RANDOMIZED MANIFOLD SMOOTHING FOR FORMAL ROBUSTNESS TO SEMANTIC PERTURBATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Model predictions can be sensitive to small, unrelated variations in the input, which questions the reliability of the model. Randomized smoothing provides robustness guarantees against small, adversarial perturbations, giving a certifiably robust predictor for such perturbations. However, these perturbations are applied across all directions of the ambient space, irrespective of the underlying data manifold. Consequently, its guarantees largely cover perturbations of the input that are not seen in reality rather than semantically meaningful variations in the data. To address this, we introduce **Randomized Manifold Smoothing**, which adapts the smoothing distribution to local data geometry by shaping Gaussian perturbations along principal directions of a locally estimated subspace. This yields anisotropic certificates over locally supported variations while retaining the statistical certification machinery of randomized smoothing. We evaluate manifold smoothing across image classification, semantic segmentation, interpretable concept representations, and vision-language retrieval tasks, including semantic distribution shifts. Across these settings, we show that manifold smoothing consistently preserves task performance while providing stronger robustness guarantees along semantically meaningful perturbations compared to standard randomized smoothing in high-dimensional data and representation spaces.

1 INTRODUCTION

Machine learning models often achieve strong predictive performance, yet accuracy alone does not guarantee that their predictions remain stable at deployment, where small changes in data acquisition, appearance, or representation can alter a model's prediction. For safety-critical applications, it is therefore important to establish not only whether a prediction is correct, but whether it is *provably stable* under natural variations of the input.

Randomized smoothing (RS) provides scalable, provable robustness guarantees (Lecuyer et al., 2019; Cohen et al., 2019) by certifying predictions within an ℓ_2 ball around a given input using isotropic Gaussian perturbations. While well-suited for norm-bounded adversarial robustness, this isotropic certified region around the input is a poor match for natural variations in the data, i.e., the semantically meaningful changes of the kind observed between real samples. Randomized smoothing perturbs inputs in a d-dimensional space, either the raw input space (e.g., pixels) or a learned representation space (e.g., CLIP embeddings), which we refer to as the *ambient space* $\mathbb{R}^d$. Here, data typically lie near a much lower-dimensional manifold (Fefferman et al., 2016; Pope et al., 2021). Natural data variations follow this manifold, which means they are anisotropic in the ambient space: large along a few directions and negligible along most others. Standard randomized smoothing using isotropic noise, however, perturbs all ambient space directions equally, which suits worst-case (adversarial) perturbations rather than natural variations. This creates a gap between *what randomized smoothing certifies* and *the variations we seek robustness against*. We therefore ask:

> *Can we certify robustness to variations supported by the data distribution, rather than to arbitrary perturbations in the ambient space?*

We introduce **Randomized Manifold Smoothing**, which adapts the smoothing distribution to the local data geometry (Fig. 1). Around each input x, we estimate local variation from neighboring data, identify its principal directions and scales, and shape Gaussian perturbations accordingly. This

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

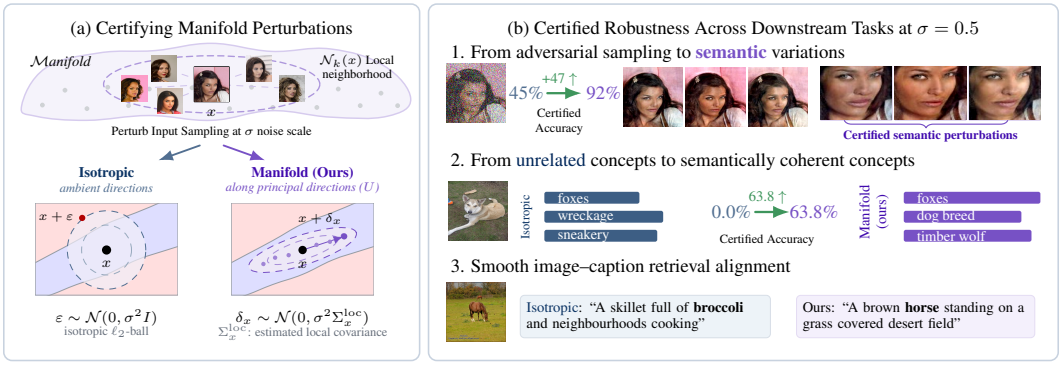


Figure 1: **Randomized Manifold Smoothing enables useful certification over data-supported variation.** (a) Aligning the certified geometry with local variation around x sustains meaningful guarantees beyond the regime effectively covered by isotropic smoothing. (b) Across downstream tasks, this yields higher certified accuracy, preserved semantic concepts, and more stable cross-modal retrieval, demonstrating certification over structured variation across representation spaces.

induces an input-dependent ellipsoidal geometry that stretches along directions of greater local variation and contracts along directions with little local support. Through a whitening transformation, we retain the statistical certification machinery of randomized smoothing while obtaining guarantees over this locally adapted geometry. Our main contributions are:

- Introduction of **Randomized Manifold Smoothing**, a framework for formal robustness certification against semantically meaningful perturbations.
- **Manifold-aligned ellipsoidal certificates.** We estimate the local geometry from neighboring data points to define an anisotropic smoothing distribution, and use whitening to certify an ellipsoidal region aligned with this geometry.
- **Empirical validation.** We evaluate manifold smoothing on image classification, semantic segmentation, sparse concept spaces, and vision–language retrieval, spanning pixel and representation spaces.

Across tasks, manifold smoothing matches isotropic smoothing at small noise and substantially outperforms it at larger noise, where isotropic smoothing collapses. Our results indicate that concentrating perturbations along data-supported directions, rather than spreading them over all ambient directions, makes manifold smoothing an effective approach for *certifying robustness to natural variations in the data*.

2 Related Work

A robustness certificate is a worst-case guarantee of prediction invariance: given a base predictor f and an input x, the certified predictor g returns a radius r such that $g(x') = g(x)$ for every x' within distance r of x, or abstains when no such r can be established. Randomized smoothing (RS) (Lecuyer et al., 2019; Cohen et al., 2019) provides scalable black-box guarantees by perturbing the input at random, aggregating the resulting predictions, and turning how consistently the top class wins into a certified radius. The region this radius carves out around x, and thus the perturbations the certificate covers, has been explored along several lines.

ℓ_p certification. Standard Gaussian RS certifies an isotropic ℓ_2-ball (Cohen et al., 2019), with subsequent work extending smoothing to other ℓ_p regions (Lee et al., 2019; Levine & Feizi, 2020; Yang et al., 2020). These regions are fixed independently of x and can be misaligned with the low-dimensional directions of natural data variations. Our region instead adapts its shape to the local data geometry.

