DIAGNOSING COUNTERFACTUAL BEHAVIOR THROUGH LOCAL REPRESENTATION GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Counterfactual explanations seek small, semantically meaningful changes to an input that alter a model's prediction, helping interpret and audit machine learning systems. In modern vision, language, and multimodal systems, pretrained encoders map inputs to representation spaces, where task-specific classifier heads impose decision boundaries. Crossing a boundary changes the prediction, but the resulting representation may lie far from observed target-class examples. In applications such as biomedical research, a prediction change can mislead if it pushes the representation far from observed alternatives. To address this gap, this work uses local representation geometry to improve counterfactual search and establish a model-level diagnostic framework. A standardized local search probe characterizes each encoder–classifier pair by its counterfactual success, minimum displacement, search effort, and endpoint support. Across multiple domains and widely-used pretrained model families in vision, medical imaging, text, multimodal data, and single-cell biology, models with comparable predictive performance exhibit different counterfactual outcomes. Our framework associates these differences with classifier boundary proximity and local representation geometry, quantified by target-class support and local curvature. Incorporating support into the search objective and adapting step sizes to curvature increases counterfactual success under matched constraints. In a real-world biology case study, the generated counterfactuals align with experimentally observed differences between low- and high-immune response in single cells. Together, these findings establish counterfactual behavior as a distinct dimension beyond predictive performance and provide a framework for model diagnosis and more reliable counterfactual search.

1 INTRODUCTION

Counterfactual explanations ask: *can a model's prediction be changed through a small, semantically meaningful change to the input?* Wachter et al. (2018); Verma et al. (2024) They provide an intuitive and actionable form of explanation, and help interpret and audit machine learning systems in domains such as healthcare and finance, where understanding which changes could alter a decision can inform subsequent actions Ustun et al. (2019); Karimi et al. (2021).

Much existing work on counterfactual explanations studies tabular data and feature-space perturbations Wachter et al. (2018); Mothilal et al. (2020). These methods typically formulate counterfactual generation as a local optimization problem: given an input, they search for a small prediction-changing perturbation subject to task-specific feasibility constraints. However, a prediction-changing perturbation may not exist within the imposed locality budget, which bounds displacement from the starting point, and boundary crossing may still lead to regions with weak data support, where few observed target-class examples are nearby. Related observations in adversarial robustness show that predictions can often be flipped by small perturbations that do not correspond to realistic nearby alternatives Fawzi et al. (2017); Tanay & Griffin (2016). Thus, whether a nearby prediction change is feasible and meaningful depends not only on reaching a decision boundary, but also on the local geometry of the data around it.

In current machine learning systems, prediction often proceeds through learned representations: a pretrained encoder maps inputs to an embedding space, and a task-specific classifier partitions that space into decision regions Devlin et al. (2019); Radford et al. (2021); Bengio et al. (2013).

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.

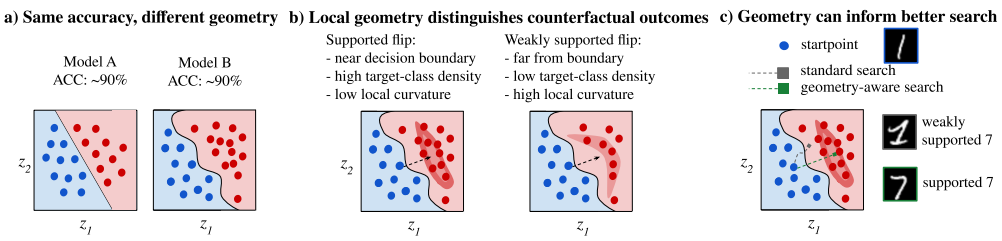

Figure 1: **Local geometry diagnoses and guides counterfactual behavior.** (a) Similarly accurate models can have different counterfactual profiles because they organize examples differently in representation space. (b) Boundary proximity, target-class support, and local curvature help explain these differences. (c) The same geometric quantities can guide local search toward better-supported counterfactual outcomes, as demonstrated by an illustrative MNIST example.

Studying counterfactual behavior in this space is natural because the classifier makes its prediction directly from the embedding. Changing that prediction requires crossing a decision boundary, with the encoder shaping the surrounding representation geometry and the classifier head placing the boundary. Yet this move across the decision boundary (captured by the predictive performance) does not guarantee whether a model admits nearby, data-supported prediction changes. Consider two models that predict whether an immune cell responds strongly to a biological signal. They may have similar predictive performance while organizing cells differently in representation space D'Amour et al. (2020); Kornblith et al. (2019). One finds a small prediction change near observed high-response cells; the other reaches a high-response prediction in a region with few observed cells. Treating the second result as a plausible biological alternative could misdirect follow-up experiments. This raises a central question: *how does local representation geometry shape counterfactual behavior?*

To address this gap, this work uses local representation geometry to improve counterfactual search and establish a model-level diagnostic framework. A standardized local search probe profiles each encoder–classifier pair by success, minimum displacement, search effort, and endpoint support. Figure 1a illustrates that models with similar predictive performance but different counterfactual profiles can organize examples differently in representation space. The diagnostic framework explains these differences by examining how target-class density and representation curvature interact with classifier boundary placement (Figure 1b). Boundary proximity measures how far a representation lies from a classifier boundary, target-class density measures the availability of nearby target examples, and local curvature captures departure from a linear approximation of the surrounding representation structure. Across a comprehensive evaluation setup of major pretrained model families and diverse datasets (vision, medical imaging, text, multimodal data, and single-cell biology), local representation geometry distinguishes counterfactual outcomes among models with comparable predictive performance and guides search toward better-supported outcomes. Figure 1c illustrates this search benefit on MNIST, where geometry-scaled search reaches a supported target-class region while standard search terminates in a weakly supported one.

Contributions This work addresses a gap in counterfactual research: despite the widespread use of pretrained representations, little is known about how their local geometry relates to counterfactual behavior. To address this, this work introduces a geometry-based diagnostic framework that separates classifier-side boundary proximity from representation-side target-class support and curvature. Second, across diverse datasets, the framework shows that state-of-the-art models with similar predictive performance can exhibit substantially different counterfactual behavior. Third, it translates these geometric quantities into improved counterfactual search mechanisms, with the goal of producing counterfactuals that are valid in input space and useful in downstream applications. Together, these contributions identify representation geometry as a missing piece in counterfactual diagnosis and a guide to model selection, robustness analysis, and search.

2 Related Work

This work connects four lines of research: representation geometry, decision-boundary analysis, counterfactual explanations, and model multiplicity. These areas provide individual pieces of the

problem, but do not jointly explain how representation geometry and classifier boundary placement shape model-level counterfactual behavior.

Geometry of learned representations Prior work has characterized learned representations through manifold radius, intrinsic dimension, and inter-class correlations Cohen et al. (2020); Chung et al. (2018), as well as anisotropy, effective rank, and neural collapse Ethayarajh (2019); Roy & Vetterli (2007); Papyan et al. (2020); Kornblith et al. (2019). Manifold-learning methods additionally estimate local structure using tangent spaces, local PCA, and curvature Álvarez Vizoso et al. (2020); Ache & Warren (2019). These lines of work characterize the organization and local geometry of learned representations, but do not study how these properties relate to counterfactual search behavior. In contrast, our work examines how local geometry, as quantified by target-class density and representation curvature, relates to counterfactual outcomes under locally constrained search.

Decision boundaries and adversarial robustness Prior work has analyzed decision-boundary geometry through adversarial robustness, relating boundary crossing to margin, curvature, and local linearity Goodfellow et al. (2015); Moosavi-Dezfooli et al. (2016); Fawzi et al. (2017); Tanay & Griffin (2016); Elsayed et al. (2018); Jiang et al. (2019). Related work has also connected curvature of neural-network level sets to changes in local gradient-based explanations Dombrowski et al. (2019). These analyses primarily concern the geometry of the classifier or its decision surfaces. In contrast, our work establishes the importance of jointly using local geometry and decision-boundary proximity to characterize counterfactual behavior in learned representations. This distinction is particularly critical given the widespread use of pretrained encoder-based models with task-specific classifier heads, which would serve as the base for counterfactual generation across domains.

Counterfactual explanations and recourse Counterfactual methods seek minimal prediction changes that remain plausible and data-supported Wachter et al. (2018); Mothilal et al. (2020); Laugel et al. (2019); Poyiadzi et al. (2020); Pawelczyk et al. (2020a). Latent-space, generative, and manifold-aware methods use learned representations or distributional constraints to encourage such changes Joshi et al. (2019); Pawelczyk et al. (2020a); Tsiourvas et al. (2024); Pegios et al. (2024); Zaher et al. (2026). Related work examines actionable recourse and adversarial perturbations Ustun et al. (2019); Karimi et al. (2021); Pawelczyk et al. (2022). These approaches focus mainly on individual instances, such as reversing one applicant's loan decision. Our work instead examines counterfactual behavior across models and representations.

Model multiplicity and underspecification Models with similar predictive performance can exhibit materially different predictions, explanations, robustness, and recourse behavior Fisher et al. (2019); Marx et al. (2020); D'Amour et al. (2020). Prior work has shown specifically that counterfactual costs and validity can vary across competing models, motivating counterfactuals that remain valid or robust under predictive multiplicity Pawelczyk et al. (2020b). Such variability can also arise from classifier boundary placement: retraining only the final layer can substantially alter behavior without modifying the underlying representations Kirichenko et al. (2023). We complement this literature by diagnosing how representation-side target-class density and curvature interact with classifier-side boundary placement to explain variation in counterfactual behavior among models with similar predictive performance. This diagnostic can enable more informed model selection for counterfactual generation.

3 Geometric Framework for Counterfactuals

This section formalizes the proposed framework. A standardized local search probe first summarizes each encoder–classifier pair through a counterfactual profile. Boundary proximity, target-class support, and curvature then help explain differences between profiles and guide counterfactual search.

Let $\mathcal{X}$ be the input space, $\mathcal{Z} = \mathbb{R}^d$ the representation space, and $\mathcal{Y} = \{1, \dots, C\}$ the label space. A pretrained encoder maps x to $z = e(x)$; a classifier with logits $h : \mathcal{Z} \to \mathbb{R}^C$ predicts $y = f(z) = \arg\max_{c \in \mathcal{Y}} h(z)_c$. For a target class $y' \neq y$, the probe seeks a nearby representation predicted as y' and records success, displacement, optimization effort, and endpoint data support. The ideal targeted minimum-distance objective is Wachter et al. (2018); Poyiadzi et al. (2020)

$$z^\star \in \arg\min_{\tilde{z} \in \mathcal{Z}} \|\tilde{z} - z\|_2 \quad \text{s.t.} \quad f(\tilde{z}) = y', \qquad y' \neq y. \tag{1}$$

The probe searches under a fixed locality radius and iteration budget, without assuming that it finds a global minimum.

