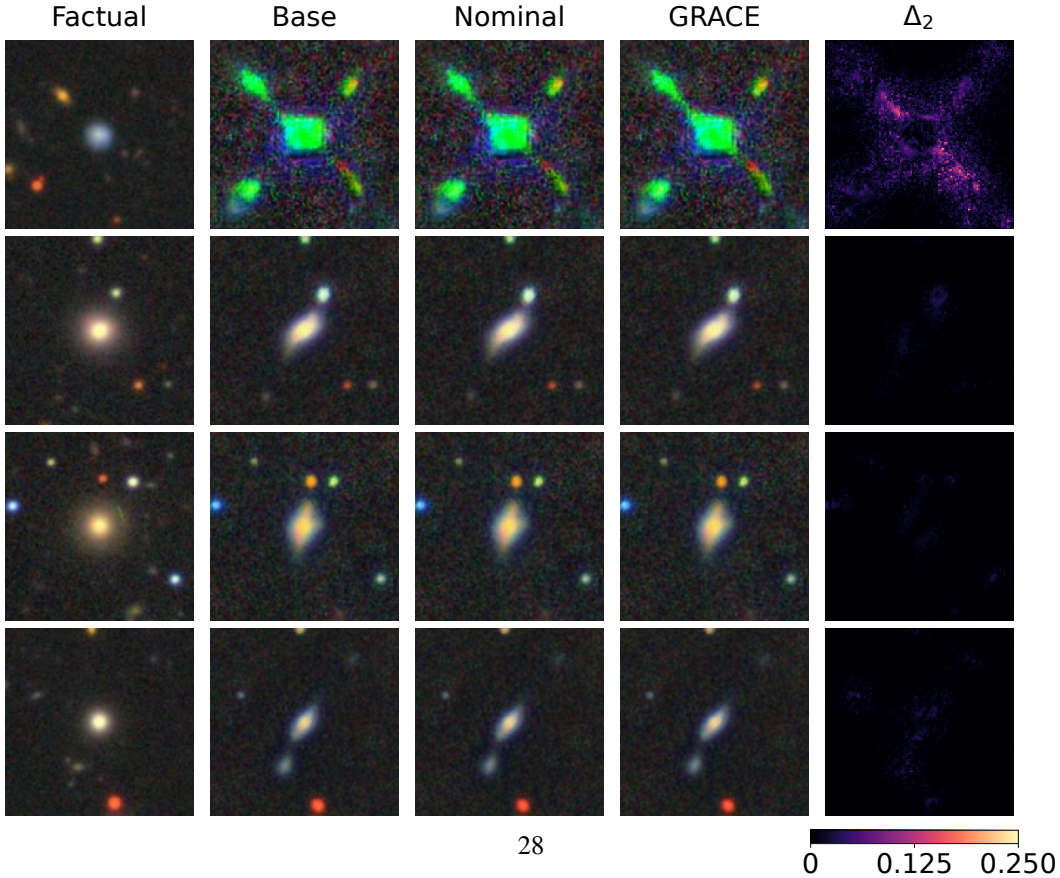
$$W_{\mathrm{CF}}(\mathbb{P}_{\mathrm{CF}}, \widetilde{\mathbb{Q}}_{r,B_{\mathrm{geo}}}) + W_{\mathrm{CF}}(\widetilde{\mathbb{Q}}_{r,B_{\mathrm{geo}}}, \mathbb{P}_1) \leq B_{\mathrm{geo}} + \varepsilon_{\mathrm{feas}}.$$

Using the ground cost underlying W_{CF}, compute the minimum-cost one-to-one assignment $\pi^\star$ between the base supports $\{z_i\}_{i=1}^n$ and the refined supports $\{w_j\}_{j=1}^n$.

return $z_i^{\mathrm{out}} \leftarrow w_{\pi^\star(i)}$ for $i = 1, \ldots, n$, together with $\widetilde{\mathbb{Q}}_{r,B_{\mathrm{geo}}}$.



(a) FFHQ (Face → Smiling Face)



(b) Galaxy10 (Round → Cigar)