Input-dependent certification. A second line relaxes isotropy by letting the smoothing covariance depend on the input. Alfarra et al. (2022b) optimize a per-input noise variance, and ANCER (Eiras

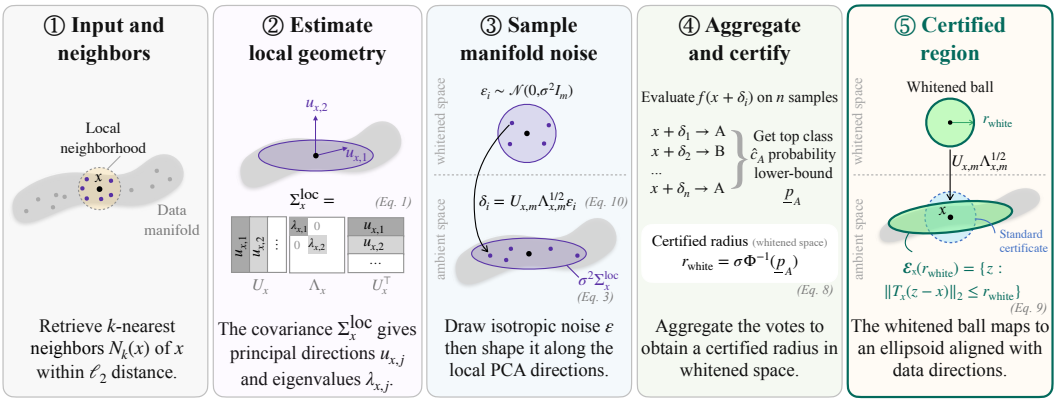


Figure 2: **Manifold-aligned certification through local geometry estimation.** Given an input x, (1) we retrieve its k nearest neighbors $N_k(x)$, (2) estimate local geometry through principal component analysis (PCA) (Eq. 1), and (3) sample isotropic Gaussian noise ε_i in whitened space and shape it along the local principal directions (Eq. 10). (4) We then aggregate predictions over n perturbed Monte Carlo samples $x+\delta_i$ and certify the predicted class $\hat{c}_A$ from the lower confidence bound $\underline{p}_A$, obtaining a certified radius r_{white} in whitened space (Eq. 8). (5) Mapping the standard certified ℓ_2-ball from the whitened space to the ambient space through the local covariance results in a manifold-aligned ellipsoid $\mathcal{E}_x(r_{\text{white}})$ around x (Eq. 9).

et al., 2022) generalizes this to full anisotropy through per-example volume maximization over ellipsoidal regions. Súkeník et al. (2022) show that input-dependent variance can break soundness unless constrained to vary slowly in x. Unlike prior input-dependent methods, which optimize the smoothing covariance against the certificate, we estimate it non-parametrically from local geometry and keep it fixed at x for certification; Theorems A.1 & A.2 establish its convergence to the true tangent-space covariance.

Transformation-specific certification. A third line certifies structured, semantically meaningful changes using predefined transformation families (Fischer et al., 2020; Li et al., 2021; Alfarra et al., 2022a). These guarantees are interpretable, but restricted to transformations fixed in advance. Our approach instead recovers the directions of variation directly from the local data density.

Beyond classification. RS has also been extended to structured outputs, including object detection (Chiang et al., 2020), general metric output spaces (Kumar & Goldstein, 2021), semantic segmentation (Fischer et al., 2021; Anani et al., 2024), attributions (Anani et al., 2025) and circuit discovery (Anani et al., 2026). These extensions concern *what* is certified and are orthogonal to the geometry in which robustness is measured.

Summary. Existing RS certificates either measure r in a fixed norm independently of the data, adapt the region by optimizing against the radius itself, or restrict it to a transformation family specified in advance. Manifold smoothing instead derives the region from the local geometry of the data: an ellipsoid aligned with the estimated tangent space at x, requiring neither a parametric family nor optimization of the smoothing distribution.

3 Randomized Manifold Smoothing

The central idea of our approach is to produce noise along directions reflecting the data geometry, shifting the goal from robustness against *adversarial* perturbations to robustness against *semantic* perturbations. Thus, instead of considering isotropic Gaussian noise, which might be added through a technical artifact or adversary but has no direct semantic meaning, we provide a sense for robustness to semantic change, such as a lesion in an X-Ray being slightly larger or smaller. Here, we determine semantic changes in a data-driven manner, considering the variation along directions in the ambient space around the target sample x. Crucially, this lets us carry over the formalism of standard RS while modeling semantic variations, in the sense of natural image changes observed in the data. We then model noise *proportional to these observed variations*, drawing Gaussian noise around x whose covariance is the local covariance of the k-neighborhood of x, scaled by a global

noise level. Instead of an ℓ_2-ball we now obtain an ellipsoid that stretches far along the principal directions with large eigenvalues, and little along those with small eigenvalues. Directions with larger eigenvalues correspond to greater observed local variation. Before we discuss this approach more formally, we introduce the necessary notation.

3.1 NOTATION

We consider inputs $\mathcal{X} = \mathbb{R}^d$, labels $\mathcal{Y} = \{1, \dots, C\}$, a training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n_\mathcal{D}} \subset \mathcal{X} \times \mathcal{Y}$, and a base predictor $f : \mathcal{X} \to \mathcal{Y}$ to certify. For $x \in \mathcal{X}$, $N_k(x) = \{i_1, \dots, i_k\}$ indexes its k nearest neighbors in ambient ℓ_2 distance, which we assume approximates the manifold distance locally. The local mean, covariance, and eigen-decomposition of the covariance over this neighborhood are

$$\mu_x = \frac{1}{k}\sum_{j=1}^{k} x_{i_j}, \qquad \Sigma_x^{\text{loc}} = \frac{1}{k-1}\sum_{j=1}^{k}(x_{i_j} - \mu_x)(x_{i_j} - \mu_x)^\top = U_x \Lambda_x U_x^\top, \tag{1}$$

where the columns $u_{x,1}, \dots, u_{x,d}$ of the orthogonal matrix U_x are the eigenvectors, i.e., the local principal directions, and $\Lambda_x = \operatorname{diag}(\lambda_{x,1}, \dots, \lambda_{x,d})$ holds the eigenvalues $\lambda_{x,1} \geq \cdots \geq \lambda_{x,d} \geq 0$, i.e., the variances along these directions; together they give a local linear approximation of the data manifold around x.

3.2 NOISE ALONG THE DATA MANIFOLD

We now introduce the noise sampling along principal directions more formally. Let $\sigma > 0$ be a global scaling parameter for the noise level, as common in RS. In the local principal basis, we draw noise independently along each principal direction with variance proportional to its eigenvalue,

$$\eta_j \sim \mathcal{N}(0,\ \sigma^2 \lambda_{x,j}), \quad j = 1, \dots, d, \qquad \text{that is} \qquad \eta \sim \mathcal{N}\left(0,\ \sigma^2 \Lambda_x\right). \tag{2}$$

Rotating η back to the original (pixel) coordinates with U_x gives

$$\delta_x = U_x \eta \sim \mathcal{N}\left(0,\ \Sigma_x\right), \qquad \Sigma_x = U_x(\sigma^2 \Lambda_x) U_x^\top.$$

Pulling out the scalar leaves the eigen-decomposition $\Sigma_x^{\text{loc}} = U_x \Lambda_x U_x^\top$, so the noise covariance is a scaled copy of the local empirical covariance,

$$\delta_x \sim \mathcal{N}\big(0,\ \sigma^2 \Sigma_x^{\text{loc}}\big). \tag{3}$$

Looking at Eq. 2, the per-direction variance $\sigma^2 \lambda_{x,j}$ ties the noise to the local data variations: perturbations are large along high-variance principal directions and vanish along directions where the neighbors barely differ. Since off-manifold directions carry near-zero eigenvalues, the noise stays close to the manifold by construction. Using this noise, we define the manifold-smoothed classifier

$$g(x) = \arg\max_{c \in \mathcal{Y}} \pi_c(x), \qquad \pi_c(x) = \Pr_{\delta_x \sim \mathcal{N}(0,\, \sigma^2 \Sigma_x^{\text{loc}})}\big[f(x + \delta_x) = c\big]. \tag{4}$$

This is the exact analogue of standard RS (Cohen et al., 2019) with the isotropic covariance $\sigma^2 I$ replaced by the *anisotropic* $\sigma^2 \Sigma_x^{\text{loc}}$. The noise is shaped by the local PCA on the kNNs of x. Lifting the formal guarantees to anisotropic noise is discussed next.

