MANCE: Manifold Aware Concept Erasure

Matan Avitan
Bar-Ilan University
matan13av@gmail.com

Yoav Goldberg
Bar-Ilan University
Allen Institute for Artificial Intelligence
yoav.goldberg@gmail.com

Yanai Elazar
Bar-Ilan University
yanaiela@gmail.com

Abstract

Concept erasure aims to remove a target concept from a representation while preserving the other information encoded in it. A surgical erasure is challenging as representations encode many concepts that are often correlated with the target concept, so removing the it risks damaging them. We propose the *Manifold Constraint Hypothesis* (MCH): if natural representations concentrate on a structured, lower-dimensional manifold, then interventions should be constrained to that manifold and better preserve other information encoded in the representation during interventions. We instantiate MCH in a new concept erasure method: MANifold aware Concept Erasure (*MANCE*). MANCE performs iterative updates to the representations while staying on the manifold using signals from classifiers that predict the target concept. We estimate the manifold using representations obtained from natural inputs, and then we project the concept removal update to the estimated manifold. We perform extensive evaluation on 119 settings spanning text and vision, including 13 language models, three NLP concepts, and 40 CelebA-CLIP attributes. Employing MANCE on top of previous methods shows consistent improved leakage results. We also introduce MANCE^{+} and MANCE^{++}, which prepend a closed-form erasure algorithm before employing MANCE, achieving better *leakage–surgicality* tradeoffs relative to matched full-space updates. MANCE^{++}, our best method, achieves state-of-the-art results on nonlinear concept erasure. These results support MCH in the erasure setting: interventions should be constrained to the natural representation manifold.[1]

1 Introduction

Neural representations encode many attributes of their inputs: semantic content, style, demographic information, task labels, and other latent factors [Conneau et al., 2018, Tenney et al., 2019, Goldberg, 2019, Hewitt and Manning, 2019, Adi et al., 2016, Rogers et al., 2020, Bolukbasi et al., 2016, Elazar et al., 2021]. In many applications, we want to remove one such attribute (e.g., gender or toxicity), while preserving everything else the representation contains [De-Arteaga et al., 2019, Avitan et al., 2025]. This is the central tension in *concept erasure*: removing information about the target attribute without unnecessarily damaging other concepts, known as *surgicality* [Singh et al., 2024].

Intervening on representations is difficult. Since concepts are often entangled with one another (e.g., profession information and gender), due to data correlation and *superposition* [Elhage et al., 2022], intervening on one concept may negatively affect another. A further challenge is that these other concepts are usually unknown. This makes it hard to formulate erasure as a standard constrained optimization problem to remove a concept while preserving others.

We propose the *Manifold Constraint Hypothesis* (MCH), formalized in Section 3.1 that circumvent these challenges. MCH assumes that representations that arrive from the natural input distribution lie on a manifold. As such, interventions constrained to that manifold will be effective for concept erasure, while also preserving other concepts encoded in the representation.

[1] Code is available at: https://github.com/MatanAvitan/mance

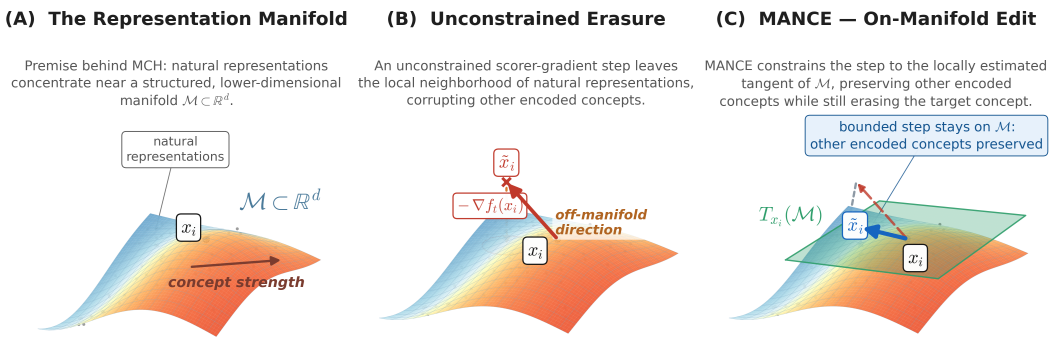


Figure 1: **MCH motivates erasure constrained by the representation manifold.** The *Manifold Constraint Hypothesis* assumes that natural representations concentrate in a structured, lower-dimensional manifold. Under this premise, an unconstrained intervention can change an encoded concept while moving a representation away from the local neighborhood of natural representations, potentially damaging other information encoded along the manifold. MANCE operationalizes MCH for concept erasure by estimating the manifold, and projecting the erasure direction onto the manifold.

We introduce MANCE, MANCE^{+}, and MANCE^{++}, a family of manifold aware methods for nonlinear concept erasure [Edwards and Storkey, 2016, Elazar and Goldberg, 2018, Ravfogel et al., 2020, 2022a,b] that operationalizes MCH. The main idea is to perform a targeted erasure by modifying the representations based on some target concept we would like to erase, and constrain the update to remain on the manifold. We implement the targeted erasure using a standard gradient-based approach (e.g., Goodfellow et al. [2015]). Then, to perform the constrained update we use local manifold estimation by its neighboring representations obtained from natural inputs [Zhang and Zha, 2004]. As part of the constrained optimization, we develop a closed-form solution that allows a flexible update based on the manifold's local geometry. We show that MANCE can be effectively added on top of any previous erasure method, by applying it on the intervened representations. We also propose two variants of MANCE by prepending closed-form linear stages: MANCE^{+} prepends LEACE [Belrose et al., 2023]; and MANCE^{++}, that further prepends a rank-2 projection that removes the leading second-moment covariance asymmetry.

We perform a comprehensive evaluation of concept erasure: 119 settings across text and vision, including 39 NLP settings spanning 13 language models and three concepts, plus 80 CelebA-CLIP settings from 40 image attributes under two surgicality regimes. By applying MANCE on previous erasure methods (INLP [Ravfogel et al., 2020], LEACE [Belrose et al., 2023], and IGBP [Iskander et al., 2023]), we show consistent erasure improvements while minimally harming other concepts. Finally, we show that MANCE^{++} achieves state-of-the-art results on concept erasure, outperforming the strongest baseline [Akbari et al., 2025] under the same limits on control accuracy degradation.

2 Previous Work

Concept erasure methods aim to remove information about a target concept from neural representations while minimizing damage to other information those representations encode. Linear methods such as INLP [Ravfogel et al., 2020] and LEACE [Belrose et al., 2023] target linearly decodable concept information. IGBP [Iskander et al., 2023] and Obliviator [Akbari et al., 2025] remove nonlinear information. IGBP iteratively trains a nonlinear probe and retracts representations toward the current decision boundary; Obliviator approximates a kernel feature map with random Fourier features and iteratively projects out the directions in that feature space that are most predictive of the target concept. A practical limitation of Obliviator is that its erased output is not guaranteed to preserve the original representation dimensionality: the method selects retained feature-space modes through an eigenvalue problem, and the resulting dimension depends on the retained modes. This does not prevent probe-based evaluation, but it limits as-is use in downstream settings where the edited vector must have the original activation dimension before it can be inserted back into the model computation, such as activation patching.

MANCE is orthogonal to the advances in concept erasure as it can be applied to any previous erasure method. Rather than move in arbitrary directions, MANCE projects the erasure update onto the natural-representation manifold. As such, we employ MANCE in addition to previous concept erasure methods, and demonstrate over a wide range of settings how the manifold constraint improves the surgicality of the edits.

3 The MANCE Algorithms Family

3.1 The Manifold Constraint Hypothesis

Hidden states produced by natural inputs do not fill the full representation space uniformly. We assume that they instead occupy a structured, lower-dimensional region shaped by the encoder and the data distribution, which we model as a manifold $\mathcal{M} \subset \mathbb{R}^d$, where d is the representation dimension.

We call an intervention $\mathbf{x} \mapsto \mathbf{x} + \boldsymbol{\delta}$ *constrained to the manifold* when it moves the representation along $\mathcal{M}$, so that the edited point $\mathbf{x} + \boldsymbol{\delta}$ remains on $\mathcal{M}$. This condition is stated with respect to $\mathcal{M}$ itself, independent of how the manifold is modeled or estimated. By contrast, an intervention is *unconstrained* when $\boldsymbol{\delta}$ may point anywhere in $\mathbb{R}^d$, at any magnitude.

The *Manifold Constraint Hypothesis* (MCH) posits that, among interventions with a matched effect on the target concept, manifold-constrained interventions preserve other concepts encoded in the representation better than unconstrained interventions. MANCE operationalizes MCH by estimating a first-order approximation of the manifold using tangent directions obtained from natural inputs.

3.2 The MANCE Method

MANifold aware Concept Erasure (MANCE) operationalizes MCH (§3.1) as a manifold-constrained erasure operation. Its inputs are the initial representations $\mathbf{X}^{(0)} \in \mathbb{R}^{N \times d}$ and target-concept labels $\mathbf{y} = (y_1, \ldots, y_N)$. Here, N is the number of samples and d is the representation dimension. Each row $\mathbf{x}_i^{(0)}$ contains the initial representation of sample i. We assume binary target labels.[2]

MANCE (Algorithm 1) is an iterative algorithm for nonlinear concept erasure. At each round t, it uses the current nonlinear concept probe $f_t : \mathbb{R}^d \to \mathbb{R}$ to compute gradients with respect to the current representations. Crucially, it restricts the erasure direction to an estimated local tangent space of the natural representation manifold. We estimate this space around each current representation, using nearby natural representations from the fixed $\mathbf{X}^{(0)}$.

Concretely, each round projects the probe's gradient into the local tangent basis and weights each component by its corresponding singular value. We use this direction to remove target-concept information through small edits. Each edit is limited to a fixed fraction of the sample's local neighborhood radius. We recompute the tangent bases at every round and retrain the probe every τ rounds on the edited representations. After H rounds, we return the edited representations $\mathbf{X}^{(H)}$.

Notation. Representations $\mathbf{x}_i$ lie in $\mathbb{R}^d$. We write $\mathcal{M}$ for the representation manifold and $T_{\mathbf{x}_i}(\mathcal{M})$ for its *tangent space* at $\mathbf{x}_i$: the linear subspace that approximates $\mathcal{M}$ to first order near $\mathbf{x}_i$, i.e., the directions one can move along while remaining on $\mathcal{M}$ [Tu, 2011]. The method includes a few hyperparameters: H is the number of erasure rounds; τ is the probe-refit period; k is the neighborhood size; r is the number of local directions retained (the tangent rank); λ_i is the per-sample step size; ε is the update size as a fraction of the local neighborhood radius; $\lambda_{\max}$ caps each step.

Step 1: Estimate the manifold locally. The first step is to estimate the representation manifold that will constrain the edits. While measuring the global manifold is challenging, we can estimate it for a local point. For each representation $\mathbf{x}_i$, we instead estimate the local tangent space $T_{\mathbf{x}_i}(\mathcal{M}) \subset \mathbb{R}^d$, the linear subspace that first-order approximates $\mathcal{M}$ near $\mathbf{x}_i$ from nearby natural representations.

Let $\mathcal{N}_k(\mathbf{x}_i)$ be the k nearest neighbors of $\mathbf{x}_i$ among the natural (unedited) representations $\mathbf{X}^{(0)}$, and let $\bar{\mathbf{x}}_{\mathcal{N}_i}$ be their mean. These neighbors provide local samples of how natural representations vary around $\mathbf{x}_i$. We form a local PCA matrix by mean-centering the neighbor positions, with one row $\mathbf{x}_j - \bar{\mathbf{x}}_{\mathcal{N}_i}$ for each neighbor $\mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i)$. Arranging these k centered local variation samples as

[2] Multiclass labels can be handled by fitting one-vs-rest probes.

rows gives $\mathbf{S}_i \in \mathbb{R}^{k \times d}$. We then compute SVD on $\mathbf{S}_i$ to summarize them into the dominant directions of nearby variation:

$$\mathrm{SVD}(\mathbf{S}_i) = \mathbf{L}_i \,\mathrm{diag}(\sigma_{i,1}, \ldots, \sigma_{i,k})\, \mathbf{V}_i^\top. \tag{1}$$

The right singular vectors are the local directions in which neighbors extend, and the singular values measure how far the centered neighborhood extends along each direction. Following local tangent-space approximation methods such as LTSA [Zhang and Zha, 2004], we keep the top r right singular vectors as a tangent-basis matrix whose column span is our estimate of $T_{\mathbf{x}_i}(\mathcal{M})$:

$$\mathbf{B}_i \stackrel{\text{def}}{=} [\mathbf{v}_i^{(1)}, \ldots, \mathbf{v}_i^{(r)}] \in \mathbb{R}^{d \times r}, \qquad \boldsymbol{\sigma}_i \stackrel{\text{def}}{=} (\sigma_{i,1}, \ldots, \sigma_{i,r}).$$

$\mathrm{span}(\mathbf{B}_i)$ is the estimated tangent space $T_{\mathbf{x}_i}(\mathcal{M})$. The columns of $\mathbf{B}_i$ form the basis for that subspace, so the edit is constructed by first expressing the probe's gradient in this local tangent basis and then mapping the resulting tangent-basis coordinates back to representation space. The rank r is the number of tangent-basis directions retained. It should be small enough to avoid unsupported directions while large enough to capture the local geometry; when r is not set directly, we estimate the intrinsic dimension of the current representations once with the TwoNN method of Facco et al. [2017] and keep that rank fixed throughout the editing loop. Although r is fixed, we recompute the local tangent basis $\mathbf{B}_i$ for each sample at every erasure round. The basis is re-estimated at the sample's *current* (edited) position $\mathbf{x}_i^{(t-1)}$, but the neighborhood it is estimated from is always drawn from the fixed natural representations $\mathbf{X}^{(0)}$. Thus, as a point is edited it queries a new neighborhood of the *natural* representations: the tangent reflects how natural representations vary near the edited location, rather than how the already-edited points vary among themselves. This keeps the manifold estimate anchored to the manifold (the region MCH hypothesize we should stay within) while still tracking the moving edit location.

Step 2: Build the tangent erasure direction. After estimating the local tangent basis $\mathbf{B}_i$, we construct an erasure direction from the concept probe's gradient.

We retrain the concept probe every τ rounds, using the edited representations and the same labels. A new probe can find target concept information that remains after the previous edit. Between refits, we use the same probe to compute the erasure gradients. We fix $\tau = 8$ across all settings (Appendix F).

For sample i, let $\mathbf{u}_i$ be this gradient normalized to unit length:

$$\mathbf{u}_i \stackrel{\text{def}}{=} \frac{\nabla f_t(\mathbf{x}_i)}{\|\nabla f_t(\mathbf{x}_i)\|_2}.$$

Projecting onto the local tangent basis gives the vector of tangent-basis coordinates, i.e., the weights of the normalized gradient along the columns of $\mathbf{B}_i$:

$$\mathbf{c}_i \stackrel{\text{def}}{=} \mathbf{B}_i^\top \mathbf{u}_i, \qquad \mathbf{B}_i \mathbf{c}_i = \mathbf{B}_i \mathbf{B}_i^\top \mathbf{u}_i.$$

$\mathbf{B}_i \mathbf{c}_i$ is the projection of the normalized probe gradient onto the estimated tangent space. This removes the gradient component perpendicular to that space.

We weight each gradient component in the tangent basis by the corresponding singular value. This favors directions along which nearby natural representations vary more. Writing $c_{i,\ell}$ for the ℓth entry of $\mathbf{c}_i$, we obtain

$$\mathbf{d}_i \;\stackrel{\text{def}}{=}\; \mathbf{B}_i \,\mathrm{diag}(\boldsymbol{\sigma}_i)\mathbf{c}_i \;=\; \sum_{\ell=1}^{r} \sigma_{i,\ell}\, c_{i,\ell}\, \mathbf{v}_i^{(\ell)}. \tag{2}$$

Appendix L discusses alternative weightings.

The goal of the update is to reduce the part of the current representation aligned with its local erasure direction. We therefore normalize $\mathbf{d}_i$ to $\hat{\mathbf{u}}_i \stackrel{\text{def}}{=} \mathbf{d}_i / \|\mathbf{d}_i\|_2$ and subtract the $\hat{\mathbf{u}}_i$-component of $\mathbf{x}_i$, scaled by the per-sample step size λ_i chosen below:

$$\tilde{\mathbf{x}}_i \;\stackrel{\text{def}}{=}\; \mathbf{x}_i \;-\; \lambda_i \,\langle \mathbf{x}_i,\, \hat{\mathbf{u}}_i \rangle\, \hat{\mathbf{u}}_i. \tag{3}$$

Algorithm 1 MANCE: Manifold aware concept erasure.

Require: Representations $\mathbf{X}^{(0)} \in \mathbb{R}^{N\times d}$, concept labels $\mathbf{y}$, rounds H, neighborhood size k, local-neighborhood scale ε, probe-refit period τ, cap $\lambda_{\max}$, local-basis rank r.

for $t = 1, \dots, H$ **do**

$f_t \leftarrow \begin{cases} \text{fit MLP probe on } (\mathbf{X}^{(t-1)}, \mathbf{y}), & t \equiv 1 \pmod{\tau} \\ f_{t-1}, & \text{otherwise.} \end{cases}$

for $i = 1, \dots, N$ **do**

Step 1: Estimate the manifold locally.

$\mathcal{N}_i^{(t)} \leftarrow k\mathrm{NN}(\mathbf{x}_i^{(t-1)}; \mathbf{X}^{(0)})$, $\quad \bar{\mathbf{x}}_{\mathcal{N}_i}^{(t)} \leftarrow \frac{1}{k}\sum_{j\in\mathcal{N}_i^{(t)}} \mathbf{x}_j^{(0)}$.

$\mathbf{S}_i^{(t)} \stackrel{\text{def}}{=} [\,\mathbf{x}_j^{(0)} - \bar{\mathbf{x}}_{\mathcal{N}_i}^{(t)}\,]_{j\in\mathcal{N}_i^{(t)}}$.

$\mathrm{SVD}(\mathbf{S}_i^{(t)}) = \mathbf{L}_i \,\mathrm{diag}\big(\underbrace{\sigma_{i,1}, \dots, \sigma_{i,r}}_{\stackrel{\text{def}}{=}\boldsymbol{\sigma}_i}, \dots, \sigma_{i,k}\big)\mathbf{V}_i^\top$.

$\mathbf{B}_i \stackrel{\text{def}}{=} [\mathbf{v}_i^{(1)}, \dots, \mathbf{v}_i^{(r)}]$.

Step 2: Build the tangent erasure direction.