Local target-class support Crossing a decision boundary does not guarantee that the resulting region is supported by observed target-class examples. Following manifold-aware counterfactuals and recourse Poyiadzi et al. (2020); Joshi et al. (2019); Ustun et al. (2019), we characterize target-class support through class-conditional neighborhoods.Let $\mathcal{R}_c = \{e(x_i) : y_i = c\}$ denote the reference embeddings whose observed labels are c. The support radius

$$r_k(z, c) = \mathrm{kNNdist}_k(z, \mathcal{R}_c) \tag{2}$$

is the Euclidean distance from z to its kth nearest neighbor in $\mathcal{R}_c$. Evaluated at the initial representation, $r_k(z, y')$ measures the distance from the search starting point to its kth nearest target-class example. For fixed k, class size, and representation dimension, ranking points by r_k is the reverse of ranking them by the corresponding standard k-NN density estimate. This interpretation and sensitivity to neighborhood size are detailed in Section A.1.

Local representation curvature Curvature characterizes the departure of the local representation structure from linearity. Although the ambient space is Euclidean, nearby embeddings may occupy a curved lower-dimensional structure. Let $\mathcal{N}_k(z)$ contain the k nearest reference embeddings, irrespective of class. Following standard local tangent-space estimation Aamari & Levrard (2018), the estimated tangent subspace T_z is spanned by the leading m right singular vectors of the matrix with rows $(u - z)^\top$, for $u \in \mathcal{N}_k(z)$. Let P_{T_z} and $P_{T_z^\perp}$ denote orthogonal projections onto this subspace and its complement. Motivated by the second-order normal departure of a smooth manifold from its tangent approximation Álvarez Vizoso et al. (2020), the curvature proxy is

$$\kappa_k(z) = \frac{\frac{1}{k}\sum_{u \in \mathcal{N}_k(z)} \left\| P_{T_z^\perp}(u - z) \right\|_2}{\frac{1}{k}\sum_{u \in \mathcal{N}_k(z)} \|P_{T_z}(u - z)\|_2^2 + \varepsilon}, \tag{3}$$

where $\varepsilon > 0$ stabilizes the denominator. Larger values indicate greater normal departure relative to squared tangential extent. The proxy depends on representation scale, neighborhood size, and tangent dimension, and is evaluated at the initial representation for diagnostic analyses.

Decision-boundary placement Let $h(z)$ denote the classifier logits and define the pairwise logit gap $g_{y,y'}(z) = h(z)_y - h(z)_{y'}$. A first-order approximation to the distance to the pairwise equality surface is

$$d_{\mathrm{bd}}(z; y, y') \approx \frac{|g_{y,y'}(z)|}{\|\nabla_z g_{y,y'}(z)\|_2 + \varepsilon}, \tag{4}$$

following local linearization analyses Moosavi-Dezfooli et al. (2016); Fawzi et al. (2017); Tanay & Griffin (2016); Jiang et al. (2019); Elsayed et al. (2018). For the linear heads used in the main experiments, $h(z) = Wz + b$, the pairwise distance has the exact form

$$d_{\mathrm{bd}}(z; y, y') = \frac{\left|(w_y - w_{y'})^\top z + (b_y - b_{y'})\right|}{\|w_y - w_{y'}\|_2}, \tag{5}$$

provided $w_y \neq w_{y'}$. In multiclass settings, this is the distance to the y–y' equality surface, not necessarily to a region predicted as y'.

Diagnostic workflow Candidate models are first restricted to those with comparable predictive performance. The same counterfactual search is then applied to many examples for each model. The resulting success rate, supported-success rate, and mean distance and optimization effort among successful searches form the model's counterfactual profile. Section 4 defines these metrics and describes how they are aggregated across examples. Boundary proximity, target-class support, and the curvature proxy are then used to interpret differences between profiles. For example, they indicate whether different outcomes are associated with distant boundaries, weak target-class support, or greater departure from local linearity. We then compare profiles based on the intended application. Applications seeking supported prediction changes may prefer high supported success and small displacement, whereas applications seeking prediction stability may prefer resistance to changes under the local search probe. The diagnostic workflow describes model behavior without providing a universal model ranking, and its conclusions remain conditional on the representation metric and search protocol.

4 EXPERIMENTAL SETUP

We design the experiments to test three hypotheses that our diagnostic framework (described in Section 3) motivates:

1. **H1:** Models with comparable predictive performance can exhibit substantially different counterfactual behavior.
2. **H2:** Local representation geometry, quantified by target-class support and representation curvature jointly, provides information about counterfactual behavior beyond decision-boundary proximity.
3. **H3:** Incorporating target-class support into the search objective and adapting step sizes to local representation curvature increases supported counterfactual success under matched locality and iteration constraints.

Datasets. Experiments span vision, medical imaging, text, multimodal data, and single-cell biology. **Shapes** Rudman et al. (2025) provides a controlled synthetic setting. **MNIST** LeCun et al. (1998) contains handwritten digits from ten classes. **Chest X-ray** Kermany et al. (2018) contains radiographs labeled as pneumonia or normal. **IMDb** Maas et al. (2011) contains movie reviews labeled by sentiment. **MM-IMDb** Arevalo et al. (2017) pairs movie posters with plot summaries; the experiments use a binary classification task. The **Papalexi study**, accessed through scPerturb Peidli et al. (2024); Papalexi et al. (2021), provides a real-world biological setting in which each cell has thousands of gene-expression measurements. Cells are classified by low or high immune (interferon) response, and a counterfactual asks how a low-response cell's representation must change to produce a high-response prediction. The experimentally measured perturbation response provides a **ground-truth reference** for comparing model-implied changes with the observed transcriptional transition. This enables evaluation beyond prediction change, highlights gene programs for further investigation, and tests the framework on high-dimensional, noisy measurements of a continuous biological process beyond conventional ML benchmarks.

Representations and classifier heads. Vision encoders include ResNet-50 He et al. (2016), ViT-B/16 Dosovitskiy et al. (2021), and DINOv2 ViT-B/14 Oquab et al. (2024). Language encoders include BERT Devlin et al. (2019), DistilBERT Sanh et al. (2020), and RoBERTa Liu et al. (2019). Multimodal encoders include CLIP Radford et al. (2021) and SigLIP2 Tschannen et al. (2025). Single-cell encoders include scGPT Cui et al. (2024) and Geneformer Theodoris et al. (2023). Embeddings are extracted from frozen encoders, and regularized linear heads are selected on validation data and evaluated on held-out test data. Classifier-side effects are isolated by varying head initialization and regularization while keeping embeddings fixed.

Geometry estimation. Target-class embeddings define the neighborhoods used to measure target-class support. Following Section 3, $r_k(z, c)$ is the Euclidean distance from an embedding point z to its kth nearest reference embedding from class c. We use $k = 20$. For each class, the support threshold $T_c^{(p)}$ is the pth quantile of the leave-one-out k-NN radii of its reference embeddings. With $p = 0.5$, an endpoint z' targeting class c is supported when $r_k(z', c) \leq T_c^{(0.5)}$. The local curvature proxy measures how far neighboring embeddings deviate from a fitted tangent subspace. Sensitivity analyses vary $k \in \{5, 10, 20, 50\}$ and $p \in \{0.5, 0.75, 0.9\}$. The qualitative associations and model orderings remain stable across these choices, with $k = 20$ producing the strongest associations in the tested range (Section A.1).

Baseline local search probe. For each initial embedding z_0, the target y' is the classifier's highest-scoring alternative class. The baseline uses projected-gradient search without target-class support or curvature, minimizing $\ell(z_t) = -h(z_t)_{y'} + \lambda_{\text{shift}} \|z_t - z_0\|_2$ through

$$z_{t+1} = \Pi_{\|z - z_0\|_2 \leq \tau} \left(z_t - \eta \nabla_{z_t} \ell(z_t) \right), \tag{6}$$

where $h(z_t)_{y'}$ is the target-class logit, η is the step size, and the locality radius τ bounds displacement from z_0. Search stops upon reaching y' or after $T = 300$ updates. Within each dataset, η and τ are fixed across models. Optimizer and budget sensitivity, including per-model validation tuning, are examined in Section B.3.

Evaluation metrics. For N examples, let z_i, y_i', and z_i' denote the initial embedding, target class, and returned endpoint. Let $\mathcal{S} = \{i : f(z_i') = y_i'\}$ denote successful searches. **Counterfactual success (CF-Suc)** measures the fraction reaching the target class, while **supported counterfactual success (Supp-Suc)** additionally requires endpoint support:

$$\text{CF-Suc} = \frac{|\mathcal{S}|}{N}, \tag{7}$$

$$\text{Supp-Suc} = \frac{1}{N} \sum_{i \in \mathcal{S}} \mathbf{1}\left\{ r_k(z_i', y_i') \leq T_{y_i'}^{(p)} \right\}. \tag{8}$$

Counterfactual distance (CF-Dist) and **optimization effort (OptEff)** measure mean displacement and first-target iteration among successful searches:

$$\text{CF-Dist} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \|z_i' - z_i\|_2, \tag{9}$$

$$\text{OptEff} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} t_i^*, \tag{10}$$

where t_i^* is the first iteration reaching the target class. Both quantities are undefined when $\mathcal{S}$ is empty and are interpreted alongside success rates because successful subsets can differ across models.

5 Results

The results follow the three hypotheses of the diagnostic workflow. Across domains, the findings support all three. A single-cell case study then compares representation-space counterfactuals with ground-truth perturbation responses, demonstrating the framework's applicability in a real-world biological setting.

Models with similar predictive performance can exhibit different counterfactual behavior (H1) Following the diagnostic workflow (Section 3), candidate models are first matched by predictive performance. Within each dataset, we select the pair of encoder–classifier models with the smallest validation-accuracy gap, without using counterfactual outcomes, and then compare their counterfactual profiles on held-out test examples. Table 1 reports the selected pairs. Their mean accuracy gaps are at most three percentage points, while their CF-Suc gaps range from 10 to 65 percentage points. The pairs also differ in counterfactual distance and optimization effort. This phenomenon is not limited to accuracy. Matching models by validation cross-entropy selects the same encoder pairs and preserves substantial differences in their counterfactual profiles (Section C.1).