3.3 FORMAL GUARANTEES THROUGH WHITENING

To directly transfer the formal guarantees of vanilla RS, we need isotropic noise. We achieve this through coordinate transformation by *whitening* with respect to the local covariance.

PCA-based whitening. To obtain isotropic noise, we map points x' near x to whitened coordinates

$$\widetilde{x}' = T_x(x' - x), \qquad T_x = \Lambda_x^{-1/2} U_x^\top. \tag{5}$$

For numerical stability, in practice we omit near-zero eigenvalues which carry no noise. Now we analyze what happens to noise $\delta_x \sim \mathcal{N}(0,\ \sigma^2 \Sigma_x^{\text{loc}})$ in this space. The whitened noise is

$$\widetilde{\delta}_x = T_x \delta_x = \Lambda_x^{-1/2} U_x^\top \delta_x \ .$$

Since $\delta_x = U_x\eta,\ \eta \sim \mathcal{N}(0,\ \sigma^2\Lambda_x)$, we thus get

$$\widetilde{\delta}_x = \Lambda_x^{-1/2} U_x^\top U_x \eta = \Lambda_x^{-1/2}\eta\ .$$

As $\eta \sim \mathcal{N}(0,\ \sigma^2\Lambda_x)$ we have $\widetilde{\delta}_x \sim \mathcal{N}(0,\ \sigma^2 I)$, meaning that in whitened coordinates the *noise is isotropic Gaussian*, allowing standard RS guarantees in the whitened space.

Smoothed classifier. With the whitened classifier $\widetilde{f}_x(\widetilde{x}') := f\big(x + U_x\Lambda_x^{1/2}\widetilde{x}'\big)$, the smoothed classifier becomes

$$g(x) = \arg\max_{c\in\mathcal{Y}} \Pr_{\widetilde{\delta}_x\sim\mathcal{N}(0,\ \sigma^2 I)} \big[\widetilde{f}_x(\widetilde{\delta}_x) = c\big]\ . \tag{6}$$

Hence, PCA–eigenvalue–scaled smoothing collapses to standard RS, *but in coordinates that are whitened* by the local covariance.

The induced certified region. Following standard RS (Cohen et al., 2019), the guarantee in the whitened coordinates is an ℓ_2-ball around the origin, which translates to an *ellipsoid* around x in the ambient space. Fix a candidate class $c_A \in \mathcal{Y}$, and let $\underline{p}_A$ be a lower bound on $\Pr[\widetilde{f}_x(\widetilde{\delta}_x) = c_A]$ and $\overline{p}_B$ an upper bound on the runner-up class probability. When $\underline{p}_A \geq \overline{p}_B$, the standard RS (with noise variance σ^2) certified radius in the whitened space is

$$r_{\text{white}} = \frac{\sigma}{2}\big(\Phi^{-1}(\underline{p}_A) - \Phi^{-1}(\overline{p}_B)\big), \tag{7}$$

where Φ^{-1} denotes the inverse standard normal CDF. In practice, we follow the one-sided CERTIFY procedure of Cohen et al. (2019): we estimate only the lower confidence bound $\underline{p}_A$ and set $\overline{p}_B = 1-\underline{p}_A$, a valid upper bound on the runner-up since class probabilities sum to one. With $\Phi^{-1}(1-p) = -\Phi^{-1}(p)$, the radius reduces to

$$r_{\text{white}} = \sigma\,\Phi^{-1}(\underline{p}_A), \tag{8}$$

which is positive exactly when $\underline{p}_A > 0.5$; otherwise we abstain. Hence, smoothing f at any $z = x + U_x\Lambda_x^{1/2}\widetilde{\delta}$ with $\|\widetilde{\delta}\|_2 \leq r_{\text{white}}$ using the noise of Eq. 3, i.e., with the covariance fixed at x, still predicts c_A.

The certificate above (Eq. 8) is defined in whitened coordinates. To express it in the ambient space, recall that an ambient perturbation $\delta = z - x$ maps to $\widetilde{\delta} = T_x\delta$ (Eq. 5). The condition $\|\widetilde{\delta}\|_2 \leq r_{\text{white}}$ therefore becomes $\|T_x(z-x)\|_2 \leq r_{\text{white}}$, and expanding the norm with $T_x^\top T_x = U_x\Lambda_x^{-1}U_x^\top$ gives the certified region

$$\mathcal{E}_x(r_{\text{white}}) = \big\{z : \|T_x(z-x)\|_2 \leq r_{\text{white}}\big\} = \big\{z : (z-x)^\top U_x\Lambda_x^{-1}U_x^\top(z-x) \leq r_{\text{white}}^2\big\}. \tag{9}$$

This is an ellipsoid centered at x with principal axes $u_{x,j}$ and semi-axis lengths $r_{\text{white}}\sqrt{\lambda_{x,j}}$. The certificate thus extends far along directions in which the neighbors of x vary strongly and contracts along directions with little local variation, i.e., it is a Mahalanobis ball in the metric of the inverse local covariance. Thus, for all $z \in \mathcal{E}_x(r_{\text{white}})$, smoothing with the covariance fixed at x predicts $c_A = g(x)$: as in standard RS, the smoothing distribution is held fixed while the input moves, rather than being re-estimated at z. App. A.1 further discusses the certified region for the empirical ellipsoid estimated from the neighborhood of x.

3.4 ESTIMATING THE MANIFOLD-SMOOTHED CLASSIFIER

Fig. 2 illustrates the manifold certification procedure. Since the probabilities in Eq. 4 are intractable, we estimate them by Monte Carlo sampling. We evaluate f on noisy copies of x and certify the most frequent class only if it wins by a sufficient margin. Following the CERTIFY procedure of Cohen et al. (2019), given x, we retrieve its k nearest neighbors $N_k(x)$ (Fig. 2 ①), compute the local covariance (Fig. 2 ②) Σ_x^{loc} (Eq. 1), and eigendecompose it, keeping the top-m pair $(U_{x,m}, \Lambda_{x,m})$. Noise is then drawn as (Fig. 2 ③)

$$\delta_i = U_{x,m}\Lambda_{x,m}^{1/2}\,\varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_m), \tag{10}$$

and f is evaluated on each $x + \delta_i$ (Fig. 2 ④). Each ε_i is a draw of the whitened noise $\widetilde{\delta}_x$ of Sec. 3.3, restricted to the retained m directions, so shaping it is equivalent to sampling δ_x directly. A one-sided binomial test certifies the top class $\hat{c}_A$ when its lower confidence bound exceeds a threshold

Table 1: **Certified accuracy across representative tasks** using the certification protocols of Cohen et al. (2019) and Fischer et al. (2021). Subscripts I and M denote isotropic and manifold smoothing, respectively. Δ denotes the absolute difference between manifold and isotropic certified accuracy.