$\mathbf{u}_i \stackrel{\text{def}}{=} \dfrac{\nabla f_t(\mathbf{x}_i^{(t-1)})}{\|\nabla f_t(\mathbf{x}_i^{(t-1)})\|_2}$, $\quad \mathbf{c}_i \stackrel{\text{def}}{=} \mathbf{B}_i^\top \mathbf{u}_i$.

$\hat{\mathbf{u}}_i \stackrel{\text{def}}{=} \mathbf{d}_i / \|\mathbf{d}_i\|_2$, $\quad \mathbf{d}_i \stackrel{\text{def}}{=} \mathbf{B}_i \,\mathrm{diag}(\boldsymbol{\sigma}_i)\mathbf{c}_i$.

Step 3: Choose a per-sample local-neighborhood cap.

$r_i \stackrel{\text{def}}{=} \frac{1}{k}\sum_{j\in\mathcal{N}_i^{(t)}} \|\mathbf{x}_j^{(0)} - \mathbf{x}_i^{(t-1)}\|_2$.

$\mathbf{x}_i^{(t)} \leftarrow \mathbf{x}_i^{(t-1)} - \lambda_i \,\langle \mathbf{x}_i^{(t-1)}, \hat{\mathbf{u}}_i\rangle\, \hat{\mathbf{u}}_i$, $\quad \lambda_i \stackrel{\text{def}}{=} \min\left(\lambda_{\max}, \dfrac{\varepsilon \cdot r_i}{\big|\langle \mathbf{x}_i^{(t-1)}, \hat{\mathbf{u}}_i\rangle\big|}\right)$.

end for

end for

return $\mathbf{X}^{(H)}$.

Step 3: Choose a per-sample local-neighborhood cap. A single global step size can over-edit some samples and under-edit others, so MANCE determines λ_i separately for each sample i. Given the tangent direction $\mathbf{d}_i$, λ_i is chosen to make the largest bounded erasure update that stays within a local-neighborhood cap. The cap limits the update to an ε fraction of the distance from $\mathbf{x}_i$ to its nearby representations:

$$\|\tilde{\mathbf{x}}_i - \mathbf{x}_i\|_2 \;\le\; \varepsilon \cdot r_i. \tag{4}$$

This keeps the first-order tangent approximation local: the farther the update moves from the neighborhood used to estimate $\mathbf{B}_i$, the less the local tangent basis describes the geometry around the edited point. The per-sample neighborhood radius is

$$r_i \;\stackrel{\text{def}}{=}\; \frac{1}{k}\sum_{j\in\mathcal{N}_k(\mathbf{x}_i)} \|\mathbf{x}_j - \mathbf{x}_i\|_2, \tag{5}$$

i.e., the mean distance from $\mathbf{x}_i$ to its k nearest neighbors among the natural representations $\mathbf{X}^{(0)}$. Substituting Eq. 3 into Eq. 4 and solving for the largest feasible λ_i gives the closed-form step size

$$\lambda_i \;\stackrel{\text{def}}{=}\; \min\left(\lambda_{\max}\,,\; \frac{\varepsilon \cdot r_i}{\big|\langle \mathbf{x}_i,\, \hat{\mathbf{u}}_i\rangle\big|}\right). \tag{6}$$

Thus λ_i is the largest step along $\hat{\mathbf{u}}_i$ that satisfies the local-neighborhood cap, subject to the upper bound $\lambda_{\max}$. The parameter ε sets the intended displacement as a fraction of the local neighborhood radius r_i; if $\lambda_{\max}$ is smaller than the cap-implied step, the actual displacement is smaller. Because r_i is itself a per-sample, per-panel quantity computed from the local representations, ε is dimensionless in the panel's representation scale and transfers across settings without per-setting tuning (Appendix H).

MANCE$^+$ & MANCE^{++} We propose two additional variants of MANCE that prepend a step that complements the concept removal. Before running the MANCE loop, we match the class-conditional

first and second moments of the representations using closed-form erasers. MANCE^{+} prepends *LEACE* [Belrose et al., 2023], which removes first-moment linear signal associated with classes means. MANCE^{++} additionally prepends *CovMatch*, a rank-2 specialisation of k-LEACE [Singh et al., 2024] that removes second-moment class-conditional covariance asymmetry, represented by the leading eigenvectors of $\Delta\mathbf{\Sigma} \stackrel{\text{def}}{=} \mathbf{\Sigma}_{+} - \mathbf{\Sigma}_{-}$. Both are one-shot affine projections of effective rank ≤ 3 (negligible relative to representation dimension $d \in [768, 5376]$ across our settings; implementation details for CovMatch and its preprocessing ablation are in Appendix A.

4 Experimental setup

4.1 Metrics

We use two standard evaluation metrics from the concept erasure literature: *leakage* and *surgicality*.

Target leakage S. To quantify the remaining information about an erased concept, following Elazar and Goldberg [2018] we train a new nonlinear probe on the representations and denote its accuracy on the test set by S.[3] We compare S with the majority-vote baseline on the test set, denoted S_{floor}; values near S_{floor} indicate that the erased information is no longer recoverable by this probe. We compute $D_S \stackrel{\text{def}}{=} S - S_{\text{floor}}$ as the difference between the probe accuracy and the majority-vote baseline.

For model selection we use $|D_S|$, since erasure quality depends on distance from chance. We report signed D_S to distinguish probe accuracy above and below the majority-vote baseline.

Surgicality ΔY. We measure surgicality as the damage to *control concepts*. Each setting defines a control concept, and reports the test set accuracy of a nonlinear MLP probe, trained separately on the control labels, on the original representations Y_{clean} and on the representations Y_{edit}. We measure the difference between the two: $\Delta Y \stackrel{\text{def}}{=} Y_{\text{edit}} - Y_{\text{clean}}$, where negative values indicate control degradation; an ideal surgical edit has $\Delta Y \approx 0$.

Surgicality budget D_Y. To compare erasure methods fairly, we evaluate them under fixed limits on how much they may degrade control-concept accuracy. For each budget, and for each method, we choose the trajectory step that gives the strongest target erasure while still satisfying that limit. We define the non-negative control-degradation measure $D_Y \stackrel{\text{def}}{=} \max(0, -\Delta Y)$, so only decreases in control accuracy count against the budget; improvements in control accuracy count as zero degradation. An edit is *within budget* b when $D_Y \leq b$. Iterative erasers produce a sequence of edited representations, with different points along the sequence trading off target leakage against control degradation. Among the steps that are within budget, we report the step with the smallest target leakage, measured by $|D_S|$. We report budgets $D_Y \leq b$ for $b \in \{1, 3, 5, 10\}$. We report all leakage (D_S) and surgicality (ΔY) values as differences between accuracies, in points on a 0–100 scale.

Coverage. A method may not satisfy a given budget b on every setting: an eraser may exceed b at all of its trajectory steps. We define *coverage* as the number of settings, out of the total N, on which a method has at least one within-budget edit, reported as n/N. Because $|D_S|$ is averaged only over covered settings, coverage and leakage must be read together: a small mean $|D_S|$ over few covered settings is weaker evidence of erasure than the same value over all of them.

4.2 Settings, methods, and hyperparameters

Settings and probe. We perform extensive evaluations of our methods and previous baselines on 119 settings: 39 NLP settings (13 LLMs $\times$ 3 concepts at the 50%-depth layer of each model) and 80 CelebA-CLIP settings (CLIP ViT-B/32 [Radford et al., 2021] final-layer pooled image embeddings of CelebA [Liu et al., 2015]; 40 binary facial attributes $\times$ 2 surgicality regimes). The 13 LLMs come from five model families and span 0.5B–27B parameters: Qwen2.5 (0.5B, 1.5B, 3B), Gemma-2 (2B, 9B, 27B), Gemma-3 (1B, 4B, 12B, 27B), Llama-3.2 (1B, 3B), Mistral-7B-v0.1. The three NLP concepts are *sycophancy* [Perez et al., 2023], *gender* (Bias in Bios [De-Arteaga et al., 2019]), and *safety* (PKU-SafeRLHF [Ji et al., 2025]). The control concepts for these three datasets are

[3] Unless otherwise specified, the new probe is a 2-layer MLP ($h = 128$, 200 SGD steps, patience 3).

answer preference for sycophancy, profession for gender, and helpfulness for safety, respectively. In PKU-SafeRLHF, examples are response pairs; we use the dataset's helpfulness preference within each pair to label the preserved control concept, while the safety annotation defines the erasure target.

For CelebA, we use each of the 40 attributes as the erasure target under two surgicality regimes: *highly entangled* and *highly disentangled* control concepts. These two regimes are formed from the five attributes most and least correlated with the target concept on the training split, respectively. For example, when erasing *male*, the most-correlated controls are *wearing lipstick* ($|r| \approx 0.80$), *heavy makeup*, *no beard*, *attractive*, and *5 o'clock shadow* (stubble). The resulting control sets, with their $|r|$ values, are listed in App. Tabs. 11 and 12; additional construction details are in App. I.

Baselines. In addition to the MANCE family, we compare against several baselines. LEACE [Belrose et al., 2023] and LEACE + CovMatch are closed-form one-shot linear and rank-2 covariance erasers. INLP [Ravfogel et al., 2020] is an iterative linear nullspace projection method. IGBP [Iskander et al., 2023] is an iterative 2-class probe with Newton-step boundary projection. Obliviator [Akbari et al., 2025] is the current state-of-the-art nonlinear method, which iteratively trains an encoder to minimize statistical dependence between the representation and the erased attribute while preserving task-relevant information, and then applies a disentanglement step in a reproducing kernel Hilbert space to refine the representation for the next round.

Ablation. To isolate the effect of the tangent constraint, we introduce AmbCE^{++}, a controlled variant of MANCE^{++} that uses the same preprocessing and nonlinear probe loop but removes the tangent projection. Instead of projecting the concept-probe gradient into the local tangent basis and scaling the step by the per-sample local-radius cap, AmbCE^{++} takes a regular, non constrained gradient step with a global step size, $\lambda = 29.31$. Its displacement $\lambda|\langle \mathbf{x}_i, \hat{\mathbf{u}}_i \rangle|$ still varies across samples, but no local radius bounds it. We set λ to the mean per-sample λ_i of MANCE^{++} on the NLP settings (App. Tab. 13), making the ablation comparable in effective step scale, and use the same value on CelebA. Thus, AmbCE^{++} tests whether the gains come from the tangent constraint rather than from the nonlinear probe loop alone, directly testing the MCH prediction (§3.1).

Evaluation protocol. The NLP datasets do not share a standard split, we therefore create our own split: we sample $12{,}000$ examples, and create a randomized $60/20/20$ split, resulting in $7{,}200$ examples for train, $2{,}400$ for val, and $2{,}400$ for test, per setting. For CelebA, we use the official train/val/test partition, subsampling within each to $1{,}200$ / 300 / 300 images per setting. In all cases the eraser is fit on train representations, the val split only early-stops the probes, and both S and Y are reported on the held-out test split. We compute S on each of the 40 attributes, and Y as the mean accuracy on the regime's five control attributes. We run every method with three seeds. Each setting takes the median over its within-budget seeds, and counts toward coverage (§4.1) if it has at least two.

Hyperparameters. All hyperparameters are fixed across the 119 settings with no per-setting tuning; values are reported in App. F.

5 Results

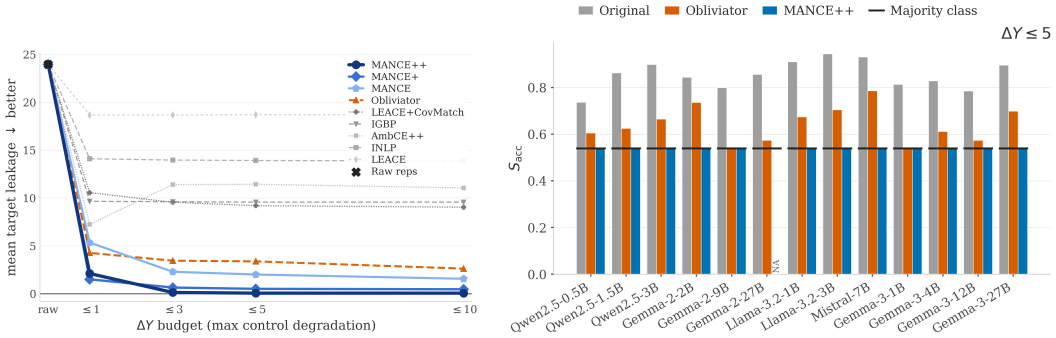
We evaluate the MANCE algorithms family (§3), and previous erasure methods as baselines. We test all methods on the 119 settings detailed earlier.

For iterative methods (e.g., INLP), we run the evaluation metrics for each iteration. Then, per surgicality budget, we select the iteration that erases the most of the target concept. In the non-iterative methods, we use the obtained, single result.

Adding MANCE improves prior erasers. Applying MANCE on intervened representations from other erasers leads to improved results across most settings. Table 1 reports leakage gains obtained by applying MANCE after each erasure method, aggregated over the 39 NLP settings (13 LLMs $\times$ three concepts: sycophancy, gender, and safety; see §4.2). Under the same surgicality budgets, adding MANCE drives each baseline's target leakage close to chance: at $D_Y \leq 1$ it falls from 19.3 to 1.5 for LEACE, 15.4 to 1.7 for INLP, and 11.2 to 1.4 for IGBP, with similar drops at the looser budgets, and it keeps 99% of the covered settings within budget (517 of the 521 in Table 1). The only exception is Obliviator, where we see no improvement in leakage. However, Obliviator's final output

Table 1: **MANCE complements prior erasers.** Mean residual leakage $|D_S|$ (points); *lower is better in both columns*. *Baseline* (↓): the baseline's leakage. *+MANCE* (↓): leakage after applying MANCE to the baseline's output. Parentheses give each method's *coverage* in that column (§4.1): the number of the 39 settings it keeps within budget. Means use only the covered settings. Values summarize 3 seeds (§4.2). App. Table 4 reports the standard deviations. Adding MANCE further lowers leakage and keeps 99% of the covered settings within budget (517 of 521). The +MANCE cells of the last two rows correspond to MANCE^{+} and MANCE^{++}. †Obliviator's final output stays within budget on only 13–20/39 settings, where it is already at chance, so MANCE has nothing to remove.

Method	$D_Y \leq 1$		$D_Y \leq 3$		$D_Y \leq 5$		$D_Y \leq 10$	
	Baseline↓	+MANCE↓	Baseline↓	+MANCE↓	Baseline↓	+MANCE↓	Baseline↓	+MANCE↓
MANCE (no preprocessing)	–	5.4 (39/39)	–	2.2 (39/39)	–	1.9 (39/39)	–	1.6 (39/39)
INLP	15.4 (36/39)	1.7 (38/39)	16.2 (38/39)	1.3 (39/39)	16.1 (39/39)	1.1 (39/39)	16.1 (39/39)	1.0 (39/39)
IGBP	11.2 (38/39)	1.4 (38/39)	11.2 (39/39)	0.8 (39/39)	11.2 (39/39)	0.6 (39/39)	11.2 (39/39)	0.5 (39/39)
Obliviator	0.0† (13/39)	0.0† (13/39)	0.0† (13/39)	0.0† (13/39)	0.0† (13/39)	0.0† (13/39)	0.0† (20/39)	0.0† (19/39)
LEACE	19.3 (38/39)	1.5 (37/39)	19.2 (39/39)	0.6 (39/39)	19.2 (39/39)	0.5 (39/39)	19.2 (39/39)	0.5 (39/39)
LEACE + CovMatch	–	2.1 (30/39)	–	0.1 (37/39)	–	0.1 (38/39)	–	0.1 (39/39)



(a) Mean NLP leakage envelope across 39 settings; MANCE variants form the lowest curves and MANCE^{++} is below 1 point from $D_Y \leq 3$ onward. Each setting takes the median over its seeds.

(b) Per-model gender nonlinear-probe accuracy at $D_Y \leq 5$; black ticks mark majority class. MANCE^{++} reaches the majority class on almost every model, while Obliviator stays above it on most. NA marks fewer than 2 of 3 seeds within budget.

Figure 2: **MANCE variants improve the leakage-surgicality tradeoff in NLP.**

stays within budget on only 13–20/39 settings. On those settings its leakage is already at chance. This apparent strength is an artifact of *which* settings it reaches. Next, we will show how for harder, highly correlated settings, MANCE^{++} outperforms Obliviator (and other baselines).

MANCE^{++} gives the strongest leakage–surgicality tradeoff. Table 2 and Fig. 2 present the results on the NLP datasets, and Figure 3 on the vision dataset; complete per-setting results appear in the appendix (NLP in Table 5 and Figs. 4 to 6, vision in Table 6 and Fig. 7). Both modalities show the same trend: MANCE^{++} outperforms all baselines, setting a new state of the art in nonlinear concept erasure. On the NLP settings, MANCE^{+} and MANCE^{++} are the only methods that stay near chance at every budget: from $D_Y \leq 1$ to $D_Y \leq 10$, MANCE^{+} goes from 1.5 to 0.4 and MANCE^{++} from 2.0 to 0.1. The strongest baseline, Obliviator, retains $4.3 \rightarrow 2.6$. On the settings both methods keep within budget, MANCE^{++} leaves 2.0 against 3.9 for Obliviator at $D_Y \leq 1$, and 0.1 against 2.6–3.5 at the looser budgets. In the vision setting, under the same configuration and preprocessing as on NLP, MANCE^{++} is the only method that combines high coverage with near-floor leakage (Figure 3).

In the least-correlated regime, MANCE^{++} covers 39/40 and 40/40 attributes at $D_Y \leq 1$ and $D_Y \leq 3$, respectively, compared with 15/40 and 29/40 for Obliviator. In the most-correlated regime, MANCE^{++} covers 17/40 and 35/40 attributes, respectively, compared with 1/40 and 13/40 for Obliviator. Obliviator's low leakage in Figure 3 is biased by coverage: it is averaged only over the easy attributes where Obliviator stays in budget. On the other hand MANCE^{++} remain close to zero leakage, but remain a much higher coverage. Thus, across settings, manifold aware erasure is more effective under the same surgicality budgets, providing empirical support for MCH (Sec. 3.1).