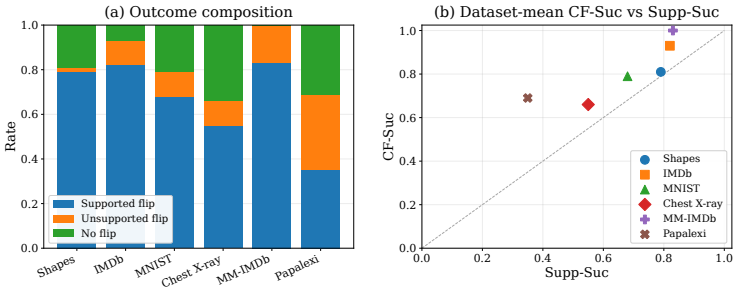

Figure 2: **Counterfactual outcomes per dataset.** (a) Fractions of test-example searches ending in a supported target-class flip, an unsupported target-class flip, or no target-class flip. (b) Supported counterfactual success (Supp-Suc) versus counterfactual success (CF-Suc).

Endpoint support provides a further distinction beyond prediction change. Successful searches whose endpoints do not meet the target-class support criterion are considered unsupported. Figure 2 shows this outcome breakdown for one representative encoder–classifier model per dataset. Chest X-ray and Papalexi have supported-success rates of 0.55 and 0.35, respectively.

Table 1: **Closest-accuracy encoder pairs within each dataset.** Pairs are selected by validation accuracy without using counterfactual outcomes. Mean absolute gaps in test accuracy, CF-Suc, CF-Dist, and OptEff are reported with standard deviations over five seeds. CF-Dist and OptEff include only successful searches.

Dataset	Model A	Model B	ACC Gap	CF-Suc Gap	CF-Dist Gap	OptEff Gap
Shapes	ResNet50	ViT	0.01 ± 0.007	0.53 ± 0.009	0.38 ± 0.007	56.55 ± 1.720
IMDb	BERT	RoBERTa	0.00 ± 0.003	0.19 ± 0.009	0.49 ± 0.007	43.75 ± 1.409
MNIST	ResNet50	ViT	0.01 ± 0.001	0.65 ± 0.002	0.23 ± 0.001	46.72 ± 0.054
Chest X-ray	ResNet50	ViT	0.01 ± 0.029	0.27 ± 0.044	0.16 ± 0.035	29.39 ± 3.096
MM-IMDb	CLIP	SigLIP2	0.02 ± 0.002	0.10 ± 0.004	0.11 ± 0.003	6.34 ± 1.047
Papalexi	scGPT	Geneformer	0.03 ± 0.007	0.56 ± 0.004	0.24 ± 0.003	13.07 ± 2.497

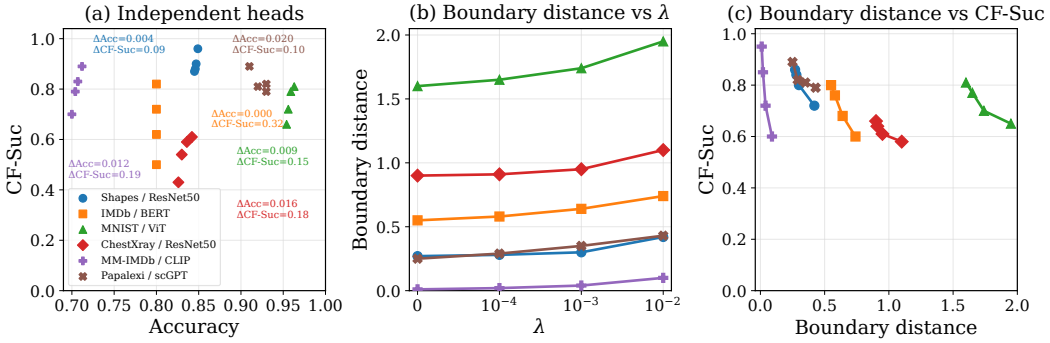

Figure 3: **Classifier boundary placement changes counterfactual behavior under fixed representations.** With encoder embeddings fixed, varying the classifier head changes (a) CF-Suc despite similar accuracy and (b) mean boundary distance through regularization. (c) Larger boundary distance is associated with lower CF-Suc within each dataset.

Together, the accuracy-matched comparisons support H1. Similarly accurate models can differ in counterfactual search success. Figure 2 shows separately that a successful prediction change need not reach a region supported by observed target-class examples.

Local representation geometry provides information beyond boundary proximity (H2) Counterfactual differences may arise from either the representation geometry or the classifier boundary. To isolate classifier-side effects, each representation is held fixed while classifier heads are varied through initialization and regularization. Among the configurations shown in Figure 3a, CF-Suc differs by as much as 32 percentage points despite accuracy differences of at most 2 percentage points. Across the broader set of fixed-representation experiments, CF-Suc ranges reach 59 percentage points (Table 7). Varying regularization changes boundary placement (Figure 3b), and larger boundary distances are associated with lower CF-Suc within each dataset (Figure 3c). Additional experiments with maximum-margin and nonlinear heads, as well as partial fine-tuning, show similar patterns. Changes in classifier boundaries or representations can substantially alter counterfactual behavior without corresponding changes in predictive performance (Sections C.3 and C.4).

These results establish the importance of classifier boundary placement. The next analysis examines whether local representation geometry provides additional information about counterfactual behavior. Before evaluating its predictive value, the curvature proxy is validated against local departure from linearity. A tangent subspace is estimated from a subset of neighboring embeddings and used to approximate held-out neighbors that were not involved in its estimation. Approximation error increases from 0.11 in the lowest-curvature quartile to 0.28 in the highest, indicating that larger values of $\kappa_k(z)$ identify regions that are less accurately described by a local linear approximation (Figure 4a). Adding target-class support and the curvature proxy to boundary proximity then improves prediction of counterfactual outcomes on held-out examples. Mean R^2 increases by 0.10 for CF-Suc and 0.07 for CF-Dist (Figure 4b–c). These associations remain stable across the tested neighborhood sizes,

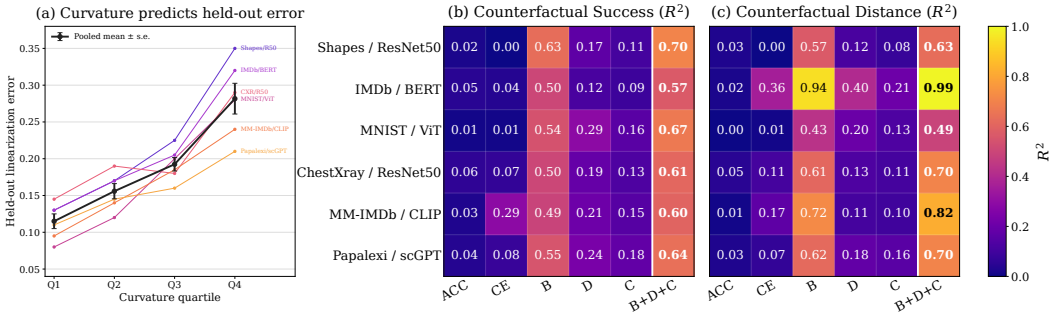

Figure 4: **Representation geometry adds information beyond classifier-side boundary proximity.** (a) Held-out linearization error increases with the curvature proxy κ_k. (b–c) Held-out CF-Suc and CF-Dist prediction across models and datasets using accuracy (ACC), cross-entropy (CE), boundary proximity (B), target-class k-NN radius (D), curvature (C), and their combinations. Adding D and C to B improves prediction.

Table 2: **Geometry-scaled search ablation.** CF-Suc and Supp-Suc are averaged equally across the six datasets. Variants use matched locality radii and iteration budgets.

Variant	Support	Curvature step	CF-Suc	Supp-Suc
Baseline	×	×	0.81	0.67
Support only	✓	×	0.87	0.76
Curvature only	×	✓	0.88	0.71
Combined	✓	✓	**0.91**	**0.79**

support thresholds, and scale normalizations, although the curvature proxy remains dependent on representation scale and parameterization (Sections A.1 and A.2).

Together, these results support H2: target-class support and curvature provide information beyond boundary proximity. This matters because crossing a nearby boundary does not ensure a data-supported alternative, a distinction that can inform model comparison and search.

Geometry-scaled search increases supported success (H3) H3 tests whether target-class support and curvature can improve search. For target class y' with reference embeddings $\mathcal{R}_{y'}$, geometry-scaled search adds a support term to the baseline objective:

$$\ell(z_t) = \ell_{\text{clf}}(z_t) + \lambda_{\text{shift}}\|z_t - z_0\|_2 + \frac{\lambda_{\text{knn}}}{k} \sum_{u \in \text{kNN}_k(z_t; \mathcal{R}_{y'})} \|z_t - u\|_2^2. \tag{11}$$

The term encourages proximity to target-class examples. Endpoint support is evaluated separately using the criterion in Section 4. Search retains the baseline locality radius and adapts its step size as

$$\eta_t = \frac{\eta_0}{1 + \alpha \widetilde{\kappa}_k(z_t)}, \qquad \widetilde{\kappa}_k(z_t) = \frac{\kappa_k(z_t)}{\text{median}_{u \in \mathcal{V}}\, \kappa_k(u) + \varepsilon}, \tag{12}$$

where $\mathcal{V}$ is the validation split. Higher curvature reduces the step size.Table 2 isolates the two components under matched locality radii and iteration budgets. The support term improves supported success, curvature-dependent steps improve reachability, and their combination performs best on both metrics. Per-dataset results and budget sensitivity appear in Sections B.3 and B.5. These gains show that geometric diagnostics can guide search toward more data-supported prediction changes.