σ	**Classification** (CelebA)			**Classification** (CelebA-HQ)			**Segmentation** (CelebAMask-HQ)			**Concept Space** (ImageNet)		
	$\mathrm{CAcc_I}\uparrow$	$\mathrm{CAcc_M}\uparrow$	Δ	$\mathrm{CAcc_I}\uparrow$	$\mathrm{CAcc_M}\uparrow$	Δ	$\mathrm{CAcc_I}\uparrow$	$\mathrm{CAcc_M}\uparrow$	Δ	$\mathrm{CAcc_I}\uparrow$	$\mathrm{CAcc_M}\uparrow$	Δ
0.02	**93.0**	92.0	1.0↓	**94.0**	93.0	1.0↓	94.30	**95.09**	0.79 ↑	67.08	**68.08**	1.00 ↑
0.05	87.0	**92.0**	5.0 ↑	91.0	**92.0**	1.0 ↑	92.82	**94.74**	1.92 ↑	64.00	**68.06**	4.06 ↑
0.10	72.0	**92.0**	20.0 ↑	83.0	**92.0**	9.0 ↑	90.46	**93.99**	3.53 ↑	45.8	**68.6**	22.8 ↑
0.50	45.0	**92.0**	47.0 ↑	50.0	**93.0**	43.0 ↑	26.87	**84.27**	57.40 ↑	0.0	**63.8**	63.8 ↑
1.00	45.0	**86.0**	41.0 ↑	50.0	**86.0**	36.0 ↑	26.99	**67.77**	40.78 ↑	0.0	**47.2**	47.2 ↑

$\tau \in [0.5, 1)$ with confidence $1-\alpha$, and abstains otherwise ; with $c_A = \hat{c}_A$, the certified region is then the ellipsoid (Fig. 2 ⑤) in Eq. 9 with radius r_{white} from Eq. 8 for classification ($\tau = 0.5$), and the fixed radius $\sigma\,\Phi^{-1}(\tau)$ for segmentation (Fischer et al., 2021), where $\underline{p}_A$ is the one-sided lower confidence bound on $\pi_{\hat{c}_A}(x)$. Since only the sampling distribution differs from standard RS (Eq. 6), the statistical test is unchanged, and the guarantee transfers directly. Pseudo-code is given in App. B.2.

4 Experimental Setup

Certification. Across all experiments, we compare manifold smoothing against standard isotropic randomized smoothing, under identical certification settings. For classification, we follow the CERTIFY procedure of Cohen et al. (2019), while for semantic segmentation we adapt SEGCERTIFY (Fischer et al., 2021) for pixel-wise certification. For certification, we use the standard $\underline{p}_A > 0.5$ criterion for classification and $\tau = 0.75$ for semantic segmentation. We use $n_0 = 30$ Monte Carlo samples for class selection, $n = 100$ for certification, and failure probability $\alpha = 0.001$. For manifold smoothing, we retrieve $k = 500$ nearest neighbors via approximate nearest-neighbor search and estimate the local PCA basis U and covariance spectrum Λ. For the main experiments, we retain the full PCA basis and use the resulting covariance to shape Gaussian noise, and perturb the input as $x' = x + \delta_i$. We use the same global noise scale σ for isotropic and manifold smoothing; for manifold smoothing, σ scales the locally estimated covariance Σ_x^{loc}. Main certification results use a randomly selected subset of 100 test examples. We ablate neighborhood size k, retained dimension m, Monte Carlo sample size n, and test subset size N in App. D.

Models and datasets. For image classification, we evaluate ResNet-18 smile prediction on CelebA and CelebA-HQ, including distribution shifts on CelebA. For segmentation, we use pretrained BiSeNet (Yu et al., 2018) on CelebAMask-HQ. We apply smoothing only at certification time, without model retraining, to isolate perturbation geometry. For concept-space classification, we evaluate a pretrained ImageNet classifier on a smooth sparse SAE concept representation (CFM (Wittenmayer et al., 2026)). For image-to-text retrieval, we smooth CLIP ViT-B/32 image embeddings and evaluate on RoCOCO (Park et al., 2024). Implementation details are provided in App. Sec. B.

Evaluation metrics. We evaluate robustness with task-specific metrics. For classification, we report certified accuracy ($\mathrm{CAcc}\uparrow$) that is, the fraction of examples that are both correctly classified and certified, and abstention rate ($\varnothing\downarrow$). For semantic segmentation, pixel-wise CAcc, certified mIoU ($\mathrm{mIoU}\uparrow$), and pixel abstention rate ($\varnothing\downarrow$), the fraction of pixels for which no certified prediction is returned. For image-to-text retrieval, we report Recall@1 ($\mathrm{R@1}\uparrow$), Retrieval Safety Misalignment Score ($\mathrm{RSMS}\downarrow$), and retrieval drop ($\mathrm{Drop}\downarrow$). Subscripts I and M denote isotropic and manifold smoothing. Evaluation metrics details are discussed in App. Sec. C.

5 Empirical results

We evaluate manifold smoothing across four tasks, each probing an aspect of **robustness certification** under meaningful **semantic perturbations** across different **representation spaces**. Hence, we investigate: (i) *Image Classification*: how far can the perturbation magnitude increase while preserving certified predictive structure, including under semantic distribution shift? (Sec. 5.1); (ii) *Semantic Segmentation*: do these guarantees extend from image-level predictions to dense, pixel-level certification? (Sec. 5.2); (iii) *Interpretable Concept Representations*: does smoothing along the

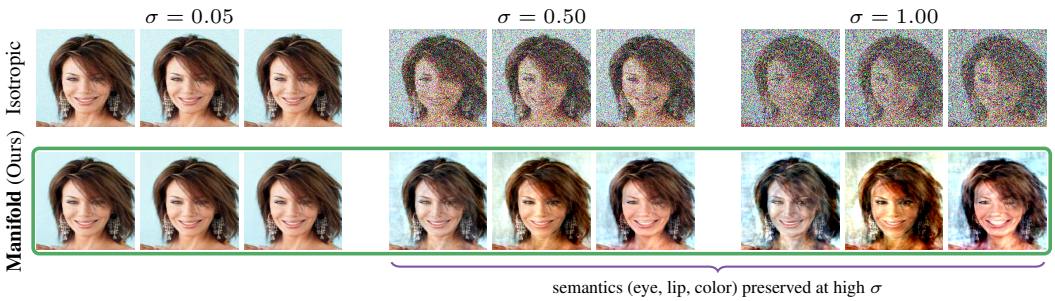


Figure 4: **Smoothing on CelebA at increasing noise levels.** Each column group shows three samples under isotropic and manifold smoothing at the same noise scale σ. Isotropic perturbations increasingly distort facial structure as σ grows, whereas manifold perturbations remain semantically meaningful.

local manifold preserve semantic structure in sparse concept spaces? (Sec. 5.3); and (iv) *Image-to-Text Retrieval*: does preserving local geometry maintain semantic neighborhoods and cross-modal alignment in multimodal embedding spaces? (Sec. 5.4).