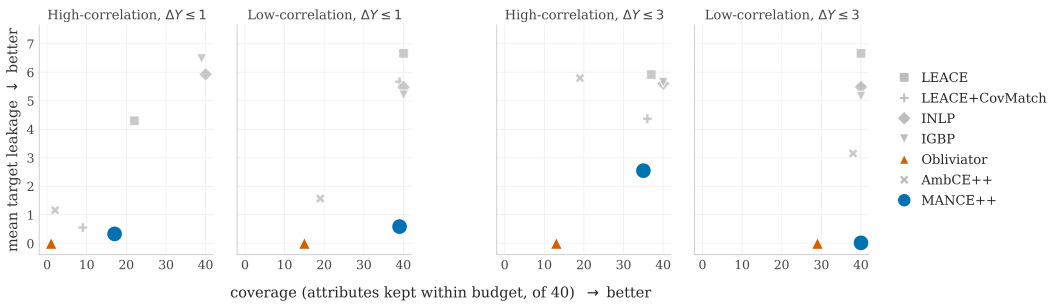


Figure 3: **Vision (CelebA): coverage vs. mean target leakage.** Each point is one erasure method: x-axis is coverage (the number of the 40 attributes kept within the surgicality budget), y-axis is its mean target leakage over those covered attributes; lower-right is better. Panels pair the two budgets ($D_Y \leq 1$, $D_Y \leq 3$) with the high- and low-correlation control regimes. MANCE^{++} shows consistent strong tradeoffs between target leakage and coverage, and marks a new pareto curve. Obliviator (orange) matches its low leakage only at far lower coverage, while LEACE, INLP, and IGBP cover many attributes but leave 4–7 points of leakage.

Table 2: **Main NLP erasure result.** *Leakage* (↓): how much target information survives, measured as D_S in points at the within-budget step closest to chance, so 0 means the probe is at majority-class chance and larger values mean more leakage. *Coverage* (↑): the number of the 39 settings a method keeps within budget (§4.1). Leakage is averaged over the covered settings. Values summarize 3 seeds (§4.2). App. Table 4 reports the standard deviations. Per-panel detail is in App. Tab. 5.

	$D_Y \leq 1$		$D_Y \leq 3$		$D_Y \leq 5$		$D_Y \leq 10$	
Method	Leakage↓	Coverage↑	Leakage↓	Coverage↑	Leakage↓	Coverage↑	Leakage↓	Coverage↑
LEACE	18.7	39/39	18.7	39/39	18.7	39/39	18.7	39/39
LEACE+CovMatch	10.5	29/39	9.5	35/39	9.2	37/39	9.0	39/39
INLP	14.1	39/39	14.0	39/39	13.9	39/39	13.9	39/39
IGBP	9.6	38/39	9.6	39/39	9.5	39/39	9.5	39/39
Obliviator	4.3	39/39	3.4	39/39	3.4	39/39	2.6	39/39
AmbCE^{++}	6.6	22/39	10.9	31/39	11.0	32/39	10.8	33/39
MANCE	5.2	37/39	2.3	39/39	2.0	39/39	1.5	39/39
MANCE^{+}	**1.5**	37/39	0.6	39/39	0.5	39/39	0.4	39/39
MANCE^{++}	2.0	30/39	**0.1**	37/39	**0.1**	38/39	**0.1**	39/39

The gain is largest where the target is hardest to remove surgically. MANCE^{++}'s advantage over the baselines is largest where the target is hardest to remove without damaging the control concepts. In NLP this is clearest on gender, where profession strongly correlates with gender: on the models both methods keep within budget, MANCE^{++} leaves 6.2 against 11.7 for Obliviator at $D_Y \leq 1$, and the gap widens sharply to 0.0 against 10.7 at $D_Y \leq 5$ (Fig. 2b; App. Fig. 4 and Table 5). Safety is easier: both methods reach near chance on all 13 models at $D_Y \leq 1$, with a mean $|D_S|$ of 0.2 for MANCE^{++} and 0.0 for Obliviator (Fig. 6 in the Appendix). On the other hand, Sycophancy is already solved by linear erasure methods, leaving nothing for MANCE to add (App. Fig. 5 and Table 3). Finally, the vision setting shows the same pattern, where the difficulty is measured directly as target–control correlation: under the most-correlated controls MANCE^{++} keeps far more attributes within budget than Obliviator while staying near the target leakage floor (Figure 3).

The manifold constraint makes erasure precise and surgical. To test the effect of the manifold constraint, we compare MANCE^{++} with an unconstrained variant, AmbCE^{++}. Both use the same preprocessing and nonlinear probe-refit loop. However, AmbCE^{++} uses the full-space probe gradient without tangent projection or singular-value weighting. Its global step-size parameter, $\lambda = 29.31$, matches the mean per-sample λ_i of MANCE^{++} on the NLP settings (App. Tab. 13). Without the manifold constraint it leaves 7–11 points of leakage, and on many settings no step of its trajectory stays within the surgicality budget; by the coverage protocol (§4.1) those out-of-budget settings are excluded from its leakage mean, so even that 7–11 is measured only over the easier settings it keeps

in budget. Compared with MANCE^{++}'s $2.0 \rightarrow 0.1$, therefore, the gains come from the manifold constraint, not the closed-form preprocessing or the nonlinear probe loop alone.

The latency tradeoff. We benchmark the latency of the different methods on the Gender concept. MANCE takes about eight minutes (458.8–474.9 s across the three variants on one NVIDIA B200), with the closed-form preprocessing in MANCE^{+}/MANCE^{++} adding only a few seconds on top; the baselines range from a few seconds (LEACE, CovMatch) to a few minutes (INLP) (Table 10, §G). Profiling attributes roughly 50% of the runtime to the per-round local SVDs and 40% to CPU–GPU transfers and CPU-side projection arithmetic. Both are properties of the implementation rather than of the algorithm, and a GPU-resident implementation would reduce the runtime severalfold.

Summary. Across the 119 settings, MANCE improves prior erasers under the same surgicality budget while keeping 99% of their covered settings within budget (517 of 521, Table 1), and MANCE^{++} achieves state-of-the-art concept erasure across models, datasets, and modalities. These results support the Manifold Constraint Hypothesis (§3.1): constraining edits to locally supported directions improves surgical concept erasure.

6 Conclusion

We proposed the *Manifold Constraint Hypothesis*: if natural representations concentrate on a manifold, then interventions constrained to the manifold should preserve other information encoded in the representation. MANCE instantiates this hypothesis for concept erasure: it estimates the manifold from neighboring representations, follows the nonlinear concept probe toward reduced leakage, and confines the update to the manifold.

Across 119 settings in text and vision, MANCE^{++} achieves state-of-the-art concept erasure under the same surgicality budget. Applied as an additional step, MANCE also improves prior erasers, and the tangent constraint improves the tradeoff between leakage and surgicality relative to unconstrained updates. These results support MCH for erasure.

We expect the same argument to hold for other representation edits, such as activation steering [Subramani et al., 2022, Turner et al., 2023]. Such interventions should also respect the geometry that the model and the data induce jointly. MANCE estimates $\mathcal{M}$ with a local, first-order tangent space, so a better estimate of $\mathcal{M}$ should allow more surgical interventions. Future work can develop such estimators and formally link their quality to guarantees on interventions.

Limitations

Measurement scope. We quantify erasure by retraining a nonlinear MLP probe on the edited representations, following the recoverability protocol of Elazar and Goldberg [2018], and we quantify preservation by accuracy on the labeled control concepts. Both target leakage S and surgicality ΔY are thus empirical measurements under a fixed probe protocol rather than guarantees that the target concept is fully unrecoverable. Surgicality in particular speaks only to the control concepts that we can measure (Sec. 4.2), and not to the full information content of the representation.

Local manifold estimation. MANCE constrains each edit to a local tangent space $T_{\mathbf{x}_i}(\mathcal{M})$ estimated by local PCA over the k nearest natural representations $\mathbf{X}^{(0)}$, with the tangent rank r taken from a TwoNN intrinsic-dimension estimate. This estimate degrades where the natural representations $\mathbf{X}^{(0)}$ are sparse or the manifold is strongly curved. The tangent constraint also helps only to the extent that MCH holds, that is, that representations concentrate on a structured, lower-dimensional manifold, so its advantage over full-space editing is expected to shrink as the intrinsic dimension approaches the representation dimension d.

Computational cost. Unlike closed-form linear erasers such as LEACE, which apply a single affine map, MANCE recomputes a k-NN query and a local SVD for every representation at every round and periodically refits the nonlinear probe. Applying the fitted eraser to new inputs likewise requires querying the natural representations $\mathbf{X}^{(0)}$ for their nearest neighbors, so both fitting and deployment are heavier than one-shot projection methods.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *CoRR*, abs/1608.04207, 2016. URL http://arxiv.org/abs/1608.04207.

Ramin Akbari, Milad Afshari, and Vishnu Naresh Boddeti. Obliviator reveals the cost of nonlinear guardedness in concept erasure. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://neurips.cc/virtual/2025/poster/118956.

Matan Avitan, Ryan Cotterell, Yoav Goldberg, and Shauli Ravfogel. A practical method for generating string counterfactuals. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3267–3286. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.180. URL https://aclanthology.org/2025.findings-naacl.180/.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems 36*, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/d066d21c619d0a78c5b557fa3291a8f4-Abstract-Conference.html.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357, 2016.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, 2018.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 120–128. ACM, 2019. doi: 10.1145/3287560.3287572.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *International Conference on Learning Representations*, 2016.

Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1002. URL https://aclanthology.org/D18-1002/.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10/.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7:12140, 2017. doi: 10.1038/s41598-017-11873-y.

Yoav Goldberg. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*, 2019.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 4129–4138, 2019. URL https://aclweb.org/anthology/papers/N/N19/N19-1419/.

Shadi Iskander, Kira Radinsky, and Yonatan Belinkov. Shielded representations: Protecting sensitive attributes through iterative gradient-based projection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5961–5977, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.369. URL https://aclanthology.org/2023.findings-acl.369/.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Juntao Dai, Boren Zheng, Tianyi Qiu, Jiayi Zhou, Kaile Wang, Boxuan Li, Sirui Han, Yike Guo, and Yaodong Yang. PKU-SafeRLHF: Towards multi-level safety alignment for LLMs with human preference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Main Conference)*. Association for Computational Linguistics, 2025. URL https://arxiv.org/abs/2406.15513. arXiv:2406.15513.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. doi: 10.1109/ICCV.2015.425.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.847. URL https://aclanthology.org/2023.findings-acl.847.

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. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647/.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. Linear adversarial concept erasure. In *International Conference on Machine Learning*, pages 18400–18421. PMLR, 2022a.

Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. Adversarial concept erasure in kernel space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6034–6055. Association for Computational Linguistics, 2022b. doi: 10.18653/v1/2022.emnlp-main.405. URL https://aclanthology.org/2022.emnlp-main.405/.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 2020. doi: 10.1162/tacl_a_00349. URL https://aclanthology.org/2020.tacl-1.54/.

Shashwat Singh, Shauli Ravfogel, Jonathan Herzig, Roee Aharoni, Ryan Cotterell, and Ponnurangam Kumaraguru. Representation surgery: Theory and practice of affine steering. In *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2024. URL https://arxiv.org/abs/2402.09631. arXiv:2402.09631.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.48. URL https://aclanthology.org/2022.findings-acl.48/.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the Conference of the Association for Computational Linguistics, ACL*, pages 4593–4601, 2019. URL https://www.aclweb.org/anthology/P19-1452/.

Loring W. Tu. *An Introduction to Manifolds*. Universitext. Springer, 2 edition, 2011. doi: 10.1007/978-1-4419-7400-6.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Zhenyue Zhang and Hongyuan Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM journal on scientific computing*, 26(1):313–338, 2004.

A Closed-form preprocessing details

The two MANCE preprocessing variants (MANCE^{+} and MANCE^{++}) exhaust 1st- and 2nd-moment concept signal in closed form so that the subsequent manifold-constrained loop only has to attack the residual nonlinear pocket.

LEACE (1st moment). Fit a LEACE eraser [Belrose et al., 2023] on $(\mathbf{X}_{\text{train}}, \mathbf{y})$ and apply to all splits. LEACE removes the rank-1 cross-covariance $\mathbf{\Sigma}_{xz}$, leaving no *linear* classifier above majority-class accuracy with provably minimum representation damage among affine projections.

Covariance-asymmetry projection (2nd moment). LEACE matches class means but leaves the within-class asymmetry

$$\Delta\mathbf{\Sigma} \stackrel{\text{def}}{=} \mathbf{\Sigma}_{+} - \mathbf{\Sigma}_{-}. \tag{7}$$

A nonlinear probe still recovers concept information from variance asymmetry along axes where $\mathbf{\Sigma}_{+}$ and $\mathbf{\Sigma}_{-}$ differ. We append a rank-2 projection: the top-2 eigenvectors of $\Delta\mathbf{\Sigma}$ ranked by $|\lambda|$, orthonormalized with a residual mean direction via QR,

$$\mathbf{D} = \mathrm{QR}\Big([\, \hat{\mathbf{d}}_{\text{mean}} \;\big|\; \mathbf{e}_1, \mathbf{e}_2 \,] \Big) \in \mathbb{R}^{d\times 3}, \tag{8}$$

$$\tilde{\mathbf{x}} = \mathbf{x} - \mathbf{D}\mathbf{D}^{\top}\mathbf{x}. \tag{9}$$

Because LEACE has aligned class means, $\hat{\mathbf{d}}_{\text{mean}} \approx \mathbf{0}$ and the effective rank removed is two. This is a special case of the k-LEACE framework of Singh et al. [2024] restricted to second-order eigenvectors, which are stable to estimate from $n \approx 7000$ examples while higher-order tensors are not. We keep the top two eigenvectors across all 119 settings: for a binary concept, these directions capture the two axes of class-conditional variance asymmetry; subsequent eigenvectors describe noise rather than class-discriminative signal. Empirically, keeping three or more eigenvectors adds no leakage reduction and begins to erode control accuracy (see Tab. 5).

The two stages erase structurally distinct signal (mean shift vs. variance asymmetry) and together remove an effective rank of at most 3, negligible relative to the representation dimension d across our settings ($d = 768$ for the CLIP ViT-B/32 image embedding, and $d \in [896, 5376]$ across the 13 LLM hidden sizes), exposing the residual nonlinear structure that the manifold-constrained loop attacks.

B Closed-form preprocessing ablation

Tab. 3 reports a new nonlinear probe (h=128, 200 SGD steps, patience 3) trained on three representation states per setting: clean, post-LEACE, post-(LEACE + CovMatch). Sycophancy is fully exhausted by closed-form preprocessing (mean NL 0.563 against floor 0.560 and MANCE^{++}'s 0.560). Gender is not: LEACE + CovMatch leaves 14.6 points above floor that the manifold-constrained loop closes to 0.0. Safety follows the gender pattern with $12.1 \rightarrow 0.2$.

C Variation across seeds

We run every method with three random seeds (§4.2). Table 4 reports every leakage value of the main tables with its standard deviation over the seeds. On the NLP settings, MANCE^{++} has a lower mean leakage than every baseline on each seed and at each budget.

D NLP full per-setting breakdown

Table 5 reports the per-setting version of the main-text Tab. 2, restricted to the 7 method columns shared with the main-text breakdown (the pure MANCE and MANCE^{+} ablation rows of the main-text table are not repeated per-setting). For each of the 39 NLP settings and each of the four surgicality budgets, the signed deviation D_S in points is measured at the in-budget trajectory step with smallest $|D_S|$. Bold marks the row-best per setting per budget block (every method within numeric precision of the row-best is bolded). The trailing row reports the per-method mean across the 39 settings.

E CelebA full per-attribute breakdown

Tab. 6 is the full per-method × per-budget summary across both surgicality regimes summarized in §5. Cells report signed mean D_S in points at the in-budget trajectory step with smallest $|D_S|$, plus coverage (§4.1).

Fig. 7 reports the per-attribute nonlinear-probe accuracy S for each of the 40 CelebA attributes at all four surgicality budgets, in the same bar-chart style as the NLP per-model figures (App. Figs. 4–6): three bars per attribute (Clean / MANCE^{++} / Obliviator) and a horizontal tick at the per-attribute majority-class floor. Hatched bars annotated "N/A" indicate that the method had no in-budget step at the shown D_Y budget; we substitute its clean (un-erased) S so the visual height communicates "no valid in-budget point" rather than a real measurement. A clear win for MANCE^{++} reads as a low solid blue bar paired with a high (often hatched-N/A) orange Obliviator bar, common at $D_Y \leq 1$ in the most-correlated regime, where Obliviator overshoots the budget on 38 of 40 high-correlation targets while MANCE^{++} keeps 19 in budget and drives 15 of them to floor.

Tabs. 7–9 report the per-attribute numerical version of App. Tab. 6, one table per surgicality budget $D_Y \in \{1, 3, 5\}$: signed D_S in points at the in-budget trajectory step with smallest $|D_S|$. Bold marks the row-best per regime block. The trailing row reports the per-regime per-method mean. The tightest budget (≤ 1) is where the gap between MANCE^{++} and the baselines is largest; it is also where Obliviator overshoots most often and shows "–" on the hardest entangled targets.

F Hyperparameters

The per-sample local-radius-capped λ_i (Eq. 6) eliminates the global λ from the proposed method. We fix the remaining step-controlling hyperparameters across all 119 settings (39 NLP × 80 CelebA) with no per-setting tuning. Values and the appendix sections that justify them:

Symbol	Value	Role / justification
ε	0.1	local-radius fraction (App. H)
h	512	concept-probe hidden dimension
probe training steps	800	per probe refit
probe-refit interval	8	every 8 outer iterations
H	60	outer iterations
$k_{\min}$	8	TwoNN intrinsic-dimension floor
$\lambda_{\max}$	64	safety cap on per-row λ_i (App. J)
CovMatch rank	2	number of covariance-asymmetry eigenvectors (App. A)

Every method in Tab. 2 runs at these fixed values.

Local-basis rank r and the floor $k_{\min} = 8$. The rank r of the local tangent basis is set from the intrinsic dimension of the clean representations $\mathbf{X}^{(0)}$, estimated once with TwoNN [Facco et al., 2017] and kept fixed throughout the iterative loop:

$$r = \max\big(k_{\min},\ \lceil \mathrm{TwoNN}(\mathbf{X}^{(0)}) \rceil\big), \qquad k_{\min} = 8.$$

We use TwoNN as a data-driven default for the rank so that the basis is wide enough to capture the local variance of nearby representations but not so wide that it absorbs noise directions, both of which would degrade the spectrally-weighted direction in Eq. 2. The floor of 8 is a guard against pathological TwoNN outputs: on small or noisy representation pools the estimator can collapse to single-digit values (sometimes 1–3), at which point a rank-r basis is too short for the local singular spectrum to support a meaningful σ-weighting. The floor sets a conservative minimum that prevents this collapse without interfering with reasonable estimates; on our 119 settings TwoNN typically returns values well above 8, so the floor does not bind in practice.

G Computational cost

Table 10 reports the wall-clock cost of fitting and applying each method on a representative NLP panel (Qwen2.5-1.5B, gender, layer 14), using each method's canonical configuration from Table 2.

All nine methods are timed sequentially in a single job on one NVIDIA B200 GPU with 8 CPU cores. We time only the fit-and-apply path: diagnostic probes used solely to log the S/Y trajectories are excluded, while probes internal to a method's own stopping rule (INLP's convergence check, MANCE's scorer margin) are included.