Real-world single-cell perturbations H3 shows that geometry-scaled search reaches better-supported endpoints, but representation-space support alone does not show whether those endpoints reflect observed changes in gene expression. The Papalexi Perturb-seq dataset includes experimentally measured perturbation responses, providing a ground-truth reference for this comparison. Starting from a low-response cell, a counterfactual estimates the expression changes the model associates with a stronger interferon response, so counterfactual search starts from low-response cell embeddings

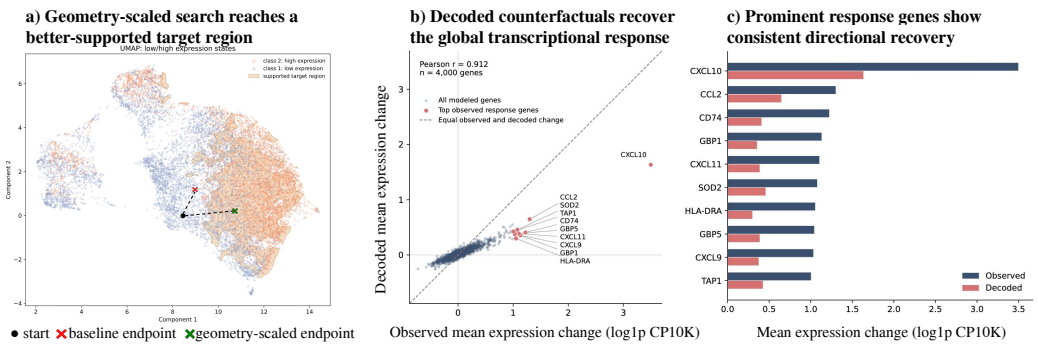

Figure 5: **Geometry-scaled search improves endpoint support and recovers structure in an observed transcriptional transition.** (a) Starting from the same cell embedding, geometry-scaled search reaches a more strongly supported target region than baseline search. (b) Across $4{,}000$ genes, decoded expression changes align with the observed population-level difference between low- and high-response cells. (c) Selected response genes generally change in the observed direction, although the predicted changes have smaller magnitudes.

and targets the high-response class. Comparing those changes with the observed response tests whether the counterfactual captures the biological pattern and suggests gene programs for further study. We use frozen embeddings from scGPT Cui et al. (2024), a widely used pretrained single-cell foundation model, for cells in the Papalexi Perturb-seq dataset Papalexi et al. (2021). Search starts from low-response cells and targets the high-response class. An MLP decoder trained on reference cells predicts log-normalized expression for $4{,}000$ selected genes from each embedding. For a starting embedding z_0 and counterfactual endpoint z', the difference $d(z') - d(z_0)$ estimates the expression change induced by search. We compare this change with the experimentally observed population-level difference (Section B.1). Baseline and geometry-scaled search use the same encoder–classifier pair, starting cells, targets, displacement penalty, locality radius, and iteration budget.

Geometry-scaled search reaches regions closer to observed target-class cells and increases Supp-Suc from 30.1% to 62.4% (Figure 5a). Across $4{,}000$ genes, its decoded changes correlate with the observed population-level difference at $r = 0.91$, compared with $r = 0.74$ for baseline search, and agree in direction for 76.2% of genes, compared with 62.7%. They recover 75% of the top 100 observed upregulated genes and 67% of the top 100 observed downregulated genes, compared with 64.5% and 56.8%, respectively (Figure 5b). Figure 5c highlights genes experimentally associated with the high interferon-response state and shows how their decoded changes compare with the observed changes.

Together, these results show that the framework serves both as a model diagnostic and as a practical tool for guiding counterfactual search in a real-world settings.

6 Conclusion

This work studies counterfactual behavior through the lens of representation geometry. Across diverse domains, results show that models with similar predictive performance differ substantially in counterfactual success, distance, optimization effort, and endpoint support. Changing only the classifier head alters these outcomes under fixed representations, while target-class support and curvature provide information beyond boundary proximity. Incorporating these geometric quantities into search increases supported counterfactual success under matched locality and iteration constraints. In a single-cell case study, decoded counterfactual changes align with observed differences between low- and high-response cells, demonstrating the framework's use in a real biological setting. Together, these findings show how local representation geometry can guide model choice and search toward counterfactuals that change predictions while remaining near observed target-class examples. Extending the results to jointly trained models remain directions for future work. Further limitations appear in Appendix D.

AI USE STATEMENT

LLM tools assisted with manuscript editing, figure preparation, and code formatting. They were not used as a component of the proposed method. All AI-assisted material was reviewed and verified by the authors, who take responsibility for the full content of the submission.

ETHICS STATEMENT

The experiments use previously published datasets, including de-identified medical images and single-cell gene-expression measurements, and introduce no new collection of human-subject data. The single-cell counterfactuals may suggest gene programs for experimental follow-up, but their decoded changes require biological validation. Representation-space counterfactuals reflect associations learned by a model and do not establish causal or realizable interventions for individual cells or patients. The biomedical results are not clinical recommendations. The diagnostics support model comparison and auditing, but could also reveal model weaknesses or decision boundaries.

REPRODUCIBILITY STATEMENT

The main text defines the geometric quantities, search protocol, evaluation metrics, and geometry-scaled search method. The appendix provides dataset and preprocessing details, model configurations, hyperparameters, ablations, sensitivity analyses, and additional results. Code is available at https://anonymous.4open.science/r/cf-geometry-E82B/

REFERENCES

Eddie Aamari and Clément Levrard. Non-asymptotic rates for manifold, tangent space, and curvature estimation, 2018. URL https://arxiv.org/abs/1705.00989.

Antonio G. Ache and Micah W. Warren. Ricci curvature and the manifold learning problem. *Advances in Mathematics*, 342:14–66, 2019. ISSN 0001-8708. doi: https://doi.org/10.1016/j.aim.2018.11.001. URL https://www.sciencedirect.com/science/article/pii/S000187081830450X.

John Arevalo, Thamar Solorio, Manuel Montes y Gómez, and Fabio A. González. Gated multimodal units for information fusion, 2017. URL https://arxiv.org/abs/1702.01992.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, August 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2013.50. URL https://doi.org/10.1109/TPAMI.2013.50.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, NY, 2006. ISBN 978-0-387-31073-2. URL https://www.microsoft.com/en-us/research/wp-content/uploads/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf.

SueYeon Chung, Daniel D. Lee, and Haim Sompolinsky. Classification and geometry of general perceptual manifolds. *Physical Review X*, 8(3), July 2018. ISSN 2160-3308. doi: 10.1103/physrevx.8.031003. URL http://dx.doi.org/10.1103/PhysRevX.8.031003.

Uri Cohen, SueYeon Chung, Daniel D. Lee, and Haim Sompolinsky. Separability and geometry of object manifolds in deep neural networks. *Nature Communications*, 11(1), February 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-14578-5. URL http://dx.doi.org/10.1038/s41467-020-14578-5.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(8):1470–1480, February 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02201-0. URL http://dx.doi.org/10.1038/s41592-024-02201-0.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh,

Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning, 2020. URL https://arxiv.org/abs/2011.03395.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Ann-Kathrin Dombrowski, Maximilian Alber, Christopher J. Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame, 2019. URL https://arxiv.org/abs/1906.07983.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Gamaleldin Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/42998cf32d552343bc8e460416382dca-Paper.pdf.

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 55–65, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1006. URL https://aclanthology.org/D19-1006/.

Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Analysis of classifiers' robustness to adversarial perturbations. *Machine Learning*, 107(3):481–508, August 2017. ISSN 1573-0565. doi: 10.1007/s10994-017-5663-3. URL http://dx.doi.org/10.1007/s10994-017-5663-3.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously, 2019. URL https://arxiv.org/abs/1801.01489.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015. URL https://arxiv.org/abs/1412.6572.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Yiding Jiang, Dilip Krishnan, Hossein Mobahi, and Samy Bengio. Predicting the generalization gap in deep networks with margin distributions, 2019. URL https://arxiv.org/abs/1810.00113.

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

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 353–362, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445899. URL https://doi.org/10.1145/3442188.3445899.

Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C.S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K. Prasadha, Jacqueline Pei, Magdalene Y.L. Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Runze Zhang, Lianghong Zheng, Rui Hou, William Shi, Xin Fu, Yaou Duan, Viet A.N. Huu, Cindy Wen, Edward D. Zhang, Charlotte L. Zhang, Oulan Li, Xiaobo Wang, Michael A. Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M. Anthony Lewis, Huimin Xia, and Kang Zhang. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131.e9, 2018. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2018.02.010. URL https://www.sciencedirect.com/science/article/pii/S0092867418301545.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Zb6c8A-Fghk.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited, 2019. URL https://arxiv.org/abs/1905.00414.

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. The dangers of post-hoc interpretability: Unjustified counterfactual explanations, 2019. URL https://arxiv.org/abs/1907.09294.

Yann LeCun, Corinna Cortes, and Christopher J. C. Burges. The mnist database of handwritten digits. http://yann.lecun.com/exdb/mnist/, 1998.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015.

Charles T. Marx, Flavio du Pin Calmon, and Berk Ustun. Predictive multiplicity in classification, 2020. URL https://arxiv.org/abs/1909.06677.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks, 2016. URL https://arxiv.org/abs/1511.04599.

Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 607–617, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Efthymia Papalexi, Eleni P. Mimitou, Andrew W. Butler, Samantha Foster, Bernadette Bracken, William M. Mauck, Hans-Hermann Wessels, Yuhan Hao, Bertrand Z. Yeung, Peter Smibert, and Rahul Satija. Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nature Genetics*, 53(3):322–331, March 2021. ISSN 1546-1718. doi: 10.1038/s41588-021-00778-2. URL http://dx.doi.org/10.1038/s41588-021-00778-2.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020. doi: 10.1073/pnas.2015509117. URL https://www.pnas.org/doi/abs/10.1073/pnas.2015509117.

Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of The Web Conference 2020*, WWW '20, pp. 3126–3132. ACM, April 2020a. doi: 10.1145/3366423.3380087. URL http://dx.doi.org/10.1145/3366423.3380087.

Martin Pawelczyk, Klaus Broelemann, and Gjergji. Kasneci. On counterfactual explanations under predictive multiplicity. In Jonas Peters and David Sontag (eds.), *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pp. 809–818. PMLR, 03–06 Aug 2020b. URL https://proceedings.mlr.press/v124/pawelczyk20a.html.

Martin Pawelczyk, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4574–4594. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/pawelczyk22a.html.

Paraskevas Pegios, Aasa Feragen, Andreas Abildtrup Hansen, and Georgios Arvanitidis. Counterfactual explanations via riemannian latent space traversal, 2024. URL https://arxiv.org/abs/2411.02259.

Stefan Peidli, Tessa D. Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J. Schumacher, Jake P. Taylor-King, Debora S. Marks, Augustin Luna, Nils Blüthgen, and Chris Sander. scperturb: harmonized single-cell perturbation data. *Nature Methods*, 21(3):531–540, January 2024. ISSN 1548-7105. doi: 10.1038/s41592-023-02144-y. URL http://dx.doi.org/10.1038/s41592-023-02144-y.

Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. Face: Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pp. 344–350, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375850. URL https://doi.org/10.1145/3375627.3375850.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European Signal Processing Conference*, pp. 606–610, 2007.

William Rudman, Michal Golovanevsky, Amir Bar, Vedant Palit, Yann LeCun, Carsten Eickhoff, and Ritambhara Singh. Forgotten polygons: Multimodal large language models are shape-blind. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11983–11998, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.620. URL https://aclanthology.org/2025.findings-acl.620/.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. URL https://arxiv.org/abs/1910.01108.

Thomas Tanay and Lewis Griffin. A boundary tilting perspective on the phenomenon of adversarial examples, 2016. URL https://arxiv.org/abs/1608.07690.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, May 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06139-9. URL http://dx.doi.org/10.1038/s41586-023-06139-9.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Asterios Tsiourvas, Wei Sun, and Georgia Perakis. Manifold-aligned counterfactual explanations for neural networks. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3763–3771. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/tsiourvas24a.html.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 10–19. ACM, January 2019. doi: 10.1145/3287560.3287566. URL http://dx.doi.org/10.1145/3287560.3287566.

Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan Hines, John Dickerson, and Chirag Shah. Counterfactual explanations and algorithmic recourses for machine learning: A review. *ACM Comput. Surv.*, 56(12), October 2024. ISSN 0360-0300. doi: 10.1145/3677119. URL https://doi.org/10.1145/3677119.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box. *Harvard Journal of Law & Technology*, 31(2):842–887, 2018.

Eslam Zaher, Maciej Trzaskowski, Quan Nguyen, and Fred Roosta. Counterfactual explanations on robust perceptual geodesics, 2026. URL https://arxiv.org/abs/2601.18678.

Javier Álvarez Vizoso, Michael Kirby, and Chris Peterson. Local eigenvalue decomposition for embedded riemannian manifolds. *Linear Algebra and its Applications*, 604:21–51, 2020. ISSN 0024-3795. doi: https://doi.org/10.1016/j.laa.2020.06.006. URL https://www.sciencedirect.com/science/article/pii/S0024379520302974.

A Geometric Quantities and Estimation

A.1 Target-Class Support and Density Estimation

Local target-class density is measured using the k-nearest-neighbor (k-NN) radius in representation space. For target class c, let n_c denote the number of reference embeddings from class c, and let $r_k(z, c)$ denote the distance from z to its kth nearest target-class neighbor. This quantity is the adaptive radius appearing in the classical k-NN density estimator Bishop (2006):

$$\widehat{p}_c(z) = \frac{k}{n_c V_d r_k(z, c)^d}, \tag{13}$$

where V_d is the volume of the d-dimensional unit ball. For fixed k, n_c, and representation dimension d, the estimated density decreases monotonically with $r_k(z, c)^d$. Thus, although $r_k(z, c)$ is distance-valued, it has a direct density interpretation: a smaller radius means that k target-class examples occupy a smaller local volume, corresponding to higher estimated target-class density. For any fixed k, ranking points by $r_k(z, c)$ is therefore exactly the reverse of ranking them by $\widehat{p}_c(z)$.

The baseline counterfactual search does not optimize $r_k(z, c)$ or the corresponding density estimate. Instead, $r_k(z, c)$ is used as a diagnostic of the target-class data available near the source or counterfactual endpoint. Since changing k changes the spatial scale at which density is estimated, the following analyses test whether the conclusions depend on this choice.

Sensitivity to neighborhood size We vary $k \in \{5, 10, 20, 50\}$ and measure (i) the Spearman correlation between $r_k(z)$ and CF-Suc, and (ii) the held-out R^2 of the post hoc models in Section 5 that use decision-boundary proximity and local target-class density as features. Results are summarized in Figure 6.

Across datasets, both quantities vary only mildly across neighborhood sizes. The curves remain close over the tested range, with $k = 20$ giving the strongest values in this sweep. These results suggest that the main trends are not sensitive to the precise neighborhood scale used to estimate local density.

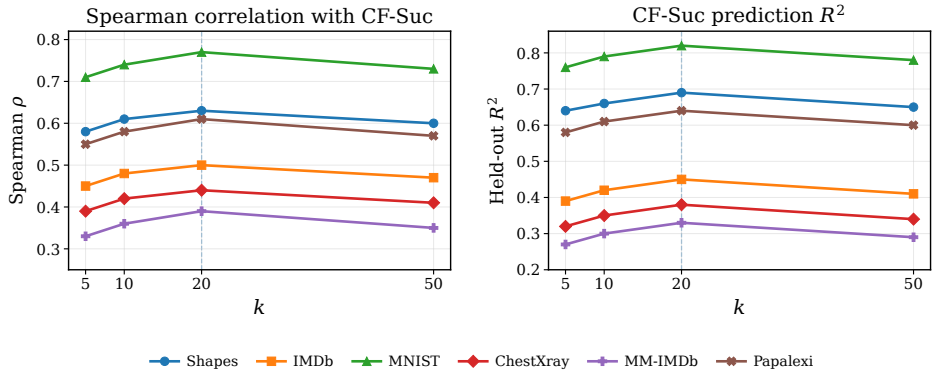

Figure 6: **Sensitivity to neighborhood size.** Spearman correlation between target-class k-NN radius and CF-Suc (left), and held-out R^2 for CF-Suc prediction using boundary proximity and target-class k-NN radius (right), across $k \in \{5, 10, 20, 50\}$. Results vary mildly across the tested neighborhood sizes, with $k = 20$ strongest in this sweep.

Supported-counterfactual threshold To determine whether a successful counterfactual terminates in a region that is representative of observed target-class data, a class-specific density threshold is defined from the reference split. For each class c, let

$$\mathcal{D}_c = \{r_k(z_i, c) : z_i \in \mathcal{R}_c\} \tag{14}$$

denote the leave-one-out k-NN radii of reference embeddings from class c. The support threshold is

$$T_c^{(p)} = \text{Quantile}_p(\mathcal{D}_c). \tag{15}$$

A successful counterfactual z' targeting class c is considered supported when

$$r_k(z', c) \leq T_c^{(p)}. \tag{16}$$

The main experiments use $k = 20$ and $p = 0.5$, so a counterfactual is supported when its local target-class density is at least as high as that of the median reference example from the target class. The class-specific threshold accounts for differences in scale and concentration across classes.

Sensitivity to this criterion is evaluated using $p \in \{0.5, 0.75, 0.9\}$. Although supported-flip rates increase as the threshold becomes less restrictive, the qualitative conclusions and model ordering remain stable across thresholds.

Table 3: **Sensitivity to the supported-counterfactual threshold.** Supported-flip rates under increasingly permissive class-specific density thresholds.

Support threshold p	IMDb	MNIST
0.50	0.62	0.65
0.75	0.69	0.68
0.90	0.76	0.74

Covariance-aware density estimate The Euclidean k-NN radius assumes an isotropic local distance metric. To test whether the conclusions depend on this assumption, we additionally evaluate a class-conditional Mahalanobis support measure. For each target class, a counterfactual is considered supported when its Mahalanobis distance to the target-class distribution is below the median within-class Mahalanobis distance on the reference split. Under a class-conditional Gaussian model, smaller Mahalanobis distance corresponds to higher class-conditional density.

The resulting support assignments agree strongly with the Euclidean k-NN criterion: agreement is 0.996 on MNIST, 0.950 on MM-IMDb, and 0.770 on IMDb. The qualitative conclusions are preserved, suggesting that the observed relationship between target-class density and counterfactual behavior is not specific to the Euclidean k-NN estimator.

Scale normalization To account for differences in representation scale, we evaluate normalized variants of $r_k(z)$, including (i) standardization by global feature variance and (ii) normalization by the median pairwise distance within each dataset. Both transformations leave correlations and R^2 values effectively unchanged (differences < 0.02), suggesting that the observed relationships are not driven by representation scale.

A.2 Local Representation Curvature

Tangent-subspace estimation Although the ambient representation space is Euclidean, observed embeddings may occupy a curved lower-dimensional structure. For each query z, let $\mathcal{N}_k(z)$ denote its k nearest reference embeddings, irrespective of class. Following local tangent-space estimation Aamari & Levrard (2018), the matrix with rows $(u - z)^\top$, for $u \in \mathcal{N}_k(z)$, is decomposed using singular value decomposition. The leading m right singular vectors form an orthonormal basis V_m for the estimated tangent subspace T_z, with projections

$$P_{T_z} = V_m V_m^\top, \qquad P_{T_z^\perp} = I - V_m V_m^\top. \tag{17}$$

The neighborhood is centered at the query z, so the resulting subspace approximates the surrounding structure through that point.

Curvature proxy and interpretation The proxy in Equation (3) divides the mean normal residual by the mean squared tangential displacement, with a positive stabilizer in the denominator. This construction is motivated by the second-order normal departure of a smooth manifold from its tangent approximation Álvarez Vizoso et al. (2020). Larger values indicate greater normal departure relative to the tangential extent of the neighborhood.

The quantity is a finite-sample geometric diagnostic rather than an intrinsic differential-geometric curvature estimate. Normal residuals can also reflect noise, heterogeneous neighborhoods, or an

imperfect choice of tangent dimension. For diagnostic prediction, curvature is evaluated at the initial embedding. For geometry-scaled search, it is recomputed at the current iterate. Away from observed data, it may additionally reflect displacement from the local data structure and is interpreted as an approximation-residual proxy.

Held-out validation To assess whether the proxy captures departure from local linearity, we compare it with linearization error on neighboring points excluded from both tangent fitting and proxy computation. This separation evaluates how well the estimated local structure extends to points not used to construct the diagnostic.

Neighborhood size and tangent dimension Neighborhood size determines the spatial scale of the approximation, while tangent dimension determines which directions are treated as tangential or normal. An underspecified tangent subspace can assign genuine variation to the normal component, whereas an overly large subspace can absorb departures from local linearity. Consequently, the proxy must be interpreted relative to both choices, and sensitivity analyses should assess whether its associations with counterfactual outcomes persist across them.

Representation scale The proxy is not invariant to representation scale. Under uniform rescaling $z \mapsto az$ with $a > 0$, normal residuals scale by a and squared tangential displacements by a^2. Ignoring the stabilizer, the proxy therefore scales as a^{-1}. The validation-median normalization in Equation (12) approximately cancels this common factor when the stabilizers are negligible. It does not establish invariance to anisotropic rescaling or general reparameterization. The diagnostic therefore characterizes local structure under the chosen representation metric.