5.1 OOD-Robust Image Classification

Robustness guarantees. *Can manifold smoothing provide guarantees at larger, semantically meaningful perturbation levels?* At small noise ($\sigma = 0.02$), Tab. 1 shows that, both smoothers achieve comparable certified accuracy on CelebA and Celeb-HQ. At $\sigma = 1.0$, manifold smoothing retains 86% ($\uparrow 91.1\%$) certified accuracy versus 45% in isotropic collapsing to "No-Smile" prediction; we analyzed this in App. Sec. E.1, showing that larger perturbations can be certified along the data manifold without unnecessarily extending the certified region into off-manifold directions. We provide extended noise scale perturbation results in App. Sec. E.2

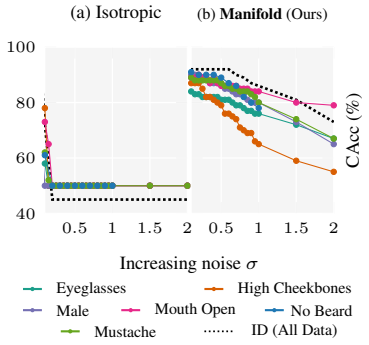


Figure 3: Certified robustness under attribute-OOD CelebA shifts. The dotted curve denotes the in-distribution certified accuracy.

Perturbation structure. *What do these larger perturbations correspond to in image space?* Fig. 4 reveals how isotropic noise produces diffuse corruption while manifold perturbations instead induce coherent variations in *altered smiles, open/closed mouths, eye shape and color, skin tone, hair, and lip shape*, while preserving the meaningful facial structure at even higher perturbations. This highlights the contrast between the two approaches: randomized smoothing was designed to provide robustness guarantees in settings typical of adversarial perturbations, whereas here we seek guarantees when semantically meaningful perturbations occur.

Robustness under semantic distribution shift. *Do guarantees over semantic variations persist under semantic distribution shifts?* We evaluate six attribute-defined CelebA shifts by training with a target attribute absent ($a = 0$) and evaluating where it is present ($a = 1$). Fig. 3 shows that manifold smoothing preserves non-trivial certification across all OOD shifts over larger perturbations. Relative to the in-distribution reference, the OOD curves exhibit varying degradation, most prominently for *High Cheekbones*, showing that the underlying distribution shift remains challenging. However, manifold smoothing robustness guarantees persist under OOD shift.

> Manifold smoothing provides better **robustness guarantees**, preserves **semantic coherence** across *larger perturbations*, and remains **robustly certifiable** under *semantic distribution shift*.

5.2 Image Segmentation

Dense Certification and Semantic Preservation. *Does the advantage of manifold smoothing extend from image-level classification to dense prediction?* Tab. 1 shows that it does. At $\sigma = 0.5$, manifold smoothing provides 84.27% certified pixel accuracy versus 26.87% isotropically (Tab. 1),

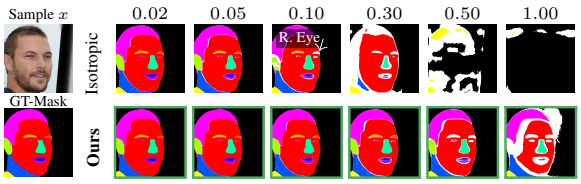


Figure 5: **Certified segmentation under increasing perturbation levels.** CelebAMask-HQ sample under isotropic and manifold smoothing; white pixels denote abstention.

Table 2: **Dense certification on CelebAMask-HQ.** Certified mIoU (%) and pixel abstention rate (∅) under isotropic (I) and manifold (M) smoothing.

σ	$\text{mIoU}_\text{I} \uparrow$	$\text{mIoU}_\text{M} \uparrow$	$\varnothing_\text{I} \downarrow$	$\varnothing_\text{M} \downarrow$
0.02	74.03	**74.60**	1.88	**0.48**
0.05	73.10	**74.54**	3.60	**1.19**
0.10	60.76	**73.37**	3.60	**2.45**
0.30	29.87	**67.62**	23.61	**7.52**
0.50	5.39	**57.44**	40.07	**13.81**
1.00	2.12	**35.58**	**8.81**	30.60

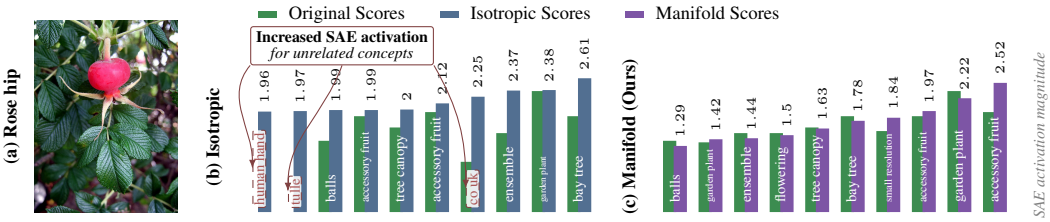


Figure 6: **SAE concept activations under smoothing.** Top-10 concepts for a Rose hip example; bars report SAE activation magnitudes before smoothing (green) and after isotropic (blue) or manifold (purple) smoothing at $\sigma = 1.0$. Isotropic smoothing activates unrelated or weakly activated concepts(mutedred), whereas manifold smoothing largely retains the original activation profile.

and certified mIoU remains 57.44% versus 5.39%, with substantially lower pixel abstention (13.81% vs. 40.07%; Tab. 2). At $\sigma = 1.0$, isotropic abstention decreases despite near-zero mIoU, revealing prediction collapse rather than meaningful robustness. Qualitatively, Fig. 5 reveals *which **semantic** structures remain preserved*: isotropic smoothing first loses *fine features* (eyes, brows, lips) and, at higher noise, *broader facial regions*, while manifold smoothing preserves both. Class-wise ΔIoU gains across the full perturbation range are provided in App. Sec. E.3

> Manifold smoothing provides **dense robustness guarantees** over **semantic variation** at *larger perturbation scales*, from *fine to broader (facial) features*.

5.3 Interpretable Latent Spaces Smoothing

Unlike the preceding pixel-space experiments, here we smooth the sparse SAE concept representation of the input image x while retaining the ImageNet classification task, to study: *Can robustness guarantees be maintained under perturbations in **interpretable sparse concept spaces**?* Tab. 1 shows that sparse concept representations are much more sensitive to isotropic noise, limiting certification at larger perturbations. Isotropic CAcc reaches 0% by $\sigma = 0.5$, whereas manifold smoothing respects local embedding geometry, retaining 63.8% CAcc at $\sigma = 0.5$ and 47.2% at $\sigma = 1.0$. This extends robustness guarantees over larger manifold-aligned variations in concept space.

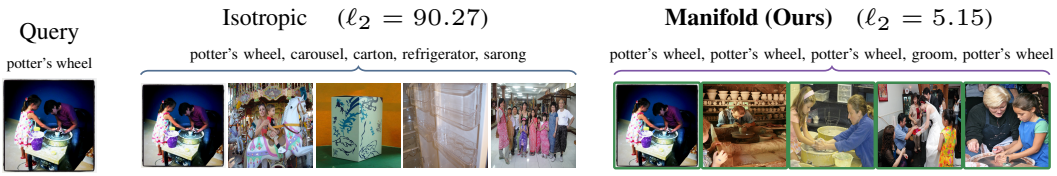


Figure 7: **Nearest-neighbor retrieval after smoothing in SAE space.** At $\sigma = 1.0$, manifold smoothing stays closer to the original representation ($\ell_2 = 5.15$ vs. 90.27) and retrieves more semantically matching top classes (potter's wheel retrieved as 4/5) than isotropic smoothing.

Semantic concept consistency. *Does prediction-level robustness reflect stability of the underlying semantic representation?* Fig. 6 shows that, despite identical top-1 predictions, isotropic perturbations activate weak or absent concepts, while manifold perturbations remain closer to the original activation profile. Consistently, z_M remains much closer to the original embedding than z_I (ℓ_2: 5.15

vs. 90.27) and better preserves its nearest-neighbor structure (Fig. 7) indicating semantic consistency under manifold smoothing. Results for initially inactive concepts are provided in App. E.5.