MANCE is the most expensive method, at roughly eight minutes per panel (458.8–474.9 s across the three variants): each round re-estimates a local tangent basis per row (kNN against the fixed natural reference $\mathbf{X}^{(0)}$, then a batched SVD), and periodically refits the nonlinear scorer. The closed-form preprocessing in MANCE^{+} and MANCE^{++} adds only a few seconds on top of the loop ($458.8 \to 470.5 \to 474.9$ s). The baselines are far cheaper: LEACE and CovMatch are one-shot at a few seconds (3.7 and 7.3 s); the boundary and gradient erasers Obliviator, IGBP, and AmbCE^{++} run in 20–24 s; and INLP takes 286.9 s, dominated by its CPU-bound per-round logistic-regression fits. This is a one-time, per-panel offline cost and does not affect inference: every method outputs an edited representation set, or an affine map for the closed-form erasers.

H Why $\varepsilon = 0.1$ transfers across panels

Section 3.2 chooses the per-sample step size λ_i so that the displacement magnitude satisfies $\|\tilde{\mathbf{x}}_i - \mathbf{x}_i\|_2 \leq \varepsilon \cdot r_i$, where r_i is the mean Euclidean distance from $\mathbf{x}_i$ to its k nearest neighbors among the natural representations $\mathbf{X}^{(0)}$ (Eq. 5). This appendix explains why a single value $\varepsilon = 0.1$ can be used across all 39 NLP settings (13 language models $\times$ 3 concepts at 50%-depth) and all 80 CelebA settings (40 attributes $\times$ 2 surgicality regimes) without per-setting tuning of the local-radius cap.

Reading the constraint. With $\|\tilde{\mathbf{x}}_i - \mathbf{x}_i\|_2 \leq \varepsilon \cdot r_i$, the budget reads "the displacement L2 norm is at most ε fraction of the local Euclidean neighborhood radius." The local radius r_i is recomputed every round: the neighbors are drawn from the fixed natural representations $\mathbf{X}^{(0)}$ but measured from the sample's current (edited) position $\mathbf{x}_i^{(t-1)}$, so as a point moves under earlier edits its distance to the surrounding natural neighborhood (and hence the budget) tracks the new location. The same ε corresponds to a small move when the neighborhood is dense (small r_i) and a larger move when the neighborhood is sparse (large r_i), which is the behavior we want from a local first-order step.

Why ε is setting-independent. A single global step size λ has setting-specific units: applied verbatim to a Gemma-3-27B representation with $d = 5376$ it moves the point by a much larger Euclidean distance than the same constant applied to a Qwen2.5-0.5B representation with $d = 896$. The local r_i used here is computed from each panel's own representations, so the displacement magnitude $\varepsilon \cdot r_i$ is automatically expressed in the panel's representation scale. The ratio of the displacement to the local neighborhood radius is the same across panels for a fixed ε, even though the absolute Euclidean distance moved is very different. This is the property that makes $\varepsilon = 0.1$ behave consistently across the 119 settings.

Per-row, not just per-panel. Because r_i is computed at the level of each individual sample, the budget also adapts within a panel. Rows that lie in dense regions (where neighbors are close) get small allowed displacements; rows in sparse regions get larger ones. This is what distinguishes the per-row λ_i in Eq. 6 from a global λ: a global step size has to be small enough not to over-edit the densest rows, which leaves the sparse rows under-edited. The per-row local-radius cap addresses both at once.

I CelebA control-set construction (least- vs. most-correlated five attributes)

For each of the 40 binary CelebA attributes used as a target concept, we report two complementary surgicality regimes (Section 5, Tabs. 7–9, Figure 7): the *5 least-correlated* attributes and the *5 most-correlated* attributes, both selected from the remaining 39 attributes. This appendix records exactly how the two control sets are computed.

Computation. Let $\mathbf{Y} \in \{0,1\}^{N_{\text{train}} \times 40}$ be the train-split label matrix, with $\mathbf{Y}_{i,k}$ the label of training example i on attribute k. We compute the empirical Pearson correlation matrix

$$\mathbf{R}_{k\ell} \;=\; \frac{\mathrm{Cov}(\mathbf{Y}_{:,k}, \mathbf{Y}_{:,\ell})}{\sqrt{\mathrm{Var}(\mathbf{Y}_{:,k}) \cdot \mathrm{Var}(\mathbf{Y}_{:,\ell})}} \quad \text{for } k, \ell \in \{1, \ldots, 40\}. \tag{10}$$

For each target attribute k, we then rank the remaining 39 attributes by absolute correlation $|\mathbf{R}_{k\ell}|$:

- the *5 least-correlated* controls are the bottom 5 by $|\mathbf{R}_{k\ell}|$. They share *minimal* information with the target and isolate clean surgicality.
- the *5 most-correlated* controls are the top 5 by $|\mathbf{R}_{k\ell}|$. They share *substantial* information with the target and form a stricter test of whether erasure can preserve genuinely related downstream features.

The target attribute itself is excluded from the candidate pool (i.e., $\ell \neq k$). Pairs are computed in float64 with no shrinkage.

Concrete examples. For *male*, the 5 most-correlated attributes are *wearing lipstick* ($|r| \approx 0.80$), *heavy makeup*, *no beard*, *attractive*, and *5 o'clock shadow* (stubble); the 5 least-correlated are *narrow eyes*, *blurry*, *straight hair*, *receding hairline*, and *pale skin* (all $|r| < 0.05$). Erasing *male* while preserving the most-correlated controls is genuinely hard because the controls are partial proxies for the target; erasing while preserving the least-correlated controls tests whether the intervention is geographically narrow but not whether it disentangles correlated information. We therefore report both regimes side by side: each one isolates a different failure mode.

Full per-target listing. Tabs. 11–12 list the two control sets for every one of the 40 target attributes, each with its absolute Pearson $|r|$ on the train split.

Why two regimes per concept. Reporting only one control set hides whichever regime the method happens to be weak in. A method that destroys the entire representation will pass the least-correlated test (controls are essentially independent of the edit anyway) and fail the most-correlated test catastrophically; conversely, a method that performs no erasure will pass both. The difference between the two regimes per method is therefore informative even when both individual numbers look acceptable. Section 5 reads the difference as a measure of how much an erasure method confuses "avoid this concept" with "avoid everything correlated with it".

J Closed-form per-row λ_i statistics

For each setting and each outer iteration of the proposed method (the manifold-constrained loop with fixed local-radius fraction $\varepsilon = 0.1$), the closed-form rule of Eq. 6 produces a per-row step magnitude λ_i (capped at $\lambda_{\max} = 64$). Table 13 reports the distribution of these per-row λ_i averaged across all 60 iterations of each trajectory and then aggregated across settings.

The mean across the entire grid, $\bar{\lambda} \approx 29.31$, is the value at which AmbCE^{++} appears in Tab. 2: forcing the unconstrained full-space method to take the same effective step magnitude that the closed-form per-row rule produces under MANCE^{++} isolates the gradient-direction constraint (tangent-projected vs. full-space) as the only remaining axis of variation between the two methods.

K Empirical support for the manifold premise

The geometric premise used by MCH (Section 3.1) states that natural representations concentrate on a low-dimensional manifold. We check this directly, independent of any erasure result. Tab. 14 reports the local intrinsic dimension (TwoNN, Facco et al., 2017) and the mean angle between a single global concept direction and the local per-sample concept directions, for the primary nonlinear case (sycophancy) and the diagnostic linear baseline (gender). For sycophancy the intrinsic dimension is 34–53, far below the dimension $d \in [768, 5376]$ of the Euclidean representation space, and the large local/global angles show that a single global direction is a poor first-order summary across neighborhoods, both consistent with concentration on a curved, low-dimensional structure rather than a global linear subspace.

L Anisotropy regimes for the local-σ spectrum

Where $\boldsymbol{\sigma}$ enters the update. The main method weights each tangent component by its singular value (Eq. 2). We can generalize this weighting by raising each singular value to a power α:

$$\mathbf{d}_i(\alpha) = \mathbf{B}_i \operatorname{diag}(\boldsymbol{\sigma}_i^{\alpha})\mathbf{c}_i. \tag{11}$$

The following choices illustrate how this exponent changes the weighting:

- $\alpha = 0$: the tangent direction $\mathbf{d}_i = \mathbf{B}_i\mathbf{c}_i$ is the bare projection of the unit gradient onto the local tangent basis; the singular spectrum plays no role in shaping the direction.
- $\alpha = 1$: each tangent direction is reweighted by $\sigma_{i,k}$. **This is the default in this paper.** It uses the local PCA spectrum directly: high-σ tangent axes receive more step mass, while thin axes receive less.
- $\alpha = 2$: each tangent direction is reweighted by $\sigma_{i,k}^2$. This is a stronger SVD-spectrum weighting with a Mahalanobis-style intuition, but it is not derived by solving a covariance-constrained optimization: we do not form or invert a local covariance matrix. Geometrically, it suppresses low-variance tangent directions most strongly, which may also suppress residual concept information carried by thin directions.

Mahalanobis-style intuition (and what $\alpha = 1$ is *not*). A reader familiar with Mahalanobis whitening might wonder whether our $\alpha = 1$ is the same operation. It is not. Classical Mahalanobis whitening explicitly uses a covariance inverse square root, $\mathbf{x}^{\text{white}} = \mathbf{C}^{-1/2}\mathbf{x}$; in the local-PCA coordinates this corresponds to scaling each tangent direction by $\sigma_{i,k}^{-1}$, i.e. $\alpha = -1$ in our notation. Whitening makes every direction equal-weighted, which would dilute the concept-specific signal we want to suppress and force every tangent axis to receive equal step magnitude.

We use a *positive* power, $\sigma_{i,k}^{+1}$, which does the opposite: it reweights the tangent step *toward* high-σ principal axes (the directions where the local manifold has the most support and where the linear approximation is most reliable) and away from low-σ thin tail directions where the linear approximation is less reliable and motion most easily leaves the manifold. The three non-negative values $\alpha \in \{0, 1, 2\}$ form a one-parameter family interpolating between (i) treating every tangent direction as equally reliable ($\alpha = 0$, isotropic), (ii) reweighting by $\sigma_{i,k}$ to respect the local ellipsoid mildly ($\alpha = 1$, our default), and (iii) a stronger spectral weighting that follows the local ellipsoid more aggressively and suppresses thin directions most ($\alpha = 2$). Whitening ($\alpha = -1$) sits outside this range for our problem because its sign is opposite the geometric correction we want; we therefore do not include it.

Table 3: **Closed-form preprocessing ablations on the 39-panel NLP grid (50%-depth layer).** A new nonlinear probe ($h = 128$, 200 SGD steps, patience 3) is trained on each post-erasure representation. *LEACE* is the affine LEACE eraser. *LEACE + CovMatch* additionally projects out the top-2 eigenvectors of $\Delta\Sigma = \Sigma_+ - \Sigma_-$ (rank-3 closed-form erasure total). *MANCE*$^{++}$ is the proposed full pipeline (LEACE + CovMatch preprocessing + iterative manifold-constrained loop, ε=0.1); we report the best $|S-\text{floor}|$ trajectory step subject to $D_Y \leq 5$. NL = nonlinear probe accuracy on test (lower is better; floor is the per-panel majority class). ΔY is the signed control-task change in points from the clean representation. Each setting takes the median over its 3 seeds, and for MANCE^{++} over the seeds within budget. '–' means fewer than 2 of 3 seeds are within budget. Bold marks the row's NL minimum.

Model	Concept	Floor (NL*)	LEACE NL	LEACE ΔY	LEACE + CovMatch NL	LEACE + CovMatch ΔY	MANCE^{++} ε=0.1 NL	MANCE^{++} ε=0.1 ΔY
Qwen2.5-0.5B	Sycophancy	0.550	**0.550**	0.0	**0.550**	-2.1	**0.550**	-1.9
	Gender	0.539	0.710	-0.1	0.668	-0.5	**0.539**	-2.7
	Safety	0.510	0.736	0.0	0.529	0.0	**0.510**	0.0
Qwen2.5-1.5B	Sycophancy	0.550	**0.550**	-0.1	**0.550**	-1.6	**0.550**	-1.9
	Gender	0.539	0.822	-0.3	0.789	-0.8	**0.539**	-3.0
	Safety	0.510	0.742	1.6	**0.510**	1.6	**0.510**	1.6
Qwen2.5-3B	Sycophancy	0.675	**0.675**	0.0	**0.675**	-3.9	**0.675**	-5.0
	Gender	0.539	0.863	-0.2	0.779	-0.3	**0.539**	-2.0
	Safety	0.510	0.713	0.5	**0.510**	0.5	**0.510**	0.5
Gemma-2-2B	Sycophancy	0.543	0.643	0.0	**0.543**	-0.4	**0.543**	-0.8
	Gender	0.539	0.722	0.0	0.655	-1.1	**0.539**	-2.2
	Safety	0.510	0.721	0.5	0.647	0.5	**0.510**	0.5
Gemma-2-9B	Sycophancy	0.569	0.756	0.0	**0.569**	-0.7	**0.569**	-0.7
	Gender	0.539	0.648	-0.1	0.562	-2.5	**0.539**	-4.4
	Safety	0.510	0.631	0.0	0.586	0.0	**0.510**	0.0
Gemma-2-27B	Sycophancy	0.553	0.695	-0.2	**0.553**	-0.9	**0.553**	-1.5
	Gender	0.539	0.676	-0.3	**0.598**	-5.2	–	–
	Safety	0.510	0.736	0.7	0.643	1.4	**0.510**	-1.0
Llama-3.2-1B	Sycophancy	0.551	0.852	0.0	**0.551**	-0.2	**0.551**	-0.2
	Gender	0.539	0.856	-0.3	0.738	-0.1	**0.539**	-1.4
	Safety	0.510	0.817	0.3	0.778	0.3	**0.514**	0.3
Llama-3.2-3B	Sycophancy	0.551	0.896	0.0	**0.551**	0.0	**0.551**	0.0
	Gender	0.539	0.897	-0.3	0.812	-0.4	**0.539**	-1.6
	Safety	0.510	0.837	0.1	0.820	0.2	**0.524**	0.1
Mistral-7B	Sycophancy	0.551	0.920	0.0	**0.551**	0.0	**0.551**	0.0
	Gender	0.539	0.852	-0.2	0.815	-0.4	**0.539**	-2.0
	Safety	0.510	0.825	0.2	0.811	0.2	**0.515**	0.2
Gemma-3-1B	Sycophancy	0.551	0.757	-0.1	**0.551**	-1.5	**0.551**	-1.5
	Gender	0.539	0.734	-0.9	0.613	-0.8	**0.539**	-1.1
	Safety	0.510	0.713	0.0	0.736	0.0	**0.510**	0.0
Gemma-3-4B	Sycophancy	0.551	**0.551**	0.0	**0.551**	-0.6	**0.551**	-1.2
	Gender	0.539	0.722	-0.5	0.598	-3.3	**0.539**	-3.4
	Safety	0.510	0.698	0.7	0.623	1.1	**0.510**	1.0
Gemma-3-12B	Sycophancy	0.556	0.601	0.0	0.561	-0.5	**0.556**	-0.8
	Gender	0.539	**0.539**	3.5	0.598	-8.7	**0.539**	-0.8
	Safety	0.510	0.587	0.0	**0.510**	0.0	**0.510**	0.0
Gemma-3-27B	Sycophancy	0.526	0.578	-0.2	0.566	-0.1	**0.526**	-0.2
	Gender	0.539	0.707	0.2	0.678	-2.6	**0.539**	-1.2
	Safety	0.510	0.686	0.9	**0.510**	0.7	**0.510**	0.9
Mean sycophancy		0.560	0.694	0.0	0.563	-1.0	**0.560**	-1.2
Mean gender		0.539	0.750	0.0	0.685	-2.1	**0.539**	-2.2
Mean safety		0.510	0.726	0.4	0.632	0.5	**0.512**	0.3

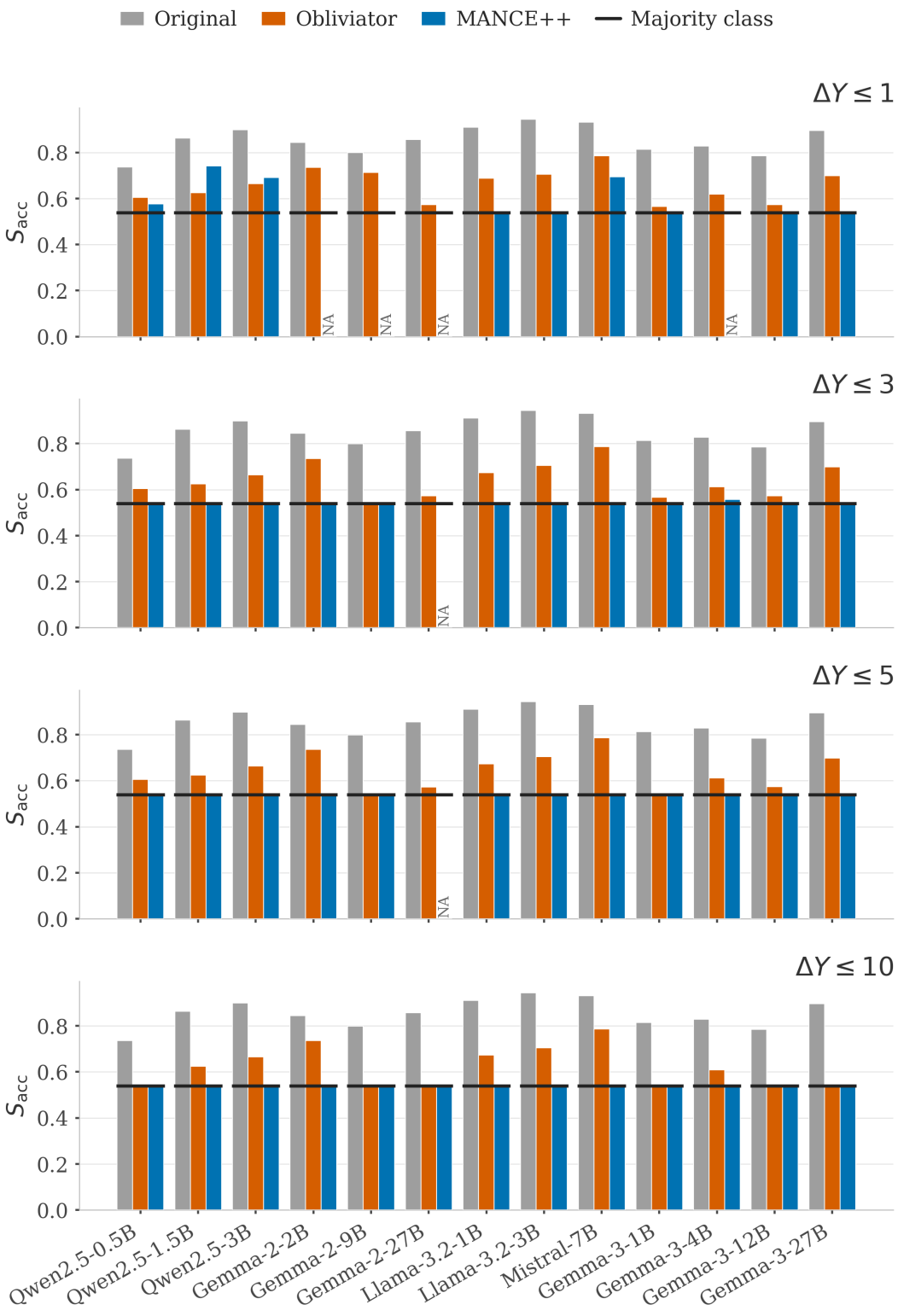


Figure 4: **Gender: per-model nonlinear-probe accuracy at all four surgicality budgets.** Each row is one budget $D_Y \in \{1, 3, 5, 10\}$. Each bar is the median over the seeds that stay within budget, and NA means fewer than 2 of 3 seeds do.