B Experimental Methods

B.1 Datasets, Tasks, and Data Splits

Shapes Shapes Rudman et al. (2025) is a synthetic image classification dataset containing simple geometric objects on a uniform background. Shape type, size, position, and color are systematically controlled, yielding low intra-class variation. It provides a useful controlled setting for studying counterfactual search and representation geometry.

MNIST MNIST LeCun et al. (1998) consists of grayscale 28×28 images of handwritten digits from 0 to 9. The dataset is visually simple but includes natural variation in handwriting style and stroke shape. It serves as a standard benchmark between synthetic and more complex real-image settings.

Chest X-ray (Pneumonia) The Chest X-ray dataset Kermany et al. (2018) contains frontal chest radiographs labeled as pneumonia or normal. Visual differences between classes are subtle and can be affected by imaging variability. This makes it a challenging medical-imaging setting for counterfactual analysis.

IMDb IMDb Maas et al. (2011) contains $50{,}000$ movie reviews evenly split between positive and negative sentiment. The task is defined over high-dimensional language representations capturing lexical and semantic structure. It provides a text-domain benchmark for studying whether counterfactual behavior varies across language encoders with similar predictive performance.

MM-IMDb MM-IMDb Arevalo et al. (2017) contains approximately $15{,}600$ movie posters paired with plot summaries. The original multi-label task is restricted here to binary classification for controlled comparison. It enables evaluation across unimodal image models, unimodal text models, pretrained multimodal encoders, and simple fusion methods.

Papalexi dataset The Papalexi Perturb-seq dataset Papalexi et al. (2021), accessed through scPerturb Peidli et al. (2024), contains single-cell gene-expression measurements with genetic perturbation annotations. The case study examines low-to-high interferon-response prediction changes, enabling evaluation of counterfactual support in biological representations and comparison with observed population-level expression differences.

Decoding representation-space counterfactuals An MLP decoder $d : \mathcal{Z} \to \mathbb{R}^G$ is trained on paired frozen scGPT embeddings and log-normalized expression profiles from the reference split. Its output contains expression values for $G = 4{,}000$ selected variable genes. The decoder is trained without the test cells used in the case study and is held fixed for both baseline and geometry-scaled search.

For a low-response starting cell i with embedding z_i and counterfactual endpoint z'_i, the decoded gene-wise change is

$$\widehat{\Delta}_{i,g} = d(z'_i)_g - d(z_i)_g. \tag{18}$$

Decoding both embeddings places them on the same output scale and avoids adding the starting cell's reconstruction error by subtracting its observed expression from the decoded endpoint.

Let $\mathcal{I}$ denote the low-response cells included in the gene-level evaluation. Their mean predicted change for gene g is

$$\widehat{\Delta}_g = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \widehat{\Delta}_{i,g}. \tag{19}$$

The experimental reference Δ_g^{obs} averages observed high-minus-low differences in log-normalized expression within matched sample and perturbation groups, weighted by the number of eligible starting cells in each group. Both vectors cover the same G genes. We compare them using correlation across genes, agreement in change direction, and recovery of genes with the largest observed changes.

Data partitions and reference sets. Classifier heads are selected using validation data and evaluated on held-out test data. Class-conditional reference embeddings define target-class neighborhoods and support thresholds, with thresholds computed from leave-one-out k-nearest-neighbor radii. MM-IMDb is restricted to a binary classification task.

B.2 Representations and Classifier Training

For each dataset, a linear classifier is trained on top of frozen pretrained embeddings. All encoders remain frozen throughout training to isolate the effects of representation geometry.

Classifier heads are trained with cross-entropy loss using Adam. Unless otherwise specified, training uses a learning rate of 10^{-3}, batch size 128, and at most 100 epochs. Early stopping is based on validation accuracy.

For experiments with multiple classifier heads, the encoder and dataset are held fixed while heads are retrained from independent random initializations. For controlled boundary-shift experiments, the ℓ_2 weight-decay parameter is varied while all other training settings remain fixed. Hyperparameters are kept consistent within each dataset to enable fair comparison across encoders and heads.

Feature extraction and normalization All encoders remain frozen, and their embeddings are cached before classifier training and counterfactual analysis. For torchvision-based vision encoders, images are resized to 224×224, converted to RGB, and normalized using ImageNet channel statistics. Processor-based encoders use their pretrained image processors. No stochastic image augmentation is applied.

Classifier-head comparisons For each dataset–encoder pair, classifier heads are trained on the same frozen embeddings from independent random initializations. Regularization experiments vary the ℓ_2 weight-decay coefficient while keeping the remaining training settings fixed. These comparisons characterize variation in counterfactual behavior across classifier heads without changing the underlying representations.

B.3 Local Search and Optimization Sensitivity

The main experiments use a standardized local search probe in representation space, with a shared step size and locality radius across models within each dataset and a maximum budget of $T = 300$ iterations. The probe measures finite-budget search behavior without assuming globally optimal or realizable input-space counterfactuals.

Objective and updates For an initial embedding z_0 with predicted class $y = f(z_0)$, the target is the highest-scoring alternative class,

$$y' = \arg\max_{c \neq y} h(z_0)_c, \tag{20}$$

where $h(z)$ denotes the classifier logits. The baseline minimizes

$$\ell(z_t) = -h(z_t)_{y'} + \lambda_{\text{shift}} \|z_t - z_0\|_2, \tag{21}$$

using projected-gradient updates

$$z_{t+1} = \Pi_{\|z - z_0\|_2 \leq \tau} \left(z_t - \eta \nabla_{z_t} \ell(z_t) \right). \tag{22}$$

Here, η is the step size, τ is the locality radius, and Π is Euclidean projection onto the ball centered at z_0. Search terminates when $f(z_t) = y'$ or the iteration budget is exhausted. A prediction change to another class does not constitute targeted success.

Shared search configuration Within each dataset, the same η and τ are applied across candidate models on the held-out test split. These settings define a common diagnostic probe, although equal numerical radii need not represent equivalent geometric constraints across embedding spaces. Targets are selected separately for each classifier and can therefore differ across models.

Search-parameter sensitivity The optimizer (SGD, Adam, AdamW), step size, locality radius, and iteration budget are varied individually from the default configuration. Figure 7 reports Spearman rank correlations between model-level CF-Suc under the default and alternative settings. Model rankings remain broadly stable, with optimizer changes producing smaller effects than changes to the step size or locality radius. Agreement decreases with shorter iteration budgets, while the observed trends stabilize by $T = 300$.

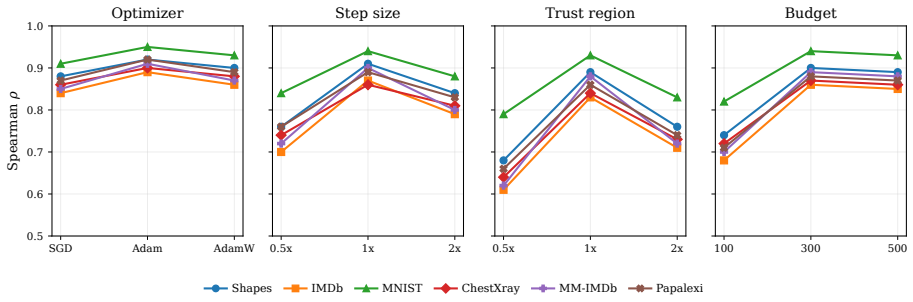

Figure 7: **Sensitivity of model rankings to search settings.** Spearman rank correlation of model-level CF-Suc between the default configuration and alternative settings. Rankings are less sensitive to optimizer choice than to step size or locality radius, and larger iteration budgets yield similar rankings.

Per-model validation tuning As an additional sensitivity analysis, η and τ are selected separately for each model using validation data. Configurations are selected by highest validation CF-Suc, with ties broken by lowest validation CF-Dist, and then fixed for test evaluation. The maximum iteration budget remains $T = 300$. Table 4 compares this procedure with the shared configuration used in the main analysis. Per-model tuning increases CF-Suc on five of the six datasets, with MM-IMDb already at 1.00 under the shared configuration. The largest increase is on Papalexi (0.69 to 0.82), followed by Shapes (0.81 to 0.90) and MNIST (0.79 to 0.88). CF-Dist decreases on most datasets but increases slightly on IMDb. Absolute counterfactual metrics therefore depend on search configuration. These comparisons complement the ranking sensitivity analysis in Figure 7.

B.4 Diagnostic Prediction and Evaluation Splits

Post hoc predictors assess whether local representation geometry provides information about counterfactual outcomes beyond boundary proximity. Target-class support $r_k(z, y')$ and curvature $\kappa_k(z)$ are computed at the initial embedding, so the predictors do not use information from the generated endpoint.

Table 4: **Shared versus per-model validation-tuned search parameters.** For each dataset, CF-Suc and CF-Dist are averaged across the same evaluated encoder–classifier models. The shared configuration is the baseline used in the geometry-scaled search ablation. Tuned settings select the step size and locality radius separately for each model on the validation split. Results are evaluated on the held-out test split.

Dataset	CF-Suc (Shared)	CF-Suc (Tuned)	CF-Dist (Shared)	CF-Dist (Tuned)
Shapes	0.81	0.90	0.30	0.27
IMDb	0.93	1.00	0.32	0.35
MNIST	0.79	0.88	0.92	0.79
Chest X-ray	0.66	0.69	0.75	0.68
MM-IMDb	1.00	1.00	0.47	0.34
Papalexi	0.69	0.82	0.45	0.40

Predictors and outcomes The analyses evaluate predictive accuracy, cross-entropy, boundary proximity, target-class support, curvature, and combinations of these quantities for predicting counterfactual success and distance. Model and dataset identity are included as controls. Counterfactual distance is evaluated only for successful searches, consistent with its definition in Section 4.

Held-out evaluation Post hoc predictors are fitted on one partition of the counterfactual results and evaluated on a disjoint held-out partition. Performance is reported as R^2 on the held-out outcomes. The incremental contribution of representation geometry is the difference in held-out R^2 between a predictor using boundary proximity and an otherwise identical predictor that also uses target-class support and curvature.