> Manifold-aligned **guarantees** extend **semantic coherence** in interpretable sparse spaces.

5.4 Safety Alignment in Foundation Models

Vision-language models rely on semantic alignment in a shared embedding space. We ask *whether manifold smoothing can preserve this **alignment** under semantic perturbations*. We evaluate CLIP ViT-B/32 on RoCOCO (Park et al., 2024), which augments COCO with semantically altered captions. We report the *Danger* split, where security-related concepts replace low-influence nouns; all corruption bank sweeps are provided in App. Sec. E.4. At $\sigma = 0.1$, manifold smoothing limits the retrieval drop to 17.20%, compared with 93.90% under isotropic smoothing, indicating substantially better preservation of retrieval-relevant semantic information in embedding space. Notably, smoothing can itself increase RSMS relative to unsmoothed CLIP, but this increase is consistently smaller under manifold smoothing. We provide a qualitative example to visualize semantic alignment in the App. Fig. E.5.

Table 3: RoCOCO Danger retrieval under smoothing. Higher R@1 and lower RSMS and Drop are preferred.

	Isotropic			**Manifold** (Ours)		
σ	R@1 ↑	RSMS ↓	Drop ↓	R@1 ↑	RSMS ↓	Drop ↓
0.02	35.23	26.46	29.70	**41.26**	**22.98**	**17.70**
0.05	8.54	32.20	66.00	**41.34**	**22.90**	**17.60**
0.10	3.08	32.96	93.90	**41.52**	**22.80**	**17.20**
0.50	0.02	27.48	100.00	**35.14**	**25.26**	**29.90**
1.00	0.02	25.00	100.00	**21.32**	**28.14**	**57.50**

> Local covariance smoothing retains **cross-modal retrieval robustness.**

6 Discussion and Conclusion

Discussion. Manifold smoothing outperforms isotropic smoothing across four tasks and three representation spaces (pixels, sparse concepts, multimodal embeddings). It retains certified accuracy where isotropic smoothing collapses, including under semantic distribution shifts. It further extends to dense segmentation with lower abstention, preserves concept activations in sparse spaces, and maintains cross-modal retrieval alignment. Together, these results support our hypothesis that isotropic certificates spend their budget on off-manifold directions in the ambient space rather than meaningful variations in the data. When robustness to such data variations, rather than random perturbations in the ambient space, is the goal, restricting perturbations to locally supported directions yields much stronger guarantees than what standard Randomized smoothing can achieve.

Limitations. Manifold smoothing adds per-input computation cost from neighborhood retrieval and local eigen-decomposition, both computed once and reused across all Monte Carlo samples, reducing cost. Approximate neighbor search reduces the retrieval cost. In App. Sec. D.1 we discuss a single-input runtime cost. The method further assumes a canonical input space in which ambient ℓ_2 distance is locally meaningful. This assumption is more natural for some learned representation spaces and normalized data domains such as CelebA than for raw natural images. For particular inputs, such as natural images, further research has to be done to ensure that local perturbations are equivariant to transformations such as translations of objects in the fixed pixel grid. Similar to standard Randomized Smoothing, certification also becomes harder as the number of classes grows, since exceeding τ is less attainable when probability mass spreads over many classes.

Summary. We introduced Randomized Manifold Smoothing, a framework that provides formal robustness guarantees against perturbations along meaningful data variations. By shaping the smoothing noise with the local covariance of an input's nearest neighbors, our method certifies an ellipsoid aligned with the data manifold rather than an isotropic ball stretching into the ambient space without any relation to the data. Empirically, we showed that manifold smoothing preserves task performance and certified accuracy at perturbation scales where isotropic smoothing collapses. Hence, we show that certification is not limited to adversarial threat models, but practitioners can obtain guarantees over meaningful input variations, bridging formal robustness and real-world reliability.

7 USE OF LLMS

The research methodology, experimental design, analysis, and scientific conclusions presented in this work were developed by the authors. We used large language models to generate text snippets for selected sections, which were subsequently reviewed and developed into the final text by the authors. LLMs were also used for language editing to improve clarity and conciseness and for LaTeX formatting of figures and tables. During implementation, LLMs assisted with project configuration and writing modular code pipelines. All generated code was reviewed and adapted before use.

8 REPRODUCIBILITY STATEMENT

All datasets used in this work are publicly available. We provide references to the pretrained models used for the classification and CFM experiments. Detailed experimental settings, including model, smoothing, sampling, and certification parameters, are provided in the appendix, together with pseudocode for our method. We will release the complete implementation code upon acceptance to facilitate the reproduction of our results.

REFERENCES

Motasem Alfarra, Adel Bibi, Naeemullah Khan, Philip H. S. Torr, and Bernard Ghanem. DeformRS: Certifying input deformations with randomized smoothing. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2022a.

Motasem Alfarra, Adel Bibi, Philip H. S. Torr, and Bernard Ghanem. Data dependent randomized smoothing. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2022b.

Alaa Anani, Tobias Lorenz, Bernt Schiele, and Mario Fritz. Adaptive hierarchical certification for segmentation using randomized smoothing. In *International Conference on Machine Learning (ICML)*, 2024.

Alaa Anani, Tobias Lorenz, Mario Fritz, and Bernt Schiele. Pixel-level certified explanations via randomized smoothing. In *International Conference on Machine Learning (ICML)*, 2025.

Alaa Anani, Tobias Lorenz, Bernt Schiele, Mario Fritz, and Jonas Fischer. Certified circuits: Stability guarantees for mechanistic circuits. In *International Conference on Machine Learning (ICML)*, 2026.

Erik Bernhardsson. Annoy: Approximate nearest neighbors in C++/Python. `https://github.com/spotify/annoy`, 2018. GitHub repository.

Ping-yeh Chiang, Michael J. Curry, Ahmed Abdelkader, Aounon Kumar, John Dickerson, and Tom Goldstein. Detection as regression: Certified object detection with median smoothing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning (ICML)*, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

Francisco Eiras, Motasem Alfarra, Philip Torr, M. Pawan Kumar, Puneet K. Dokania, Bernard Ghanem, and Adel Bibi. ANCER: Anisotropic certification via sample-wise volume maximization. *Transactions on Machine Learning Research (TMLR)*, 2022.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 2016.

Marc Fischer, Maximilian Baader, and Martin Vechev. Certified defense to image transformations via randomized smoothing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Marc Fischer, Maximilian Baader, and Martin Vechev. Scalable certified segmentation via randomized smoothing. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018.

Aounon Kumar and Tom Goldstein. Center smoothing: Certified robustness for networks with structured outputs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *IEEE Symposium on Security and Privacy (S&P)*, 2019.

Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. MaskGAN: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Guang-He Lee, Yang Yuan, Shiyu Chang, and Tommi Jaakkola. Tight certificates of adversarial robustness for randomly smoothed classifiers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Alexander Levine and Soheil Feizi. Robustness certificates for sparse adversarial attacks by randomized ablation. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 4585–4593, 2020.

Linyi Li, Maurice Weber, Xiaojun Xu, Luka Rimanic, Bhavya Kailkhura, Tao Xie, Ce Zhang, and Bo Li. TSS: Transformation-specific smoothing for robustness certification. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 535–557, 2021.