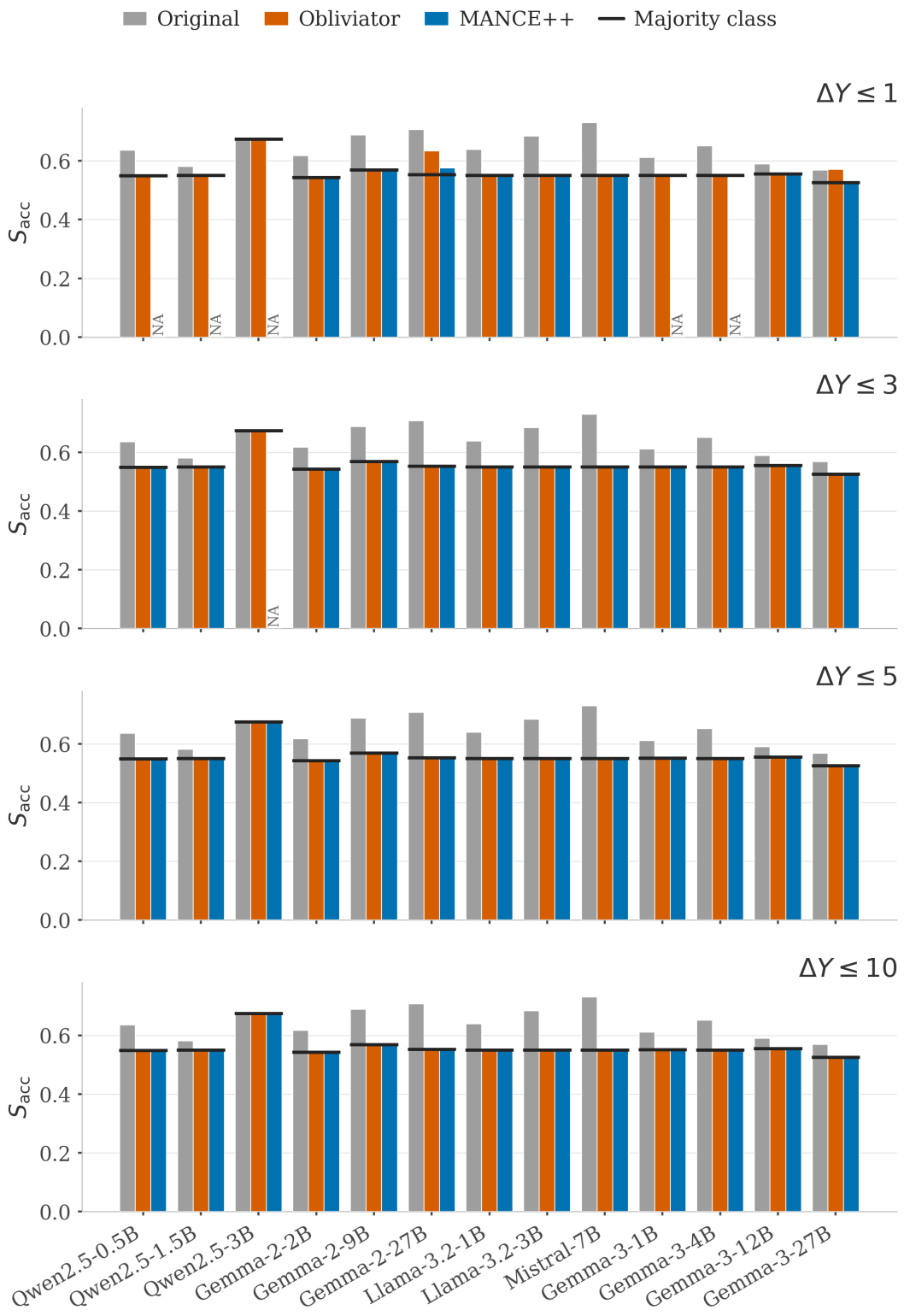


Figure 5: **Sycophancy: per-model nonlinear-probe accuracy at four surgicality budgets.** Layout matches Fig. 4.

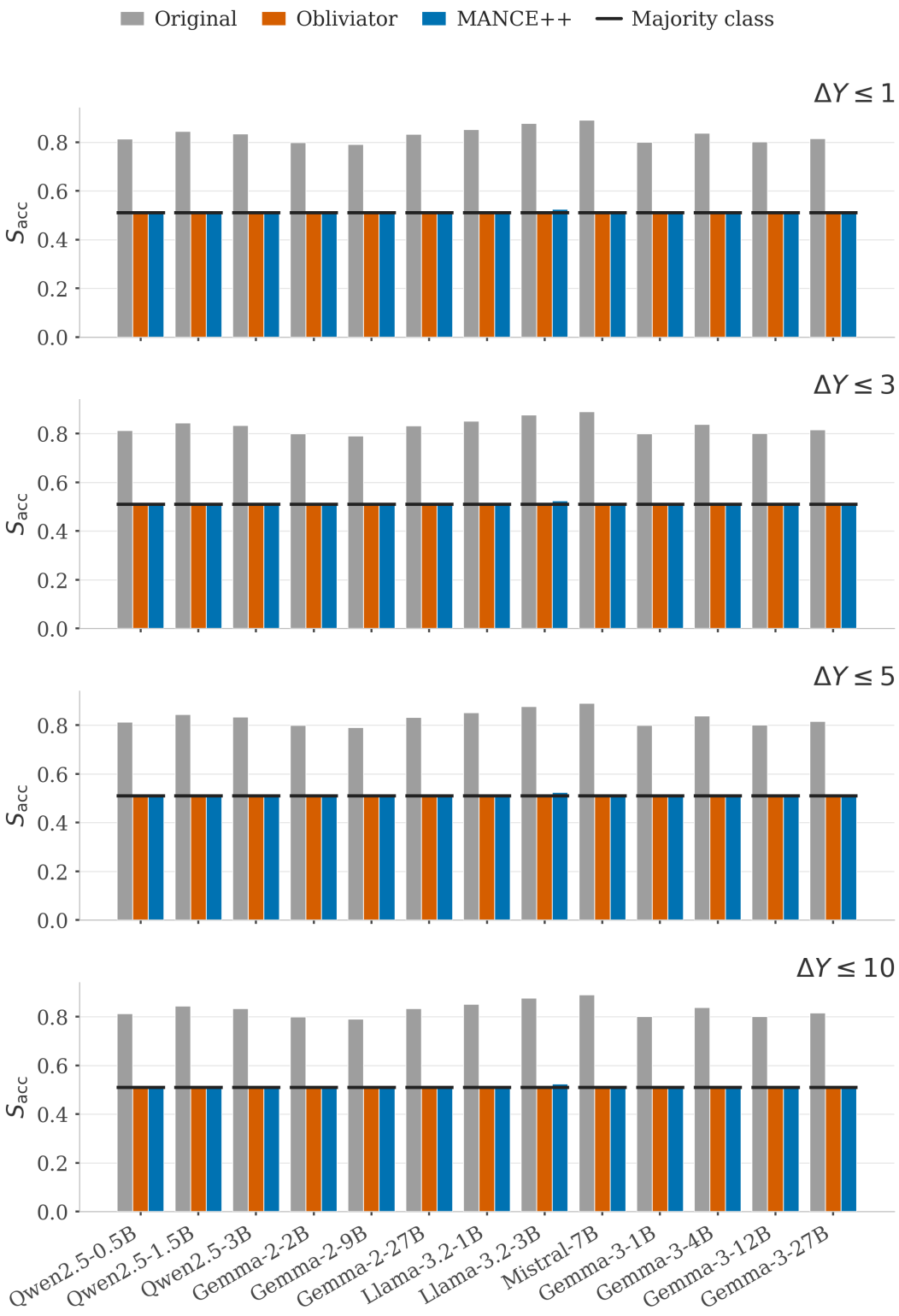


Figure 6: **Safety: per-model nonlinear-probe accuracy at four surgicality budgets.** Layout matches Fig. 4.

Table 4: **Variation across seeds.** Each leakage value of Tables 1, 2 and 6 (points) $\pm$ its standard deviation over the 3 seeds. '–' = no setting is covered.

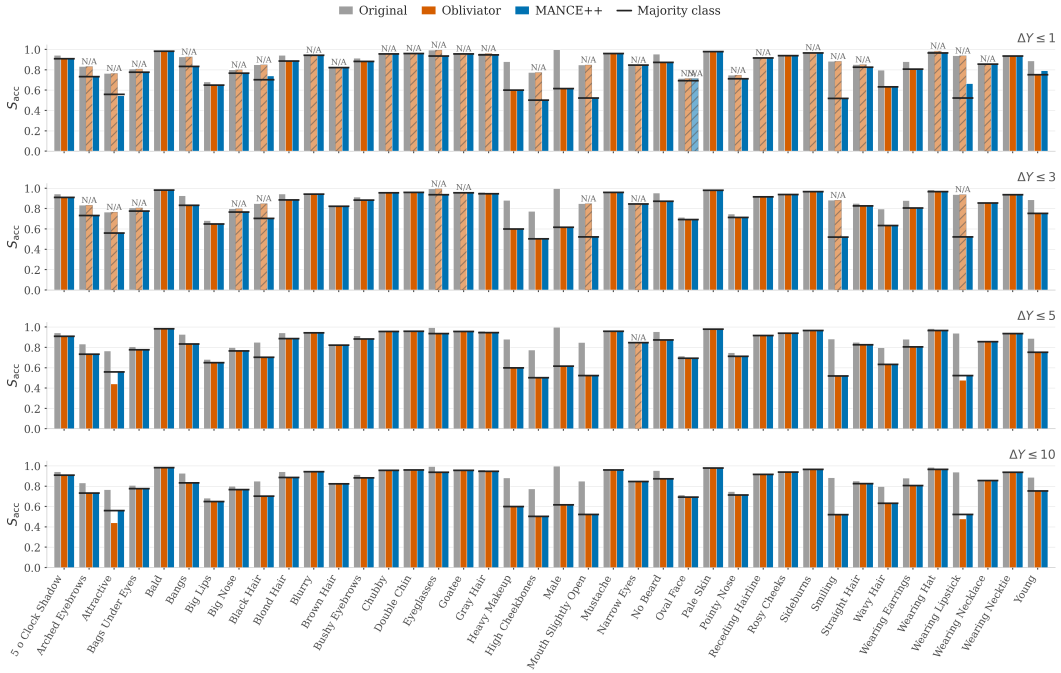
Method	$D_Y \leq 1$	$D_Y \leq 3$	$D_Y \leq 5$	$D_Y \leq 10$
NLP, Table 2 (signed D_S).				
LEACE	18.7 ± 0.3	18.7 ± 0.1	18.7 ± 0.1	18.7 ± 0.1
LEACE+CovMatch	10.5 ± 0.8	9.5 ± 0.6	9.2 ± 0.4	9.0 ± 0.3
INLP	14.1 ± 0.1	14.0 ± 0.1	13.9 ± 0.1	13.9 ± 0.1
IGBP	9.6 ± 0.2	9.6 ± 0.2	9.5 ± 0.2	9.5 ± 0.2
Obliviator	4.3 ± 0.5	3.4 ± 0.4	3.4 ± 0.4	2.6 ± 0.6
AmbCE^{++}	6.6 ± 1.1	10.9 ± 0.5	11.0 ± 0.2	10.8 ± 0.6
MANCE	5.2 ± 0.9	2.3 ± 0.1	2.0 ± 0.0	1.5 ± 0.1
MANCE^{+}	1.5 ± 0.6	0.6 ± 0.2	0.5 ± 0.1	0.4 ± 0.1
MANCE^{++}	2.0 ± 0.8	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0
NLP, Table 1, +MANCE column ($\lvert D_S \rvert$).				
MANCE (no preprocessing)	5.4 ± 0.4	2.2 ± 0.2	1.9 ± 0.3	1.6 ± 0.1
INLP	1.7 ± 0.3	1.3 ± 0.2	1.1 ± 0.2	1.0 ± 0.2
IGBP	1.4 ± 0.2	0.8 ± 0.0	0.6 ± 0.1	0.5 ± 0.0
Obliviator	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
LEACE	1.5 ± 0.6	0.6 ± 0.2	0.5 ± 0.1	0.5 ± 0.1
LEACE + CovMatch	2.1 ± 0.7	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0
CelebA, 5 least-correlated controls (signed D_S).				
LEACE	6.4 ± 0.2	6.4 ± 0.2	6.4 ± 0.2	6.4 ± 0.2
LEACE+CovMatch	5.4 ± 0.2	5.2 ± 0.1	5.2 ± 0.1	5.2 ± 0.1
INLP	5.5 ± 0.0	5.5 ± 0.0	5.5 ± 0.0	5.5 ± 0.0
IGBP	5.2 ± 0.3	5.2 ± 0.0	5.2 ± 0.0	5.2 ± 0.0
Obliviator	0.0 ± 0.0	0.0 ± 0.0	-0.4 ± 0.0	-0.4 ± 0.0
AmbCE^{++}	1.3 ± 0.7	3.0 ± 0.4	0.8 ± 0.1	0.4 ± 0.0
MANCE	1.1 ± 0.1	1.1 ± 0.1	1.1 ± 0.1	1.1 ± 0.1
MANCE^{+}	0.4 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
MANCE^{++}	0.5 ± 0.4	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
CelebA, 5 most-correlated controls (signed D_S).				
LEACE	4.3 ± 0.5	5.7 ± 0.6	6.4 ± 0.2	6.4 ± 0.2
LEACE+CovMatch	0.1 ± 1.4	4.3 ± 0.1	4.9 ± 0.4	5.2 ± 0.1
INLP	5.9 ± 0.1	5.6 ± 0.1	5.5 ± 0.0	5.5 ± 0.0
IGBP	6.5 ± 0.4	5.7 ± 0.1	5.3 ± 0.1	5.2 ± 0.1
Obliviator	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
AmbCE^{++}	-1.2 ± 1.5	5.5 ± 1.2	5.9 ± 0.3	4.6 ± 0.1
MANCE	4.2 ± 0.3	3.4 ± 0.3	2.4 ± 0.3	1.3 ± 0.3
MANCE^{+}	2.2 ± 0.1	3.1 ± 0.1	2.9 ± 0.2	1.3 ± 0.2
MANCE^{++}	0.3 ± 0.1	2.5 ± 0.1	2.4 ± 0.0	1.6 ± 0.3

Table 5: **Per-panel breakdown.** Signed D_S in points at the in-budget trajectory step with smallest $|D_S|$, subject to the surgicality budget named by each $D_Y \in \{1, 3, 5, 10\}$ block. Bold = best (smallest absolute deviation from floor) per panel per budget block. Each cell is the median over the seeds that keep the setting within budget. '–' = fewer than 2 of the 3 seeds keep the setting within budget. Final row: per-budget per-method mean across the 39 panels.