B.5 Geometry-Scaled Search and Ablation Details

Geometry-scaled search modifies the baseline in two ways. A target-support term encourages proximity to observed target-class examples, and a curvature-dependent step size adjusts the search trajectory. The ablation evaluates each component separately and then combines them. All four variants use the same examples, targets, initializations, displacement penalty, locality radius, and iteration budget within each model.

Search objective Let z_0 be the initial representation, y' the target class, and $\mathcal{R}_{y'}$ the target-class reference embeddings. All variants minimize a target-class objective with a displacement penalty:

$$\ell_{\text{base}}(z) = -h(z)_{y'} + \lambda_{\text{shift}}\|z - z_0\|_2. \tag{23}$$

Here, $h(z)_{y'}$ is the target-class logit. The displacement term encourages minimal changes and has the same weight in every ablation variant.

The support-only and combined variants add a target-support term:

$$\ell_{\text{support}}(z) = \ell_{\text{base}}(z) + \frac{\lambda_{\text{knn}}}{k} \sum_{u \in \text{kNN}_k(z;\mathcal{R}_{y'})} \|z - u\|_2^2. \tag{24}$$

The neighbor term encourages proximity to the k nearest observed target-class embeddings. We use $k = 20$, matching the endpoint-support diagnostic. It guides optimization but does not define a supported counterfactual. Endpoint support is evaluated separately using the class-specific k-NN radius threshold in Section 4. A small average squared distance to nearby target examples does not guarantee that the kth-neighbor distance satisfies that threshold.

Curvature-dependent updates The curvature-only and combined variants use the projected update in Equation (6) with the adaptive step size

$$\eta_t = \frac{\eta_0}{1 + \alpha\widetilde{\kappa}_k(z_t)}, \qquad \widetilde{\kappa}_k(z_t) = \frac{\kappa_k(z_t)}{\text{median}_{u \in \mathcal{V}}\, \kappa_k(u) + \varepsilon}, \tag{25}$$

where $\mathcal{V}$ is the validation set used to calibrate the curvature scale. The baseline and support-only variants use a constant step size, corresponding to $\alpha = 0$. For $\alpha \geq 0$, larger values of the curvature

Table 5: **Per-dataset ablation of geometry-scaled search.** CF-Suc and Supp-Suc results for the four search variants. All variants use matched examples, targets, initializations, displacement penalties, locality radii, and iteration budgets within each model.

	Baseline		Support only		Curvature only		Combined	
Dataset	CF-Suc	Supp-Suc	CF-Suc	Supp-Suc	CF-Suc	Supp-Suc	CF-Suc	Supp-Suc
Shapes	0.81	0.79	0.96	0.86	0.97	0.84	0.98	0.88
MNIST	0.79	0.68	0.94	0.84	0.95	0.82	0.97	0.86
Chest X-ray	0.66	0.55	0.68	0.62	0.70	0.58	0.76	0.65
IMDb	0.93	0.82	0.90	0.84	0.91	0.80	0.96	0.83
MM-IMDb	1.00	0.83	1.00	0.85	1.00	0.82	1.00	0.90
Papalexi	0.69	0.35	0.74	0.55	0.75	0.40	0.79	0.62

proxy yield smaller updates. Curvature is recomputed at each iterate, while projection keeps every variant within the same locality radius τ.

The proxy in Equation (3) measures normal departure from a locally fitted tangent subspace. Along a search trajectory, a large value may reflect local nonlinearity or departure from the structure occupied by reference embeddings. Step-size adaptation uses this quantity without treating it as an intrinsic curvature estimate.

Ablation variants The baseline minimizes Equation (23) with a constant step size. Support-only search minimizes Equation (24) with a constant step size. Curvature-only search minimizes Equation (23) with the adaptive step size. Combined search minimizes Equation (24) with the adaptive step size. Within each model, all variants use the same initial examples, targets, displacement penalty, locality radius, and maximum budget of $T = 300$ iterations. Thus, the support-only comparison changes the target-support term, while the curvature-only comparison changes the step-size rule.

Per-dataset results Table 5 presents the ablation results per dataset. All four variants use the same examples, targets, initializations, displacement penalty, locality radius, and iteration budget within each model.

B.6 Compute Resources

All experiments were conducted on a high-performance computing cluster. The main compute cost was feature extraction from pretrained encoders. After embeddings were cached, linear probe training, SVM baselines, and downstream analyses were relatively lightweight, typically requiring minutes per run.

Total compute for the reported experiments was on the order of tens of GPU-hours, together with additional CPU-hours for probe training, search sweeps, and analysis, excluding preliminary and discarded pilot runs.

C Additional Results and Case Studies

C.1 Loss-Matched Model Comparisons

The main analysis matches encoder pairs by predictive accuracy. To test whether the conclusion depends on this matching criterion, we repeat the analysis using validation cross-entropy (CE). For each dataset, we select the encoder pair with the smallest validation CE gap and compare its counterfactual behavior. In these experiments, loss-based matching selects the same encoder pairs as accuracy-based matching while additionally accounting for differences in predictive confidence.

Loss-based matching selects the same encoder pairs as accuracy-based matching across all evaluated datasets. Nevertheless, small validation CE gaps coincide with substantial differences in CF-Suc, CF-Dist, and OptEff. The conclusion that predictive performance does not fully characterize counterfactual behavior therefore holds under both accuracy-based and loss-based matching.

Table 6: **Closest-validation-loss encoder pairs within each dataset.** Models with similar validation cross-entropy can nevertheless exhibit distinct counterfactual profiles. Values are mean pairwise differences with standard deviations across five seeds.

Dataset	Model A	Model B	CE Gap	CF-Suc Gap	CF-Dist Gap	OptEff Gap
Shapes	ResNet50	ViT	0.08 ± 0.009	0.53 ± 0.009	0.38 ± 0.007	56.55 ± 1.720
IMDb	BERT	RoBERTa	0.01 ± 0.004	0.19 ± 0.009	0.49 ± 0.007	43.75 ± 1.409
MNIST	ResNet50	ViT	0.06 ± 0.008	0.65 ± 0.002	0.23 ± 0.001	46.72 ± 0.054
Chest X-ray	ResNet50	ViT	0.02 ± 0.018	0.27 ± 0.044	0.16 ± 0.035	29.39 ± 3.096
MM-IMDb	CLIP	SigLIP2	0.02 ± 0.002	0.10 ± 0.004	0.11 ± 0.003	6.34 ± 1.047
Papalexi	scGPT	Geneformer	0.04 ± 0.007	0.56 ± 0.004	0.24 ± 0.003	13.07 ± 2.497

C.2 Classifier-Head Variation under Fixed Representations

Table 7 shows that retraining classifier heads on the same frozen embeddings can substantially change counterfactual behavior while often leaving accuracy nearly unchanged. The effect varies across datasets and encoders: some settings exhibit little head-induced variation, whereas others show large CF-Suc shifts under minor accuracy changes. This suggests that representation geometry constrains the range of possible outcomes, but does not uniquely determine the behavior realized by the classifier.

Table 7: **Variation in counterfactual behavior under fixed representations**. Reported ranges show the maximum change in predictive accuracy and counterfactual success across independently retrained classifier heads on the same frozen embeddings. Large counterfactual shifts can arise despite minor accuracy changes.

Dataset	Encoder	ΔACC Range	ΔCF-Suc Range
Shapes	DINOv2	0.00	0.12
Shapes	ResNet50	0.00	0.13
Shapes	ViT	0.00	0.11
IMDB	BERT	0.00	0.33
IMDB	DistilBERT	0.01	0.19
IMDB	RoBERTa	0.01	0.10
MNIST	DINOv2	0.03	0.26
MNIST	ResNet50	0.01	0.12
MNIST	ViT	0.01	0.22
ChestXray	DINOv2	0.14	0.36
ChestXray	ResNet50	0.02	0.22
ChestXray	ViT	0.00	0.06
MM-IMDb	DINOv2+BERT	0.03	0.40
MM-IMDb	DINOv2+DistilBERT	0.02	0.42
MM-IMDb	DINOv2+RoBERTa	0.01	0.34
MM-IMDb	ResNet50+BERT	0.00	0.17
MM-IMDb	ResNet50+DistilBERT	0.01	0.12
MM-IMDb	ResNet50+RoBERTa	0.02	0.30
MM-IMDb	ViT+BERT	0.02	0.48
MM-IMDb	ViT+DistilBERT	0.02	0.28
MM-IMDb	ViT+RoBERTa	0.02	0.59
MM-IMDb	CLIP	0.01	0.19
MM-IMDb	SigLIP2	0.01	0.13
Papalexi	scGPT	0.02	0.10
Papalexi	Geneformer	0.03	0.19

C.3 Comparison to Maximum-margin Linear Probes

To assess whether probe-to-probe variation is arbitrary or reflects systematic boundary placement, each learned probe is compared to a linear SVM trained on the same frozen embeddings. The SVM

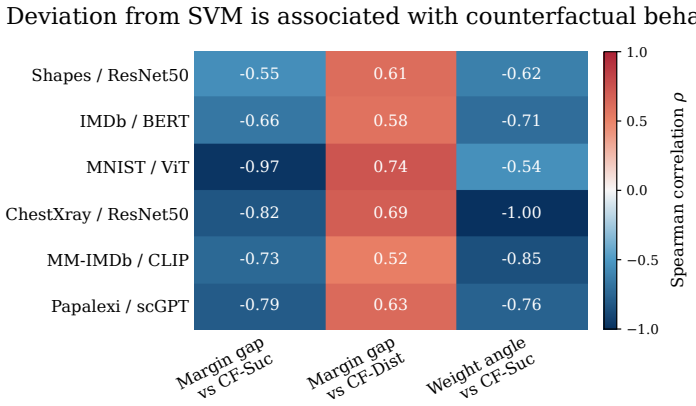

Figure 8: **Deviation from maximum-margin probes is associated with counterfactual behavior.** Spearman correlations between selected SVM-deviation metrics and counterfactual outcomes across datasets. Larger margin-gap or weight-angle deviation from the corresponding linear SVM is generally associated with lower CF-Suc and larger CF-Dist, supporting the role of classifier boundary placement in local counterfactual behavior.

provides a canonical maximum-margin separator in the shared representation space. Figure 8 reports correlations between selected deviation metrics and counterfactual outcomes across datasets. Margin gap measures the difference in average signed margin relative to the SVM, while weight angle measures the difference in orientation between the probe and SVM weight vectors. Across datasets, probes closer to the SVM solution tend to exhibit better counterfactual behavior: higher CF-Suc and lower CF-Dist. On MNIST, for example, margin gap is strongly associated with both lower CF-Suc ($\rho = -0.97$) and higher CF-Dist ($\rho = 0.74$).