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

Seulki Park, Daeho Um, Hajung Yoon, Sanghyuk Chun, Sangdoo Yun, and Jin Young Choi. RoCOCO: Robustness benchmark of ms-coco to stress-test image-text matching models. *arXiv preprint arXiv:2304.10727*, 2024.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Peter Súkeník, Aleksei Kuvshinov, and Stephan Günnemann. Intriguing properties of input-dependent randomized smoothing. In *International Conference on Machine Learning (ICML)*, pp. 20697–20743, 2022.

Kai Wittenmayer, Sukrut Rao, Amin Parchami-Araghi, Bernt Schiele, and Jonas Fischer. CFM: Language-aligned concept foundation model for vision. *arXiv preprint arXiv:2601.13798*, 2026.

Greg Yang, Tony Duan, J. Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In *International Conference on Machine Learning (ICML)*, pp. 10693–10705, 2020.

Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, 2018.

Overview of the Appendix

In this appendix, we provide additional theoretical analysis, experimental details, ablations, and extended results. Theoretical analysis and proofs are provided in Sec. A, experimental details in Sec. B, evaluation metrics in Sec. C, and ablation studies in Sec. D. Extended quantitative and qualitative results are provided in Secs. E–E.5.

A Discussion on Theory

In this section we discuss formally what the implications of the local covariance estimator are on how well we model the underlying data geometry. We first give a more abstract outline of the theorems in the next section and then provide the detailed proofs in the following sections.

A.1 Outline for the Convergence of the Empirical Manifold Ellipsoid

Our certificate relies on the empirical local covariance matrix Σ_x^{loc} Eq. 1 computed from the k nearest neighbors $N_k(x) = \{i_1, \dots, i_k\}$ of x. The resulting certified region is the ellipsoid

$$E_x(r) = \left\{ z \in \mathbb{R}^d : (z-x)^\top U_x \Lambda_x^{-1} U_x^\top (z-x) \le r^2 \right\},$$

where $\Sigma_x^{\mathrm{loc}} = U_x \Lambda_x U_x^\top$ and r the certified radius. In this section, we state a convergence result showing that this empirical ellipsoid approaches an ideal local manifold ellipsoid as the number of local samples increases, provided the neighborhood remains sufficiently local.

To distinguish the empirical quantity from its population analogue, we write $\widehat{\Sigma}_{x,k}^{\mathrm{loc}} := \Sigma_x^{\mathrm{loc}}$. Let $\mathcal{M} \subset \mathbb{R}^d$ denote the data manifold and let $T_x\mathcal{M}$ be its tangent space at x. We define the k-NN radius

$$\rho_k(x) := \max_{j \le k} \|x_{i_j} - x\|_2.$$

We can define the ideal local tangent covariance by projecting all points in the neighborhood of x on the tangent plane at x as

$$\Sigma_{x,\rho_k}^\star := \mathrm{Cov}\left(P_{T_x\mathcal{M}}(X - x) \,|\, X \in \mathcal{M} \cap B(x, \rho_k(x)) \right),$$

viewed as a $d \times d$ matrix by extending it with zeros in directions normal to $T_x\mathcal{M}$. Here, $P_{T_x\mathcal{M}}$ is the projection onto the tangent plane of the manifold at x and $B(x, \rho_k(x))$ is the euclidean ball with radius $\rho_k(x)$ centered at x.

Now let

$$\Sigma_{x,\rho_k}^\star = U_x^\star \Lambda_x^\star U_x^{\star\top}$$

be its eigendecomposition, where the nonzero eigenvalues are

$$\lambda_{x,1}^\star \ge \cdots \ge \lambda_{x,m}^\star > 0.$$

The *tangent–normal eigengap* is

$$\gamma_x := \lambda_{x,m}^\star - \lambda_{x,m+1}^\star = \lambda_{x,m}^\star,$$

since $\Sigma_{x,\rho_k}^\star$ has rank m.

We first show the convergence of covariance and local PCA. For that, we assume that the data lie locally on a smooth m-dimensional manifold $\mathcal{M} \subset \mathbb{R}^d$ whose reach is at least τ, and that the k nearest neighbors of x lie in a sufficiently small ball $B(x, \rho_k(x))$ where the manifold is well approximated by its tangent space. We also assume that the local samples are well-behaved, e.g. sub-Gaussian, in tangent coordinates.

Under these assumptions, the empirical local covariance $\widehat{\Sigma}_{x,k}^{\mathrm{loc}}$ computed from the k nearest neighbors converges to the ideal tangent-space covariance $\Sigma_{x,\rho_k}^\star$ at a rate controlled by a finite-sample term and a curvature-bias term. Consequently, the empirical eigenvalues and top-m PCA subspace converge to their ideal tangent-space counterparts, with the subspace error additionally controlled by the tangent–normal eigengap γ_x.

Theorem A.1 (Covariance and local PCA convergence)**.** *Assume that the data lie on an m-dimensional C^2 manifold $\mathcal{M} \subset \mathbb{R}^d$ with reach at least $\tau > 0$. Assume further that the local samples in $N_k(x)$ are sub-Gaussian in tangent coordinates and that $\rho_k(x) \leq c\tau$ for a sufficiently small universal constant $c > 0$.*

Then, with probability at least $1 - \delta$,

$$\left\| \widehat{\Sigma}^{\mathrm{loc}}_{x,k} - \Sigma^{\star}_{x,\rho_k} \right\|_2 \leq \eta_k(x, \delta),$$

where

$$\eta_k(x, \delta) = C_1 \rho_k(x)^2 \left(\sqrt{\frac{m + \log(1/\delta)}{k}} + \frac{m + \log(1/\delta)}{k} \right) + C_2 \frac{\rho_k(x)^3}{\tau}.$$

The first term is the finite-sample covariance estimation error, while the second term is the local curvature bias incurred by replacing the manifold by its tangent space. If $\eta_k(x, \delta) \leq \gamma_x/4$, then the empirical PCA subspace $\widehat{U}_{x,m}$ spanned by the top m eigenvectors of $\widehat{\Sigma}^{\mathrm{loc}}_{x,k}$ satisfies

$$\left\| \sin \Theta(\widehat{U}_{x,m}, U^{\star}_x) \right\|_2 \leq \frac{2\eta_k(x, \delta)}{\gamma_x}.$$

Here, $\Theta(\widehat{U}_{x,m}, U^{\star}_x)$ denotes the matrix of principal angles between the empirical PCA subspace $\mathrm{span}(\widehat{U}_{x,m})$ and the ideal tangent subspace $\mathrm{span}(U^{\star}_x) = T_x\mathcal{M}$. Thus, $\| \sin \Theta(\widehat{U}_{x,m}, U^{\star}_x) \|_2$ is the sine of the largest principal angle between these two subspaces and quantifies the error in the estimated local manifold directions. Moreover, for each $j = 1, \ldots, m$,

$$|\widehat{\lambda}_{x,j} - \lambda^{\star}_{x,j}| \leq \eta_k(x, \delta).$$

In summary, Theorem A.1 shows that the empirical ellipsoid is accurate when three quantities are favorable: the number of neighbors k is large, the local neighborhood radius $\rho_k(x)$ is small relative to the manifold reach τ, and the tangent–normal eigengap γ_x is not too small. In particular, the eigengap controls the stability of the local PCA directions.