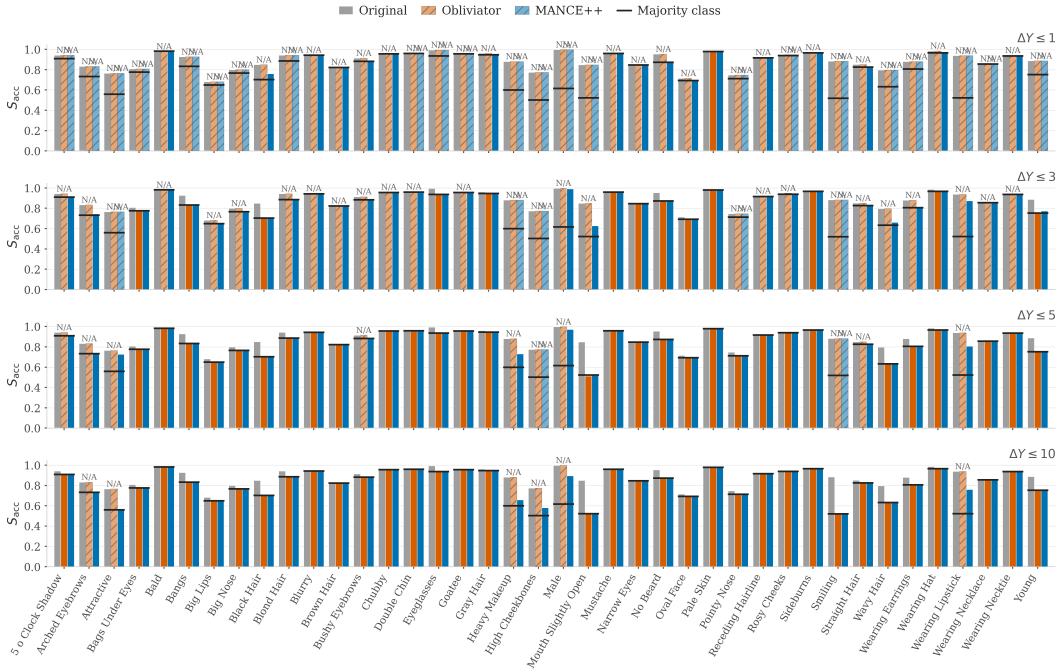
Panel	$D_Y \leq 1$							$D_Y \leq 3$							$D_Y \leq 5$							$D_Y \leq 10$						
	LE	L+C	INLP	IGBP	Obl	Amb++	MANCE++	LE	L+C	INLP	IGBP	Obl	Amb++	MANCE++	LE	L+C	INLP	IGBP	Obl	Amb++	MANCE++	LE	L+C	INLP	IGBP	Obl	Amb++	MANCE++
Gemma2-27B/Gen L23	13.7	–	13.0	12.8	**3.4**	–	–	13.7	–	11.2	12.8	**3.4**	–	–	13.7	–	11.1	12.8	**3.4**	–	–	13.7	5.9	11.1	12.8	**0.0**	10.5	**0.0**
Gemma2-27B/Saf L23	22.5	13.3	29.3	23.1	**0.0**	**0.0**	**0.0**	22.5	13.3	29.3	23.1	**0.0**	**0.0**	**0.0**	22.5	13.3	29.3	23.1	**0.0**	**0.0**	**0.0**	22.5	13.3	29.3	23.1	**0.0**	**0.0**	**0.0**
Gemma2-27B/Syc L23	14.3	**-0.1**	9.3	14.3	8.0	18.9	2.3	14.3	**0.0**	9.3	14.3	**0.0**	18.9	**0.0**	14.3	**0.0**	9.3	14.3	**0.0**	17.3	**0.0**	14.3	**0.0**	9.3	14.3	**0.0**	17.4	**0.0**
Gemma2-2B/Gen L13	18.3	–	16.0	**6.7**	19.6	–	–	18.3	11.6	16.0	6.7	19.6	17.4	**0.0**	18.3	11.6	16.0	6.7	19.6	18.4	**0.0**	18.3	11.6	16.0	6.7	19.6	18.4	**0.0**
Gemma2-2B/Saf L13	20.8	13.7	17.6	12.1	**0.0**	**0.0**	**0.0**	21.0	13.7	17.6	11.6	**0.0**	-0.1	**0.0**	21.0	13.7	17.6	11.6	**0.0**	-0.1	**0.0**	21.0	13.7	17.6	11.6	**0.0**	**0.0**	**0.0**
Gemma2-2B/Syc L13	9.9	**0.0**	1.0	0.3	**0.0**	11.4	**0.0**	9.9	**0.0**	1.0	0.3	**0.0**	11.4	**0.0**	9.9	**0.0**	1.0	0.3	**0.0**	11.4	**0.0**	9.9	**0.0**	1.0	0.3	**0.0**	10.6	**0.0**
Gemma2-9B/Gen L21	10.9	–	3.9	**2.3**	17.4	–	–	10.9	2.3	3.9	2.3	**0.0**	15.4	**0.0**	10.9	2.3	3.9	2.3	**0.0**	15.4	**0.0**	10.9	2.3	3.9	2.3	**0.0**	15.4	**0.0**
Gemma2-9B/Saf L21	12.0	7.5	17.3	8.0	**0.0**	**0.0**	**0.0**	12.0	7.5	17.3	8.0	**0.0**	**0.0**	**0.0**	12.0	7.5	17.3	8.0	**0.0**	**0.0**	**0.0**	12.0	7.5	17.3	8.0	**0.0**	**0.0**	**0.0**
Gemma2-9B/Syc L21	18.7	**0.0**	6.1	4.2	**0.0**	20.9	**0.0**	18.7	**0.0**	6.1	4.2	**0.0**	20.7	**0.0**	18.7	**0.0**	6.1	4.2	**0.0**	20.7	**0.0**	18.7	**0.0**	6.1	4.2	**0.0**	19.5	**0.0**
Gemma3-12B/Gen L24	**0.0**	–	9.2	17.5	3.4	8.8	**0.0**	**0.0**	–	9.4	17.5	3.4	8.8	**0.0**	**0.0**	–	9.4	17.5	3.4	8.8	**0.0**	**0.0**	5.9	8.0	17.5	**0.0**	8.9	**0.0**
Gemma3-12B/Saf L24	7.7	**0.0**	16.2	25.5	**0.0**	**0.0**	**0.0**	7.7	**0.0**	16.2	25.5	**0.0**	**0.0**	**0.0**	7.7	**0.0**	16.2	25.5	**0.0**	**0.0**	**0.0**	7.7	**0.0**	16.2	25.5	**0.0**	**0.0**	**0.0**
Gemma3-12B/Syc L24	4.5	0.5	**0.0**	3.4	**0.0**	6.4	**0.0**	4.5	0.5	**0.0**	3.4	**0.0**	6.5	**0.0**	4.5	0.5	**0.0**	3.4	**0.0**	6.5	**0.0**	4.5	0.5	**0.0**	3.4	**0.0**	6.5	**0.0**
Gemma3-1B/Gen L13	18.2	6.7	1.6	–	2.7	16.0	**0.2**	18.2	7.4	**0.0**	14.2	2.7	16.0	**0.0**	19.5	7.4	**0.0**	12.7	**0.0**	15.4	**0.0**	19.5	7.4	**0.0**	12.7	**0.0**	15.4	**0.0**
Gemma3-1B/Saf L13	20.2	22.5	18.0	21.3	**0.0**	**0.0**	**0.0**	20.2	22.5	18.0	21.3	**0.0**	**0.0**	**0.0**	20.2	22.5	18.0	21.3	**0.0**	**0.0**	**0.0**	20.2	22.5	18.0	21.3	**0.0**	**0.0**	**0.0**
Gemma3-1B/Syc L13	20.5	–	1.5	5.6	**0.0**	–	–	20.5	**0.0**	1.5	5.6	**0.0**	13.8	**0.0**	20.5	**0.0**	1.5	5.6	**0.0**	13.8	**0.0**	20.5	**0.0**	1.5	5.6	**0.0**	11.4	**0.0**
Gemma3-27B/Gen L31	16.4	–	25.6	24.4	15.9	–	**0.0**	16.4	13.4	25.6	24.4	15.9	–	**0.0**	16.4	13.4	23.9	24.4	15.9	–	**0.0**	16.4	13.9	23.9	24.4	**0.0**	–	**0.0**
Gemma3-27B/Saf L31	17.6	**0.0**	25.7	27.5	**0.0**	**0.0**	**0.0**	17.6	**0.0**	25.7	27.5	**0.0**	**0.0**	**0.0**	17.6	**0.0**	25.7	27.5	**0.0**	**0.0**	**0.0**	17.6	**0.0**	25.7	27.5	**0.0**	**0.0**	**0.0**
Gemma3-27B/Syc L31	5.2	4.0	1.0	2.5	4.5	7.1	**0.0**	5.2	4.0	1.0	2.5	**0.0**	7.1	**0.0**	5.2	4.0	1.0	2.5	**0.0**	7.1	**0.0**	5.2	4.0	1.0	2.5	**0.0**	6.3	**0.0**
Gemma3-4B/Gen L17	18.3	–	8.5	21.1	**7.9**	–	–	18.3	–	8.5	21.1	7.2	–	**1.8**	18.3	5.8	8.5	21.1	7.2	15.0	**0.0**	18.3	5.8	8.5	21.1	6.9	15.0	**0.0**
Gemma3-4B/Saf L17	18.8	11.3	27.7	26.9	**0.0**	**0.0**	**0.0**	18.8	11.3	27.7	26.9	**0.0**	**0.0**	**0.0**	18.8	11.3	27.7	26.9	**0.0**	**0.0**	**0.0**	18.8	11.3	27.7	26.9	**0.0**	**0.0**	**0.0**
Gemma3-4B/Syc L17	**0.0**	**0.0**	**0.0**	8.8	**0.0**	–	–	**0.0**	**0.0**	**0.0**	8.8	**0.0**	19.7	**0.0**	**0.0**	**0.0**	**0.0**	8.8	**0.0**	19.7	**0.0**	**0.0**	**0.0**	**0.0**	8.8	**0.0**	15.9	**0.0**
Llama3.2-1B/Gen L8	31.7	19.9	21.5	4.6	14.8	–	**-0.5**	31.7	19.9	19.8	2.3	13.4	25.8	**0.0**	31.7	19.9	19.8	2.3	13.4	26.0	**0.0**	31.7	19.9	19.8	2.3	13.4	26.0	**0.0**
Llama3.2-1B/Saf L8	30.7	26.7	29.1	9.3	**0.0**	-4.9	0.4	30.7	26.7	29.1	9.3	**0.0**	-4.9	0.4	30.7	26.7	29.1	9.3	**0.0**	-4.9	0.4	30.7	26.7	29.1	9.3	**0.0**	-2.1	0.4
Llama3.2-1B/Syc L8	30.1	**0.0**	**0.0**	5.2	**0.0**	–	**0.0**	30.1	**0.0**	**0.0**	5.2	**0.0**	–	**0.0**	30.1	**0.0**	**0.0**	5.2	**0.0**	–	**0.0**	30.1	**0.0**	**0.0**	5.2	**0.0**	–	**0.0**
Llama3.2-3B/Gen L14	35.8	27.3	31.3	3.3	16.5	–	**0.4**	35.8	27.3	31.3	2.6	16.5	30.3	**0.0**	35.8	27.3	31.3	2.6	16.5	30.3	**0.0**	35.8	27.3	31.3	2.6	16.5	30.3	**0.0**
Llama3.2-3B/Saf L14	32.7	31.0	33.3	10.2	**0.0**	**0.0**	1.4	32.7	31.0	32.9	10.2	**0.0**	**0.0**	1.4	32.7	31.0	32.9	10.2	**0.0**	**0.0**	1.4	32.7	31.0	32.9	10.2	**0.0**	**0.0**	1.4
Llama3.2-3B/Syc L14	34.5	**0.0**	**0.0**	4.6	**0.0**	–	**0.0**	34.5	**0.0**	**0.0**	4.6	**0.0**	–	**0.0**	34.5	**0.0**	**0.0**	4.6	**0.0**	–	**0.0**	34.5	**0.0**	**0.0**	4.6	**0.0**	–	**0.0**
Mistral-7B/Gen L16	31.3	27.6	29.5	**3.9**	24.7	–	15.5	31.3	27.6	29.5	3.7	24.7	28.9	**0.0**	31.3	27.6	29.5	3.7	24.7	28.7	**0.0**	31.3	27.6	29.5	3.7	24.7	28.7	**0.0**
Mistral-7B/Saf L16	31.5	30.0	34.8	8.8	**0.0**	-2.1	0.4	31.5	30.0	34.8	8.8	**0.0**	-2.1	0.4	31.5	30.0	34.8	8.8	**0.0**	-2.1	0.4	31.5	30.0	34.8	8.8	**0.0**	-2.1	0.4
Mistral-7B/Syc L16	37.0	**0.0**	**0.0**	10.9	**0.0**	33.1	**0.0**	37.0	**0.0**	**0.0**	10.9	**0.0**	33.0	**0.0**	37.0	**0.0**	**0.0**	10.9	**0.0**	33.0	**0.0**	37.0	**0.0**	**0.0**	10.9	**0.0**	33.0	**0.0**
Qwen2.5-0.5B/Gen L12	17.0	12.9	12.6	6.1	6.5	–	**3.7**	17.0	12.9	12.6	4.6	6.5	16.5	**-0.1**	17.0	12.9	12.6	4.6	6.5	16.8	**0.0**	17.0	12.9	12.6	4.6	**0.0**	16.8	**0.0**
Qwen2.5-0.5B/Saf L12	22.6	1.9	17.1	4.0	**0.0**	-0.1	**0.0**	22.6	1.9	17.1	4.0	**0.0**	**0.0**	**0.0**	22.6	1.9	17.1	4.0	**0.0**	**0.0**	**0.0**	22.6	1.9	17.1	4.0	**0.0**	**0.0**	**0.0**
Qwen2.5-0.5B/Syc L12	**0.0**	–	**0.0**	**0.0**	**0.0**	–	–	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	–	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	–	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	–	**0.0**
Qwen2.5-1.5B/Gen L14	28.2	25.0	22.4	**5.1**	8.5	27.1	20.2	28.2	25.0	22.4	4.1	8.5	27.3	**0.1**	28.2	25.0	22.4	4.1	8.5	27.3	**0.0**	28.2	25.0	22.4	4.1	8.5	27.3	**0.0**
Qwen2.5-1.5B/Saf L14	23.1	**0.0**	19.9	6.5	**0.0**	1.7	**0.0**	23.1	**0.0**	19.9	6.5	**0.0**	1.7	**0.0**	23.1	**0.0**	19.9	6.5	**0.0**	**0.0**	**0.0**	23.1	**0.0**	19.9	6.5	**0.0**	**0.0**	**0.0**
Qwen2.5-1.5B/Syc L14	**0.0**	–	**0.0**	**0.0**	**0.0**	–	–	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	–	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	–	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	–	**0.0**
Qwen2.5-3B/Gen L18	32.8	24.0	26.5	**4.7**	12.5	–	15.2	32.4	24.0	26.5	4.7	12.5	26.3	**0.0**	32.4	24.0	26.5	4.7	12.5	26.3	**0.0**	32.4	24.0	26.5	4.7	12.5	26.3	**0.0**
Qwen2.5-3B/Saf L18	20.3	**0.0**	23.1	9.3	**0.0**	0.1	**0.0**	20.3	**0.0**	23.1	9.3	**0.0**	**0.0**	**0.0**	20.3	**0.0**	23.1	9.3	**0.0**	**0.0**	**0.0**	20.3	**0.0**	23.1	9.3	**0.0**	**0.0**	**0.0**
Qwen2.5-3B/Syc L18	**0.0**	–	**0.0**	1.1	**0.0**	–	–	**0.0**	–	**0.0**	1.1	**0.0**	–	–	**0.0**	**0.0**	**0.0**	1.1	**0.0**	–	**0.0**	**0.0**	**0.0**	**0.0**	1.1	**0.0**	–	**0.0**
Mean	18.7	10.5	14.1	9.6	4.3	6.6	2.0	18.7	9.5	14.0	9.6	3.4	10.9	0.1	18.7	9.2	13.9	9.5	3.4	11.0	0.1	18.7	9.0	13.9	9.5	2.6	10.8	0.1

Table 6: **CelebA cross-modal results (CLIP ViT-B/32, 40 binary attributes).** *Leakage* (↓): how much target information survives, measured as D_S in points at the within-budget step closest to chance, so 0 is at chance and larger values mean more leakage. *Coverage* (↑): the number of the 40 attributes a method keeps within budget (§4.1). Leakage is averaged over the covered attributes. Each setting takes the median over the seeds that keep it within budget, and it counts as covered when at least 2 of the 3 seeds do. App. Table 4 reports the standard deviations. Each method is run on 40 attributes under each surgicality regime (5 least- vs. 5 most-correlated controls per target). The MANCE family is shown in three increments under one fixed configuration (eps=0.1, scorer h=512, 800 steps): MANCE (manifold loop only), MANCE^{+} (LEACE + manifold loop), MANCE^{++} (LEACE + rank-2 CovMatch + manifold loop, the full method). Best per column in **bold**. '–' = no in-budget step at that budget for that method.

Method	$D_Y \leq 1$		$D_Y \leq 3$		$D_Y \leq 5$		$D_Y \leq 10$	
	Leakage↓	Coverage↑	Leakage↓	Coverage↑	Leakage↓	Coverage↑	Leakage↓	Coverage↑
5 least-correlated controls (40 attributes).								
LEACE	6.4	40/40	6.4	40/40	6.4	40/40	6.4	40/40
LEACE+CovMatch	5.4	39/40	5.2	40/40	5.2	40/40	5.2	40/40
INLP	5.5	40/40	5.5	40/40	5.5	40/40	5.5	40/40
IGBP	5.2	40/40	5.2	40/40	5.2	40/40	5.2	40/40
Obliviator	**0.0**	15/40	**0.0**	29/40	−0.4	39/40	−0.4	40/40
AmbCE^{++}	1.3	19/40	3.0	38/40	0.8	40/40	0.4	40/40
MANCE	1.1	40/40	1.1	40/40	1.1	40/40	1.1	40/40
MANCE^{+}	0.4	40/40	**0.0**	40/40	**0.0**	40/40	**0.0**	40/40
MANCE^{++}	0.5	39/40	**0.0**	40/40	**0.0**	40/40	**0.0**	40/40
5 most-correlated controls (40 attributes).								
LEACE	4.3	22/40	5.7	37/40	6.4	40/40	6.4	40/40
LEACE+CovMatch	0.1	9/40	4.3	36/40	4.9	39/40	5.2	40/40
INLP	5.9	40/40	5.6	40/40	5.5	40/40	5.5	40/40
IGBP	6.5	39/40	5.7	40/40	5.3	40/40	5.2	40/40
Obliviator	**0.0**	1/40	**0.0**	13/40	**0.0**	30/40	**0.0**	34/40
AmbCE^{++}	−1.2	2/40	5.5	19/40	5.9	34/40	4.6	40/40
MANCE	4.2	40/40	3.4	40/40	2.4	40/40	1.3	40/40
MANCE^{+}	2.2	22/40	3.1	36/40	2.9	40/40	1.3	40/40
MANCE^{++}	0.3	17/40	2.5	35/40	2.4	38/40	1.6	40/40



(a) 5 least-correlated controls (highly disentangled control concepts).



(b) 5 most-correlated controls (highly entangled control concepts).

Figure 7: **CelebA per-attribute nonlinear-probe accuracy at four surgicality budgets.** Same bar-chart layout as the NLP per-model figures. For each attribute, three bars: Clean (gray, no erasure), MANCE^{++} (blue, the proposed method), Obliviator (orange, baseline). Black tick: per-attribute majority-class floor (lower bars are better; bars at the floor mean the probe is at chance). Hatched bars marked "N/A" = the method had no in-budget step at that D_Y; we substitute its clean (un-erased) S as a stand-in, so a hatched-orange tall bar next to a low solid-blue bar is a clean MANCE^{++} win on a setting Obliviator could not reach. Most-correlated regime at $D_Y \leq 1$: Obliviator stays within budget on $1/40$ attributes, while MANCE^{++} stays within budget on $17/40$. Each bar is the median over the seeds that stay within budget, and N/A means fewer than 2 of 3 seeds do.

Table 7: **CelebA per-attribute breakdown at $D_Y \leq 1$ (40 attributes $\times$ 2 surgicality regimes).** Signed D_S in points at the in-budget trajectory step with smallest $|D_S|$, subject to surgicality budget $D_Y \leq 1$ on the 5 controls. Bold = best (smallest absolute deviation from floor) per attribute per regime block. Each cell is the median over the seeds that keep the attribute within budget. '–' = fewer than 2 of the 3 seeds keep the attribute within budget. Final row: per-regime per-method mean.