These results suggest that probe variability is structured: even within fixed embeddings, boundary placement relative to a maximum-margin reference substantially influences local counterfactual behavior.

C.4 Robustness Beyond Frozen Linear Heads

The main experiments use frozen encoders with linear classifier heads to separate representation geometry from classifier boundary placement. To test whether the observed counterfactual variation is specific to this setting, we additionally evaluate a nonlinear MLP head on frozen MNIST representations and partial fine-tuning obtained by unfreezing the last one to four encoder layers.

Table 8: **Counterfactual behavior beyond frozen linear heads.** A nonlinear MLP head produces counterfactual outcomes similar to the frozen linear-head setting, while progressively unfreezing encoder layers substantially changes counterfactual behavior despite improving predictive accuracy.

MNIST variant	ACC	CF-Suc	CF-Dist
Frozen + linear head	0.95	0.93	0.92
Frozen + MLP head	0.98	0.95	0.89
Partial FT last 1 + linear head	0.98	0.62	0.99
Partial FT last 2 + linear head	0.99	0.59	0.99
Partial FT last 3 + linear head	0.99	0.38	1.21
Partial FT last 4 + linear head	1.00	0.24	1.24

The nonlinear MLP head produces outcomes close to those of the frozen linear head: accuracy increases from 0.95 to 0.98, while CF-Suc changes from 0.93 to 0.95 and CF-Dist from 0.92 to 0.89. This provides initial evidence that the diagnostic is not specific to linear decision boundaries.

Partial fine-tuning produces a substantially different pattern. As progressively more encoder layers are unfrozen, predictive accuracy increases from 0.95 to 1.00, while CF-Suc decreases from 0.93

to 0.24 and CF-Dist increases from 0.92 to 1.24. Thus, improving predictive performance does not necessarily preserve counterfactual behavior. Instead, modifying the representation itself can substantially alter the accessibility of nearby prediction changes. These results reinforce the main finding that predictive performance alone does not characterize counterfactual behavior, while also showing that the diagnostic remains informative beyond the frozen linear-head setting.

C.5 Retrieval-based Counterfactual Search

A retrieval-based baseline complements the projected-gradient probe by searching directly among observed reference embeddings. Results are available across datasets.

Retrieval procedure For each test embedding z with predicted class y, the target y' is selected using the same highest-scoring alternative-class rule as in the gradient-based probe. The procedure retrieves the nearest target-class reference embedding,

$$z' \in \arg\min_{u \in \mathcal{R}_{y'}} \|u - z\|_2. \tag{26}$$

Retrieval counts as successful only when the returned embedding lies within the locality radius and the classifier predicts the target class,

$$\|z' - z\|_2 \leq \tau \quad \text{and} \quad f(z') = y'. \tag{27}$$

The prediction requirement distinguishes an observed target-class label from the classifier's assignment. Retrieval and gradient-based search use the same classifier, target assignments, test examples, and locality radius within each comparison.

Table 9: **Counterfactual success under gradient-based and retrieval-based search.** Both procedures use the same classifier, target-selection rule, and locality radius within each comparison.

Dataset	Gradient CF-Suc	Retrieval CF-Suc
Shapes	0.81	0.73
IMDb	0.93	0.81
MNIST	0.79	0.64
Chest X-ray	0.66	0.59
MM-IMDb	1.00	0.92
Papalexi	0.69	0.60

Retrieval produces prediction-changing alternatives across all six datasets, although its success is generally lower than that of the gradient-based probe. Retrieval restricts candidate endpoints to observed reference embeddings, whereas gradient-based search may explore any point within the locality radius. These results show that the framework can be applied to a non-gradient procedure, but do not establish that model rankings or geometric associations remain unchanged across search methods.

C.6 Qualitative Counterfactual Examples

We show qualitative examples across vision, medical, text, and multimodal settings. Each compares baseline search with geometry-scaled search under identical budgets. Baseline search often fails, while geometry-scaled search guides trajectories to supported target regions and achieves successful prediction changes. Text is trimmed for display, and retrieved examples illustrate nearby target regions rather than the counterfactuals themselves.

D Limitations & Future Work

This study has several limitations. First, the main experiments use frozen pretrained encoders with linear classifier heads to distinguish representation geometry from classifier-side effects. Additional experiments with nonlinear heads and partial fine-tuning broaden this setting, but do not establish that all findings extend to jointly learned representations and classifiers. Second, counterfactual outcomes

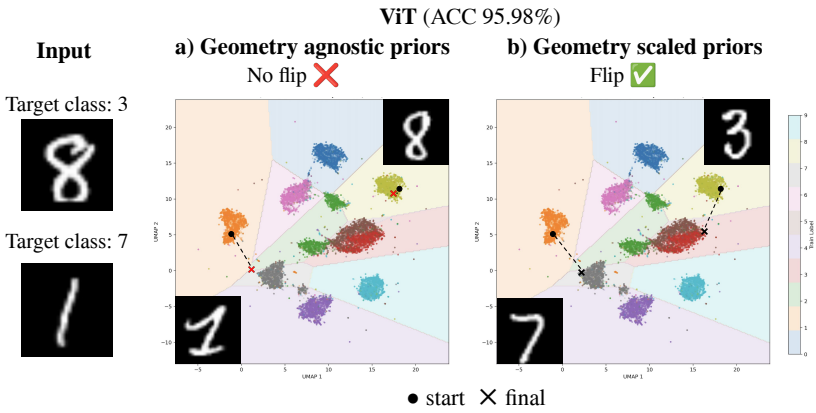

Figure 9: **Illustrative case study: geometry priors can improve local counterfactual search within the same model.** Two MNIST inputs are shown, each evaluated on the same pretrained model under the same optimization budget. For both examples, the baseline search does not reach a prediction change, whereas geometry-scaled search successfully reaches the specified target classes (3 and 7). Retrieved target class neighbors ground the final representation-space endpoints in input space. UMAP panels are shown for visualization only.

depend on the search objective, locality radius, iteration budget, and target-selection rule, with sensitivity analyses examining this dependence. CF-Dist and OptEff are conditioned on successful searches, so comparisons can involve different subsets of examples. Third, representation-space prediction changes do not guarantee semantically meaningful or actionable input-space changes. Nearby reference examples provide evidence of local support rather than proof of input-space feasibility. The single-cell case study shows that decoded counterfactual changes align with observed transcriptional responses and can highlight gene programs for follow-up. Determining whether those changes can be induced in individual cells requires experimental validation. Also, support and curvature estimates depend on the reference data and representation metric. Future work includes extending the framework to jointly trained models, separating geometric feasibility from optimization failure, and developing diagnostics that better account for representation scale and estimation uncertainty. Connecting supported endpoints to realizable input-space changes and evaluating geometry-scaled search under matched computational budgets are further directions.

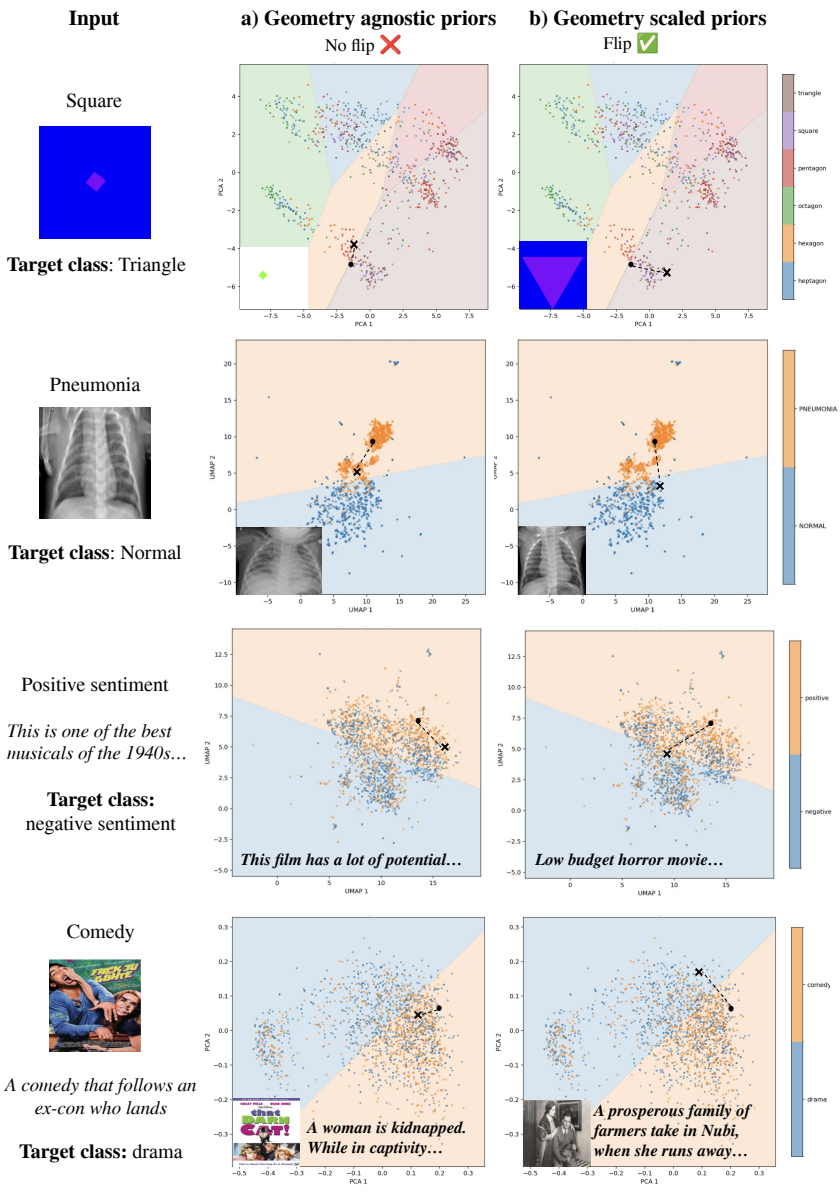

Figure 10: **Geometry-scaled search improves counterfactual search across modalities.** Across vision, medical, text, and multimodal examples, baseline search often fails (no flip), while geometry-scaled search guides trajectories toward target class regions and achieves successful prediction changes. Low-dimensional projections exhibit weak class separation. Retrieved images/text illustrate nearby target regions, not counterfactuals.