Concept	5 least-correlated controls							5 most-correlated controls						
	LE	L+C	INLP	IGBP	Obl	Amb++	MANCE++	LE	L+C	INLP	IGBP	Obl	Amb++	MANCE++
5_o_Clock_Shadow	1.3	1.3	**0.0**	3.0	**0.0**	1.0	**0.0**	–	–	**0.0**	3.0	–	–	–
Arched_Eyebrows	9.7	2.7	6.3	7.3	–	–	**0.0**	–	–	**6.3**	8.0	–	–	–
Attractive	15.7	17.0	17.0	2.7	–	–	**-1.7**	–	–	17.0	**16.7**	–	–	–
Bags_Under_Eyes	0.3	**0.0**	**0.0**	0.3	–	–	**0.0**	–	–	**0.0**	0.3	–	–	–
Bald	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	–	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	–	**0.0**	**0.0**
Bangs	4.7	**0.0**	**0.0**	4.3	–	–	**0.0**	4.2	–	**0.0**	4.3	–	–	–
Big_Lips	1.3	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	–	–	**0.0**	**0.0**	–	–	–
Big_Nose	2.7	2.3	**0.0**	3.0	–	–	**0.0**	–	–	**0.0**	2.3	–	–	–
Black_Hair	11.0	9.3	8.3	4.3	–	–	**3.7**	11.0	–	8.3	**4.3**	–	–	5.7
Blond_Hair	5.0	2.0	3.0	3.7	**0.0**	**0.0**	**0.0**	5.0	–	**3.0**	3.7	–	–	–
Blurry	**0.0**	**0.0**	**0.0**	**0.0**	–	–	**0.0**	–	–	**0.0**	**0.0**	–	–	**0.0**
Brown_Hair	0.7	-1.0	**0.0**	1.0	–	–	**0.0**	0.7	-2.0	**0.0**	1.0	–	–	**0.0**
Bushy_Eyebrows	0.3	**0.0**	0.7	3.7	**0.0**	-0.5	**0.0**	–	–	0.7	3.7	–	–	**0.0**
Chubby	**0.0**	**0.0**	**0.0**	0.3	–	0.5	**0.0**	**0.0**	**0.0**	**0.0**	0.3	–	–	**0.0**
Double_Chin	**0.0**	**0.0**	**0.0**	**0.0**	–	**0.0**	**0.0**	**0.0**	–	**0.0**	**0.0**	–	–	–
Eyeglasses	3.0	**0.0**	**0.0**	6.0	–	–	**0.0**	3.0	–	**0.0**	6.0	–	–	–
Goatee	**0.0**	**0.0**	**0.0**	**0.0**	–	–	**0.0**	**0.0**	–	**0.0**	**0.0**	–	–	–
Gray_Hair	**0.0**	**0.0**	**0.0**	2.0	–	–	**0.0**	**0.0**	**0.0**	**0.0**	2.0	–	–	**0.0**
Heavy_Makeup	26.3	26.0	26.0	22.7	**0.0**	-2.3	**0.0**	–	–	27.0	**26.3**	–	–	–
High_Cheekbones	21.7	18.7	20.7	12.0	–	–	**0.0**	–	–	**25.0**	–	–	–	–
Male	37.7	37.7	35.7	32.0	**0.0**	**0.0**	**0.0**	37.5	–	**36.0**	**36.0**	–	–	–
Mouth_Slightly_Open	16.0	11.7	14.7	5.3	–	–	**0.0**	–	–	15.7	**8.7**	–	–	–
Mustache	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	–	**0.0**	**0.0**	–	–	**0.0**
Narrow_Eyes	**0.0**	**0.0**	**0.0**	**0.0**	–	–	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	–	–	**0.0**
No_Beard	6.0	3.0	5.7	8.0	**0.0**	**0.0**	**0.0**	5.8	3.0	5.7	8.0	–	–	**0.0**
Oval_Face	0.3	–	**0.0**	1.0	–	–	–	–	–	**0.0**	**0.0**	–	–	**0.0**
Pale_Skin	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	–	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	–	**0.0**
Pointy_Nose	-5.0	-4.0	-0.3	1.0	–	**0.0**	**0.0**	–	–	**-0.3**	1.0	–	-2.3	–
Receding_Hairline	**0.0**	**0.0**	**0.0**	0.3	–	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	0.3	–	–	**0.0**
Rosy_Cheeks	0.3	**0.0**	**0.0**	1.3	**0.0**	**0.0**	**0.0**	–	–	**0.0**	1.3	–	–	–
Sideburns	**0.0**	**0.0**	**0.0**	0.7	–	**0.0**	**0.0**	0.2	–	**0.0**	0.7	–	–	**0.0**
Smiling	29.0	26.3	22.3	16.3	–	25.5	**0.0**	–	–	**33.7**	35.2	–	–	–
Straight_Hair	0.3	0.3	**0.0**	2.3	–	**0.0**	**0.0**	0.2	–	**0.0**	2.3	–	–	**0.0**
Wavy_Hair	14.0	10.7	9.3	12.3	**0.0**	**0.0**	**0.0**	14.8	–	**9.3**	15.3	–	–	–
Wearing_Earrings	2.7	**0.0**	0.3	6.7	**0.0**	**0.0**	**0.0**	–	–	**0.3**	6.7	–	–	–
Wearing_Hat	**0.0**	**0.0**	**0.0**	1.7	–	**0.0**	**0.0**	0.2	–	**0.0**	1.7	–	–	**0.0**
Wearing_Lipstick	39.7	38.0	39.3	27.7	–	–	**14.0**	–	–	39.3	**38.3**	–	–	–
Wearing_Necklace	**0.0**	**0.0**	**0.0**	**0.0**	–	–	**0.0**	–	–	**0.0**	**0.0**	–	–	–
Wearing_Necktie	**0.0**	**0.0**	**0.0**	1.7	**0.0**	–	**0.0**	0.7	**0.0**	**0.0**	1.7	–	–	**0.0**
Young	12.0	9.2	9.7	13.7	**0.0**	–	3.7	11.5	–	**9.7**	13.7	–	–	–
Mean	6.4	5.4	5.5	5.2	0.0	1.3	0.5	4.3	0.1	5.9	6.5	0.0	-1.2	0.3

Table 8: **CelebA per-attribute breakdown at $D_Y \leq 3$ (40 attributes $\times$ 2 surgicality regimes).** Signed D_S in points at the in-budget trajectory step with smallest $|D_S|$, subject to surgicality budget $D_Y \leq 3$ on the 5 controls. Bold = best (smallest absolute deviation from floor) per attribute per regime block. Each cell is the median over the seeds that keep the attribute within budget. '–' = fewer than 2 of the 3 seeds keep the attribute within budget. Final row: per-regime per-method mean.

Concept	5 least-correlated controls							5 most-correlated controls						
	LE	L+C	INLP	IGBP	Obl	Amb++	MANCE++	LE	L+C	INLP	IGBP	Obl	Amb++	MANCE++
5_o_Clock_Shadow	1.3	1.3	**0.0**	3.0	**0.0**	**0.0**	**0.0**	1.3	1.3	**0.0**	3.0	–	–	**0.0**
Arched_Eyebrows	9.7	2.7	6.3	7.3	–	4.7	**0.0**	–	2.7	6.3	7.3	–	4.2	**1.0**
Attractive	15.7	17.0	17.0	2.7	–	–	**0.0**	–	–	17.0	**2.7**	–	–	–
Bags_Under_Eyes	0.3	**0.0**	**0.0**	0.3	–	1.3	**0.0**	0.3	0.2	**0.0**	0.3	**0.0**	–	**0.0**
Bald	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	–	**0.0**	**0.0**
Bangs	4.7	**0.0**	**0.0**	4.3	**0.0**	1.0	**0.0**	4.7	**0.0**	**0.0**	4.3	**0.0**	–	**0.0**
Big_Lips	1.3	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	1.3	**0.0**	**0.0**	**0.0**	–	–	**0.0**
Big_Nose	2.7	2.3	**0.0**	2.3	–	-0.7	**0.0**	2.7	2.3	**0.0**	2.3	–	-0.7	-0.7
Black_Hair	11.0	9.0	8.3	4.3	–	7.7	**0.7**	11.0	9.0	8.3	4.3	**0.0**	7.7	0.3
Blond_Hair	5.0	2.0	3.0	3.7	**0.0**	**0.0**	**0.0**	5.0	2.0	3.0	3.7	–	–	**0.0**
Blurry	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	–	**0.0**	**0.0**
Brown_Hair	0.7	-2.0	**0.0**	1.0	**0.0**	0.7	**0.0**	0.7	-2.0	**0.0**	1.0	–	–	**0.0**
Bushy_Eyebrows	0.3	**0.0**	0.7	3.7	**0.0**	-0.3	**0.0**	0.3	**0.0**	0.7	3.7	–	–	**0.0**
Chubby	**0.0**	**0.0**	**0.0**	0.3	**0.0**	0.3	**0.0**	**0.0**	**0.0**	**0.0**	0.3	–	**0.0**	**0.0**
Double_Chin	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	–	**0.0**	**0.0**
Eyeglasses	3.0	**0.0**	**0.0**	6.0	–	**0.0**	**0.0**	3.0	**0.0**	**0.0**	6.0	**0.0**	–	**0.0**
Goatee	**0.0**	**0.0**	**0.0**	**0.0**	–	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	–	–	**0.0**
Gray_Hair	**0.0**	**0.0**	**0.0**	2.0	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	2.0	**0.0**	**0.0**	**0.0**
Heavy_Makeup	26.3	26.0	26.0	22.7	**0.0**	**0.0**	**0.0**	26.3	26.0	26.0	**22.7**	–	–	–
High_Cheekbones	21.7	18.7	20.7	12.0	**0.0**	17.3	**0.0**	–	–	23.3	**14.7**	–	–	–
Male	37.7	37.7	35.7	32.0	**0.0**	**0.0**	**0.0**	37.7	37.7	35.7	**32.7**	–	37.3	37.3
Mouth_Slightly_Open	16.0	11.7	14.7	5.3	–	13.7	**0.0**	16.0	12.7	14.7	**5.3**	–	–	10.3
Mustache	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	–	**0.0**
Narrow_Eyes	**0.0**	**0.0**	**0.0**	**0.0**	–	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
No_Beard	6.0	3.0	5.7	8.0	**0.0**	**0.0**	**0.0**	6.0	3.0	5.7	8.0	**0.0**	4.3	**0.0**
Oval_Face	0.3	**0.0**	**0.0**	**0.0**	**0.0**	-2.0	**0.0**	0.3	**0.0**	**0.0**	**0.0**	**0.0**	–	**0.0**
Pale_Skin	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	–	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	–	**0.0**
Pointy_Nose	-5.0	-4.0	-0.3	1.0	**0.0**	**0.0**	**0.0**	-4.3	–	**-0.3**	1.0	–	-2.0	–
Receding_Hairline	**0.0**	**0.0**	**0.0**	0.3	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	0.3	–	**0.0**	**0.0**
Rosy_Cheeks	0.3	**0.0**	**0.0**	1.3	**0.0**	**0.0**	**0.0**	0.3	**0.0**	**0.0**	1.3	–	–	**0.0**
Sideburns	**0.0**	**0.0**	**0.0**	0.7	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	0.7	**0.0**	**0.0**	**0.0**
Smiling	29.0	26.3	22.3	16.3	–	25.3	**0.0**	29.0	–	**24.3**	28.7	–	–	–
Straight_Hair	0.3	0.3	**0.0**	2.3	**0.0**	**0.0**	**0.0**	0.3	0.3	**0.0**	2.3	–	0.5	**0.0**
Wavy_Hair	14.0	10.7	9.3	12.3	**0.0**	**0.0**	**0.0**	14.0	11.3	9.3	12.3	–	8.0	**2.7**
Wearing_Earrings	2.7	**0.0**	0.3	6.7	**0.0**	**0.0**	**0.0**	2.7	**0.0**	0.3	6.7	–	–	**0.0**
Wearing_Hat	**0.0**	**0.0**	**0.0**	1.7	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	1.7	**0.0**	**0.0**	**0.0**
Wearing_Lipstick	39.7	38.0	39.3	27.7	–	37.3	**0.0**	39.7	38.0	39.3	**31.3**	–	37.3	35.0
Wearing_Necklace	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	–	–	**0.0**
Wearing_Necktie	**0.0**	**0.0**	**0.0**	1.7	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	1.7	–	–	**0.0**
Young	12.0	9.0	9.7	13.7	**0.0**	7.7	**0.0**	12.0	9.0	9.7	13.7	**0.0**	8.2	2.0
Mean	6.4	5.2	5.5	5.2	0.0	3.0	0.0	5.7	4.3	5.6	5.7	0.0	5.5	2.5

Table 9: **CelebA per-attribute breakdown at $D_Y \leq 5$ (40 attributes $\times$ 2 surgicality regimes).** Signed D_S in points at the in-budget trajectory step with smallest $|D_S|$, subject to surgicality budget $D_Y \leq 5$ on the 5 controls. Bold = best (smallest absolute deviation from floor) per attribute per regime block. Each cell is the median over the seeds that keep the attribute within budget. '–' = fewer than 2 of the 3 seeds keep the attribute within budget. Final row: per-regime per-method mean.

Concept	5 least-correlated controls							5 most-correlated controls						
	LE	L+C	INLP	IGBP	Obl	Amb++	MANCE++	LE	L+C	INLP	IGBP	Obl	Amb++	MANCE++
5_o_Clock_Shadow	1.3	1.3	**0.0**	3.0	**0.0**	**0.0**	**0.0**	1.3	1.3	**0.0**	3.0	–	0.7	**0.0**
Arched_Eyebrows	9.7	2.7	6.3	7.3	**0.0**	-0.3	**0.0**	9.7	2.7	6.3	7.3	–	4.7	**0.0**
Attractive	15.7	17.0	17.0	2.7	-12.0	16.0	**0.0**	15.7	17.0	17.0	**2.7**	–	16.0	16.7
Bags_Under_Eyes	0.3	**0.0**	**0.0**	0.3	**0.0**	**0.0**	**0.0**	0.3	**0.0**	**0.0**	0.3	**0.0**	1.3	**0.0**
Bald	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
Bangs	4.7	**0.0**	**0.0**	4.3	**0.0**	0.3	**0.0**	4.7	**0.0**	**0.0**	4.3	**0.0**	0.3	**0.0**
Big_Lips	1.3	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	1.3	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
Big_Nose	2.7	2.3	**0.0**	2.3	**0.0**	**0.0**	**0.0**	2.7	2.3	**0.0**	2.3	**0.0**	-0.3	**0.0**
Black_Hair	11.0	9.0	8.3	4.3	**0.0**	7.7	**0.0**	11.0	9.0	8.3	4.3	**0.0**	**0.0**	**0.0**
Blond_Hair	5.0	2.0	3.0	3.7	**0.0**	**0.0**	**0.0**	5.0	2.0	3.0	3.7	**0.0**	3.7	**0.0**
Blurry	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
Brown_Hair	0.7	-2.0	**0.0**	1.0	**0.0**	0.7	**0.0**	0.7	-2.0	**0.0**	1.0	**0.0**	–	**0.0**
Bushy_Eyebrows	0.3	**0.0**	0.7	3.7	**0.0**	-0.3	**0.0**	0.3	**0.0**	0.7	3.7	–	–	**0.0**
Chubby	**0.0**	**0.0**	**0.0**	0.3	**0.0**	0.3	**0.0**	**0.0**	**0.0**	**0.0**	0.3	**0.0**	**0.0**	**0.0**
Double_Chin	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
Eyeglasses	3.0	**0.0**	**0.0**	6.0	**0.0**	**0.0**	**0.0**	3.0	**0.0**	**0.0**	6.0	**0.0**	–	**0.0**
Goatee	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
Gray_Hair	**0.0**	**0.0**	**0.0**	2.0	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	2.0	**0.0**	**0.0**	**0.0**
Heavy_Makeup	26.3	26.0	26.0	22.7	**0.0**	**0.0**	**0.0**	26.3	26.0	26.0	22.7	–	25.3	**13.0**
High_Cheekbones	21.7	18.7	20.7	12.0	**0.0**	0.3	**0.0**	21.7	–	20.7	**12.0**	–	19.0	–
Male	37.7	37.7	35.7	32.0	**0.0**	**0.0**	**0.0**	37.7	37.7	35.7	**32.0**	–	37.3	35.3
Mouth_Slightly_Open	16.0	11.7	14.7	5.3	**0.0**	0.7	**0.0**	16.0	11.7	14.7	5.3	-1.2	13.7	**-0.7**
Mustache	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	–	**0.0**
Narrow_Eyes	**0.0**	**0.0**	**0.0**	**0.0**	–	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
No_Beard	6.0	3.0	5.7	8.0	**0.0**	**0.0**	**0.0**	6.0	3.0	5.7	8.0	**0.0**	4.3	**0.0**
Oval_Face	0.3	**0.0**	**0.0**	**0.0**	**0.0**	-0.3	**0.0**	0.3	**0.0**	**0.0**	**0.0**	**0.0**	-2.0	**0.0**
Pale_Skin	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
Pointy_Nose	-5.0	-4.0	-0.3	1.0	**0.0**	**0.0**	**0.0**	-5.0	-4.0	-0.3	1.0	**0.0**	-2.0	**0.0**
Receding_Hairline	**0.0**	**0.0**	**0.0**	0.3	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	0.3	**0.0**	**0.0**	**0.0**
Rosy_Cheeks	0.3	**0.0**	**0.0**	1.3	**0.0**	**0.0**	**0.0**	0.3	**0.0**	**0.0**	1.3	**0.0**	–	**0.0**
Sideburns	**0.0**	**0.0**	**0.0**	0.7	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	0.7	**0.0**	**0.0**	**0.0**
Smiling	29.0	26.3	22.3	16.3	**0.0**	**0.0**	**0.0**	29.0	26.2	22.3	**17.3**	–	25.3	–
Straight_Hair	0.3	0.3	**0.0**	2.3	**0.0**	**0.0**	**0.0**	0.3	0.3	**0.0**	2.3	–	0.3	**0.0**
Wavy_Hair	14.0	10.7	9.3	12.3	**0.0**	**0.0**	**0.0**	14.0	10.7	9.3	12.3	**0.0**	8.0	**0.0**
Wearing_Earrings	2.7	**0.0**	0.3	6.7	**0.0**	**0.0**	**0.0**	2.7	**0.0**	0.3	6.7	**0.0**	0.5	**0.0**
Wearing_Hat	**0.0**	**0.0**	**0.0**	1.7	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	1.7	**0.0**	**0.0**	**0.0**
Wearing_Lipstick	39.7	38.0	39.3	27.7	-4.7	**0.0**	**0.0**	39.7	38.0	39.3	30.0	–	37.3	**28.3**
Wearing_Necklace	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
Wearing_Necktie	**0.0**	**0.0**	**0.0**	1.7	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	1.7	**0.0**	–	**0.0**
Young	12.0	9.0	9.7	13.7	**0.0**	7.7	**0.0**	12.0	9.0	9.7	13.7	**0.0**	7.7	0.3
Mean	6.4	5.2	5.5	5.2	-0.4	0.8	0.0	6.4	4.9	5.5	5.3	0.0	5.9	2.4

Table 10: **Fit-and-apply wall-clock latency.** One representative panel (Qwen2.5-1.5B, gender, layer 14), every method under its canonical configuration from Table 2, all timed sequentially on the same dedicated GPU (one NVIDIA B200, 8 CPU cores). Diagnostic per-round probes used only for reporting are excluded. INLP's fit is CPU-bound (per-round logistic-regression fits), so its wall-clock scales with the CPU allocation rather than the GPU.

Method	Wall-clock (s)
LEACE	3.7
LEACE + CovMatch	7.3
AmbCE^{++}	19.8
IGBP	20.3
Obliviator	23.6
INLP	286.9
MANCE	458.8
MANCE^{+}	470.5
MANCE^{++}	474.9

Table 11: **CelebA least-correlated control sets (highly disentangled control concepts).** For each of the 40 target attributes (rows), the 5 control attributes selected as the bottom-5 by absolute Pearson $|r|$ on the train split, with each $|r|$ value in parentheses. These are the 5 attributes that share the least information with the target and isolate clean surgicality.

Target	5 least-correlated controls (with $\|r\|$)				
5_o_Clock_Shadow	Receding_Hairline (0.00)	Double_Chin (0.01)	Young (0.01)	Bald (0.01)	Blurry (0.02)
Arched_Eyebrows	Oval_Face (0.00)	Bushy_Eyebrows (0.01)	Bangs (0.01)	Receding_Hairline (0.02)	Black_Hair (0.03)
Attractive	Straight_Hair (0.00)	Bushy_Eyebrows (0.01)	Black_Hair (0.01)	Mouth_Slightly_Open (0.04)	Wearing_Necklace (0.08)
Bags_Under_Eyes	Straight_Hair (0.00)	Wearing_Necklace (0.01)	Big_Lips (0.01)	Wearing_Hat (0.01)	Black_Hair (0.02)
Bald	Wearing_Hat (0.00)	Mouth_Slightly_Open (0.01)	Smiling (0.01)	5_o_Clock_Shadow (0.01)	High_Cheekbones (0.03)
Bangs	Black_Hair (0.00)	Brown_Hair (0.01)	Mouth_Slightly_Open (0.01)	Narrow_Eyes (0.01)	Arched_Eyebrows (0.01)
Big_Lips	Chubby (0.00)	Mustache (0.00)	Goatee (0.00)	Bags_Under_Eyes (0.01)	Double_Chin (0.01)
Big_Nose	Wearing_Necklace (0.01)	Straight_Hair (0.03)	Rosy_Cheeks (0.03)	Wearing_Earrings (0.03)	Black_Hair (0.05)
Black_Hair	Bangs (0.00)	Smiling (0.00)	Mouth_Slightly_Open (0.01)	Attractive (0.01)	Chubby (0.01)
Blond_Hair	Blurry (0.00)	Straight_Hair (0.01)	Pale_Skin (0.03)	Oval_Face (0.04)	Narrow_Eyes (0.05)
Blurry	Blond_Hair (0.00)	Sideburns (0.00)	Goatee (0.00)	No_Beard (0.01)	Wearing_Hat (0.01)
Brown_Hair	Wearing_Necklace (0.00)	Mouth_Slightly_Open (0.00)	Bangs (0.01)	Straight_Hair (0.02)	High_Cheekbones (0.02)
Bushy_Eyebrows	Chubby (0.00)	Narrow_Eyes (0.00)	Smiling (0.00)	Wearing_Hat (0.01)	Arched_Eyebrows (0.01)
Chubby	Big_Lips (0.00)	Bushy_Eyebrows (0.00)	Mouth_Slightly_Open (0.01)	Wearing_Necklace (0.01)	High_Cheekbones (0.01)
Double_Chin	5_o_Clock_Shadow (0.01)	Rosy_Cheeks (0.01)	Wearing_Necklace (0.01)	Big_Lips (0.01)	Sideburns (0.02)
Eyeglasses	Straight_Hair (0.01)	Blurry (0.01)	Black_Hair (0.02)	Narrow_Eyes (0.03)	Pale_Skin (0.03)
Goatee	Blurry (0.00)	Big_Lips (0.00)	Double_Chin (0.02)	Receding_Hairline (0.03)	Black_Hair (0.03)
Gray_Hair	Sideburns (0.01)	Narrow_Eyes (0.02)	Bushy_Eyebrows (0.02)	Smiling (0.02)	Straight_Hair (0.03)
Heavy_Makeup	Pale_Skin (0.07)	Receding_Hairline (0.07)	Straight_Hair (0.07)	Narrow_Eyes (0.07)	Brown_Hair (0.10)
High_Cheekbones	Chubby (0.01)	Straight_Hair (0.02)	Brown_Hair (0.02)	Young (0.02)	Bushy_Eyebrows (0.02)
Male	Narrow_Eyes (0.02)	Blurry (0.02)	Straight_Hair (0.05)	Receding_Hairline (0.08)	Pale_Skin (0.10)
Mouth_Slightly_Open	Brown_Hair (0.00)	Chubby (0.01)	Bangs (0.01)	Bald (0.01)	Black_Hair (0.01)
Mustache	Big_Lips (0.00)	Pale_Skin (0.01)	Blurry (0.01)	Narrow_Eyes (0.02)	Bald (0.03)
Narrow_Eyes	Wearing_Earrings (0.00)	Bushy_Eyebrows (0.00)	Sideburns (0.00)	Bangs (0.01)	No_Beard (0.01)
No_Beard	Blurry (0.01)	Narrow_Eyes (0.01)	Straight_Hair (0.02)	Double_Chin (0.03)	Receding_Hairline (0.04)
Oval_Face	Pointy_Nose (0.00)	Straight_Hair (0.00)	Arched_Eyebrows (0.00)	Receding_Hairline (0.01)	Wavy_Hair (0.01)
Pale_Skin	Wearing_Hat (0.00)	Mustache (0.01)	Wearing_Necktie (0.02)	Wavy_Hair (0.02)	Bushy_Eyebrows (0.02)
Pointy_Nose	Oval_Face (0.00)	Straight_Hair (0.00)	Receding_Hairline (0.01)	5_o_Clock_Shadow (0.03)	Pale_Skin (0.03)
Receding_Hairline	5_o_Clock_Shadow (0.00)	Rosy_Cheeks (0.00)	Straight_Hair (0.01)	Oval_Face (0.01)	Pointy_Nose (0.01)
Rosy_Cheeks	Receding_Hairline (0.00)	Double_Chin (0.01)	Chubby (0.02)	Narrow_Eyes (0.03)	Bushy_Eyebrows (0.03)
Sideburns	Blurry (0.00)	Narrow_Eyes (0.00)	Gray_Hair (0.01)	Straight_Hair (0.01)	Wearing_Necktie (0.01)
Smiling	Bushy_Eyebrows (0.00)	Black_Hair (0.00)	Straight_Hair (0.00)	Bald (0.01)	Brown_Hair (0.02)
Straight_Hair	Oval_Face (0.00)	Bags_Under_Eyes (0.00)	Wearing_Necklace (0.00)	Pointy_Nose (0.00)	Smiling (0.00)
Wavy_Hair	Oval_Face (0.01)	Pale_Skin (0.02)	Narrow_Eyes (0.02)	Double_Chin (0.02)	Sideburns (0.03)
Wearing_Earrings	Narrow_Eyes (0.00)	Receding_Hairline (0.02)	Double_Chin (0.02)	Chubby (0.03)	Oval_Face (0.03)
Wearing_Hat	Pale_Skin (0.00)	Bald (0.00)	Bushy_Eyebrows (0.01)	Wearing_Necktie (0.01)	Bags_Under_Eyes (0.01)
Wearing_Lipstick	Straight_Hair (0.04)	Narrow_Eyes (0.05)	Pale_Skin (0.08)	Receding_Hairline (0.09)	Black_Hair (0.11)
Wearing_Necklace	Brown_Hair (0.00)	Straight_Hair (0.00)	Bags_Under_Eyes (0.01)	Big_Nose (0.01)	Double_Chin (0.01)
Wearing_Necktie	Wearing_Hat (0.01)	Sideburns (0.01)	Pale_Skin (0.02)	Blurry (0.02)	Smiling (0.03)
Young	5_o_Clock_Shadow (0.01)	Wearing_Necklace (0.02)	Mouth_Slightly_Open (0.02)	High_Cheekbones (0.02)	Wearing_Hat (0.03)

Table 12: **CelebA most-correlated control sets (highly entangled control concepts).** For each of the 40 target attributes (rows), the 5 control attributes selected as the top-5 by absolute Pearson $|r|$ on the train split, with each $|r|$ value in parentheses. These share substantial information with the target and form the surgicality stress test.

Target	5 most-correlated controls (with $\|r\|$)				
5_o_Clock_Shadow	No_Beard (0.53)	Male (0.42)	Wearing_Lipstick (0.35)	Heavy_Makeup (0.29)	Sideburns (0.27)
Arched_Eyebrows	Wearing_Lipstick (0.46)	Heavy_Makeup (0.44)	Male (0.42)	Attractive (0.28)	Big_Lips (0.28)
Attractive	Wearing_Lipstick (0.51)	Heavy_Makeup (0.50)	Male (0.45)	Young (0.42)	Arched_Eyebrows (0.28)
Bags_Under_Eyes	Big_Nose (0.35)	Male (0.30)	Heavy_Makeup (0.28)	Wearing_Lipstick (0.26)	Young (0.25)
Bald	Gray_Hair (0.28)	Young (0.22)	Chubby (0.22)	Big_Nose (0.22)	Wearing_Necktie (0.19)
Bangs	Male (0.20)	Wearing_Lipstick (0.20)	No_Beard (0.15)	Heavy_Makeup (0.14)	Receding_Hairline (0.13)
Big_Lips	Arched_Eyebrows (0.28)	Male (0.21)	Wearing_Lipstick (0.21)	Wavy_Hair (0.18)	Wearing_Necklace (0.17)
Big_Nose	Male (0.36)	Bags_Under_Eyes (0.35)	Young (0.33)	Double_Chin (0.31)	Chubby (0.29)
Black_Hair	Bushy_Eyebrows (0.29)	Brown_Hair (0.24)	Blond_Hair (0.24)	Young (0.15)	5_o_Clock_Shadow (0.14)
Blond_Hair	Male (0.31)	Wearing_Lipstick (0.31)	Heavy_Makeup (0.28)	Black_Hair (0.24)	Arched_Eyebrows (0.20)
Blurry	Attractive (0.18)	Heavy_Makeup (0.16)	Wearing_Lipstick (0.15)	Young (0.11)	Arched_Eyebrows (0.09)
Brown_Hair	Black_Hair (0.24)	Blond_Hair (0.17)	Male (0.14)	Wearing_Lipstick (0.13)	Wavy_Hair (0.13)
Bushy_Eyebrows	Black_Hair (0.29)	No_Beard (0.25)	Male (0.23)	5_o_Clock_Shadow (0.17)	Wearing_Lipstick (0.16)
Chubby	Double_Chin (0.46)	Young (0.30)	Big_Nose (0.29)	Wearing_Necktie (0.22)	Bald (0.22)
Double_Chin	Chubby (0.46)	Young (0.33)	Big_Nose (0.31)	Gray_Hair (0.25)	Bags_Under_Eyes (0.24)
Eyeglasses	Wearing_Lipstick (0.24)	Chubby (0.21)	Male (0.21)	Heavy_Makeup (0.20)	Attractive (0.20)
Goatee	No_Beard (0.56)	Sideburns (0.51)	Mustache (0.43)	Male (0.30)	Wearing_Lipstick (0.26)
Gray_Hair	Young (0.35)	Bald (0.28)	Double_Chin (0.25)	Wearing_Necktie (0.23)	Receding_Hairline (0.23)
Heavy_Makeup	Wearing_Lipstick (0.82)	Male (0.68)	Attractive (0.50)	Arched_Eyebrows (0.44)	Wearing_Earrings (0.37)
High_Cheekbones	Smiling (0.69)	Mouth_Slightly_Open (0.44)	Wearing_Lipstick (0.29)	Wearing_Earrings (0.29)	Male (0.27)
Male	Wearing_Lipstick (0.80)	Heavy_Makeup (0.68)	No_Beard (0.52)	Attractive (0.45)	5_o_Clock_Shadow (0.42)
Mouth_Slightly_Open	Smiling (0.53)	High_Cheekbones (0.44)	Wearing_Earrings (0.22)	Rosy_Cheeks (0.15)	Arched_Eyebrows (0.13)
Mustache	Goatee (0.43)	No_Beard (0.43)	Sideburns (0.36)	Male (0.24)	Big_Nose (0.20)
Narrow_Eyes	Big_Lips (0.16)	Oval_Face (0.14)	Big_Nose (0.13)	Double_Chin (0.12)	Mouth_Slightly_Open (0.09)
No_Beard	Goatee (0.56)	5_o_Clock_Shadow (0.53)	Sideburns (0.52)	Male (0.52)	Wearing_Lipstick (0.43)
Oval_Face	Heavy_Makeup (0.22)	High_Cheekbones (0.21)	Attractive (0.21)	Smiling (0.20)	Rosy_Cheeks (0.17)
Pale_Skin	Attractive (0.12)	Male (0.10)	Young (0.09)	High_Cheekbones (0.09)	Smiling (0.09)
Pointy_Nose	Wearing_Lipstick (0.25)	Heavy_Makeup (0.25)	Male (0.24)	Arched_Eyebrows (0.23)	Attractive (0.18)
Receding_Hairline	Gray_Hair (0.23)	Big_Nose (0.22)	Young (0.18)	Double_Chin (0.15)	Bags_Under_Eyes (0.13)
Rosy_Cheeks	Heavy_Makeup (0.28)	Arched_Eyebrows (0.26)	Wearing_Lipstick (0.25)	High_Cheekbones (0.25)	Smiling (0.24)
Sideburns	No_Beard (0.52)	Goatee (0.51)	Mustache (0.36)	Male (0.28)	5_o_Clock_Shadow (0.27)
Smiling	High_Cheekbones (0.69)	Mouth_Slightly_Open (0.53)	Rosy_Cheeks (0.24)	Wearing_Earrings (0.23)	Wearing_Lipstick (0.22)
Straight_Hair	Wavy_Hair (0.31)	Black_Hair (0.14)	Wearing_Hat (0.10)	Wearing_Earrings (0.09)	Blurry (0.09)
Wavy_Hair	Wearing_Lipstick (0.35)	Heavy_Makeup (0.33)	Male (0.33)	Straight_Hair (0.31)	Attractive (0.24)
Wearing_Earrings	Wearing_Lipstick (0.37)	Heavy_Makeup (0.37)	Male (0.37)	High_Cheekbones (0.29)	Arched_Eyebrows (0.26)
Wearing_Hat	Wearing_Lipstick (0.17)	Heavy_Makeup (0.17)	Attractive (0.13)	Big_Nose (0.12)	Male (0.12)
Wearing_Lipstick	Heavy_Makeup (0.82)	Male (0.80)	Attractive (0.51)	Arched_Eyebrows (0.46)	No_Beard (0.43)
Wearing_Necklace	Male (0.26)	Wearing_Lipstick (0.24)	Arched_Eyebrows (0.23)	Wearing_Earrings (0.17)	Big_Lips (0.17)
Wearing_Necktie	Male (0.30)	Wearing_Lipstick (0.24)	Young (0.24)	Gray_Hair (0.23)	Big_Nose (0.22)
Young	Attractive (0.42)	Gray_Hair (0.35)	Double_Chin (0.33)	Big_Nose (0.33)	Male (0.32)

Table 13: **Per-row λ_i statistics for the proposed method ($\varepsilon = 0.1$, $\lambda_{\max} = 64$) on the 39-setting NLP grid at 50%-depth.** Each cell averages the per-row λ_i first across all 60 outer iterations of one trajectory, then across the 13 model settings of the concept. The within-setting p_{10}/p_{90} spread (mean across settings) is roughly $14/45$, i.e. a $\sim$10$\times$ ratio between the least- and most-confident rows in a single iteration. The safety cap $\lambda_{\max} = 64$ is observed only sporadically (max 57.32 across 39$\times$60 = 2340 iterations); the closed-form rule rarely binds against it.

Concept	n	mean λ_i	median	min	max
Sycophancy	13	29.58	37.15	4.89	49.38
Gender	13	25.28	23.76	0.45	45.08
Safety	13	33.07	33.30	1.71	57.32
Overall	**39**	**29.31**	**30.19**	**0.45**	**57.32**

Table 14: Geometry contrast between sycophancy and gender across completed layers 4, 8, and 12. This is a focused contrast between the primary nonlinear case and the diagnostic baseline, not an all-dataset geometry census. The table reports two quantities: a local intrinsic-dimension range and the mean angle between a global concept direction and local per-sample concept directions. Under the manifold framing, large local-global angles indicate that a single global direction is a poor first-order summary across neighborhoods. They do not by themselves provide a direct estimate of curvature. Sycophancy shows substantially larger local-global misalignment than gender, supporting it as the primary case for locally adaptive, manifold-motivated erasure methods.

Concept	Local intrinsic dimension	Mean local/global angle
Sycophancy	34–53	76.9–82.6°
Gender	365–461	51.8–55.8°

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction claim that MANCE improves prior erasers under fixed surgicality budgets and that MANCE^{++} reaches state-of-the-art nonlinear erasure. Section 5 supports both claims on 119 NLP and vision settings.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Limitations section discusses the measurement scope of the probe-based metrics, the local manifold estimate, and the computational cost.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper contains no theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 3 and Algorithm 1 define the method. Section 4.2 gives the settings, data splits, and seed protocol, and App. F gives every hyperparameter value.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code is public at `https://github.com/MatanAvitan/mance` under the MIT license. All datasets are public, and Section 4.2 gives the splits and seeds.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Section 4.2 gives the data splits and the seed protocol. App. F gives all hyperparameters, which are fixed across the 119 settings with no per-setting tuning.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We run every method with three seeds. App. Table 4 reports the standard deviation over seeds.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: App. G reports the hardware (one NVIDIA B200 GPU with 8 CPU cores) and the wall-clock time of every method. MANCE takes about eight minutes per setting and seed.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses public datasets and models under their licenses and involves no human subjects.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper does not discuss societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases no data or models with a high risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite each dataset and CLIP, and we name every LLM with its version. The licenses are CelebA (non-commercial research only), Bias in Bios (MIT), PKU-SafeRLHF (CC BY-NC 4.0), the model-written evaluations of Perez et al. [2023] (CC BY 4.0), CLIP (MIT), Qwen2.5 0.5B and 1.5B (Apache 2.0), Qwen2.5 3B (Qwen Research License), Gemma-2 and Gemma-3 (Gemma Terms of Use), Llama-3.2 (Llama 3.2 Community License), and Mistral-7B (Apache 2.0).

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The only new asset is our code, released at `https://github.com/MatanAvitan/mance` under the MIT license with a README. We release no new datasets or models.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper involves no crowdsourcing and no research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper involves no research with human subjects.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are only the source of the representations that we edit. No LLM is part of the method.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.