The preceding result controls the covariance and tangent subspace. We next state the corresponding consequence for the Mahalanobis metric used by the certificate. For that we assume the conditions of Theorem A.1 and additionally require that the covariance estimation error is small compared to the smallest nonzero ideal eigenvalue $\lambda^{\star}_{x,m}$. This ensures that the inverse covariance, and hence the Mahalanobis metric that defines the certificate, is stable.

Let the empirical rank-m Mahalanobis metric be

$$\widehat{A}_x := \widehat{U}_{x,m} \widehat{\Lambda}^{-1}_{x,m} \widehat{U}^{\top}_{x,m},$$

and the ideal tangent Mahalanobis metric

$$A^{\star}_x := U^{\star}_x (\Lambda^{\star}_x)^{-1} U^{\star\top}_x.$$

Under assumptions mentioned above, the empirical Mahalanobis metric $\widehat{A}_x$ converges to the ideal tangent Mahalanobis metric $A^{\star}_x$. As a result, the empirical certified ellipsoid $\widehat{E}_x(r)$ is sandwiched between slightly smaller and slightly larger ideal ellipsoids, showing that the certificate induced by the empirical k-NN covariance converges to the ideal local manifold certificate.

Theorem A.2 (Convergence of the certified ellipsoid)**.** *Let the assumptions of Theorem A.1 hold, and assume $\eta_k(x, \delta) \leq \lambda^{\star}_{x,m}/4$. Then, with probability at least $1 - \delta$,*

$$\left\| \widehat{A}_x - A^{\star}_x \right\|_2 \leq C_3 \eta_k(x, \delta) \left(\frac{1}{(\lambda^{\star}_{x,m})^2} + \frac{1}{\lambda^{\star}_{x,m} \gamma_x} \right).$$

Consequently, for any ambient radius $R > 0$, define

$$\zeta_R := R^2 C_3 \eta_k(x, \delta) \left(\frac{1}{(\lambda^{\star}_{x,m})^2} + \frac{1}{\lambda^{\star}_{x,m} \gamma_x} \right).$$

Let estimated and ideal certified ellipsoid be

$$\widehat{E}_x(r) := \left\{ z : (z-x)^\top \widehat{A}_x (z-x) \le r^2 \right\}$$

and

$$E_x^\star(r) := \left\{ z : (z-x)^\top A_x^\star (z-x) \le r^2 \right\},$$

respectively. Then, if $r^2 > \zeta_R$, we have

$$E_x^\star\left(\sqrt{r^2-\zeta_R}\right) \cap B(x,R) \subseteq \widehat{E}_x(r) \cap B(x,R) \subseteq E_x^\star\left(\sqrt{r^2+\zeta_R}\right) \cap B(x,R).$$

Thus, the empirical certificate converges to the ideal local manifold certificate, with error controlled by the number of neighbors k, the tangent–normal eigengap γ_x, and the curvature scale τ.

In the usual local manifold regime,

$$\lambda_{x,j}^\star \asymp \rho_k(x)^2, \qquad j = 1, \dots, m,$$

and hence also $\gamma_x \asymp \rho_k(x)^2$. Here $a_k \asymp b_k$ means that a_k and b_k are equal up to constant factors, i.e., there exist constants $c_1, c_2 > 0$ such that $c_1 b_k \le a_k \le c_2 b_k$. Thus, $\lambda_{x,j}^\star \asymp \rho_k(x)^2$ means that the local tangent variance along direction j scales like the squared radius of the k-nearest neighborhood.

If this is satisfied, the ellipsoid error scales heuristically as

$$O\left(\sqrt{\frac{m + \log(1/\delta)}{k}} + \frac{\rho_k(x)}{\tau} \right).$$

For a training set of size $n_{\mathcal{D}}$ sampled from a regular density on an m-dimensional manifold, the k-NN radius typically satisfies $\rho_k(x) \asymp \left(\frac{k}{n_{\mathcal{D}}}\right)^{1/m}$. Hence the relative error is of the form

$$O\left(\sqrt{\frac{m + \log(1/\delta)}{k}} + \frac{1}{\tau}\left(\frac{k}{n_{\mathcal{D}}}\right)^{1/m} \right).$$

This displays the *bias–variance tradeoff* in choosing k: smaller neighborhoods produce noisier covariance estimates, while larger neighborhoods reduce sampling error but incur more curvature bias. We discuss the full proof next.

A.2 Detailed Proofs for Local Ellipsoid Convergence

This appendix provides the proof of Theorems A.1 and A.2. The proof separates the approximation error into three components, which are the finite-sample covariance error, the curvature bias, and the inverse-metric perturbation. We will first discuss the relevant notation.

A.2.1 Geometric and probabilistic setup

Let $\mathcal{M} \subset \mathbb{R}^d$ be an m-dimensional C^2 manifold with reach at least $\tau > 0$. Let $x \in \mathcal{M}$, and let $T_x\mathcal{M}$ denote the tangent space at x. Let

$$P_T := P_{T_x\mathcal{M}} \quad \text{and} \quad P_N := I - P_T$$

be the tangent and normal projections.

For the k nearest neighbors of x, $N_k(x) = \{i_1, \dots, i_k\}$, define the maximal distance to a neighbor as

$$\rho_k(x) := \max_{j \le k} \|x_{i_j} - x\|_2.$$

For notational simplicity, we write $\rho := \rho_k(x)$. We work locally on the assumption that this distance is sufficiently small with respect to the reach of the manifold, i.e. $\rho_k(x) \le c\tau$, for $c > 0$ sufficiently small. This ensures that $\mathcal{M} \cap B(x, \rho_k(x))$ is well approximated by the tangent space. As before, let the empirical local covariance be

$$\widehat{\Sigma}_{x,k}^{\mathrm{loc}} = \frac{1}{k-1} \sum_{j=1}^{k} (x_{i_j} - \widehat{\mu}_x)(x_{i_j} - \widehat{\mu}_x)^\top,$$

with empirical mean $\widehat{\mu}_x = \frac{1}{k}\sum_{j=1}^k x_{i_j}$. Let the ideal tangent covariance be

$$\Sigma_{x,\rho_k}^\star = \mathrm{Cov}\left(P_T(X-x) \,|\, X \in \mathcal{M} \cap B(x, \rho_k(x)) \right).$$

A.2.2 LOCAL CHART AND CURVATURE CONTROL

Since $\mathcal{M}$ has reach at least τ, for every $y \in \mathcal{M} \cap B(x, \rho)$ with $\rho \leq c\tau$, we may write

$$y = x + t + q(t),$$

where

$$t \in T_x\mathcal{M}, \qquad q(t) \in T_x\mathcal{M}^{\perp}.$$

The vector t describes tangent displacement, while $q(t)$ describes the normal deviation due to curvature. Standard reach bounds imply that, for a universal constant $C > 0$,

$$\|q(t)\|_2 \leq C\frac{\|t\|_2^2}{\tau}.$$

Since $\|y - x\|_2 \leq \rho$, we also have $\|t\|_2 \leq C\rho$, and therefore

$$\|q(t)\|_2 \leq C\frac{\rho^2}{\tau}.$$

Thus local normal displacement is second order in the neighborhood radius.

A.2.3 CURVATURE BIAS OF THE POPULATION COVARIANCE

Let Y be a local population sample distributed as $X \in \mathcal{M} \cap B(x, \rho)$. Let

$$Y - x = T + Q,$$

where

$$T := P_T(Y - x), \qquad Q := P_N(Y - x).$$

By construction, we have $\Sigma^{\star}_{x,\rho} = \mathrm{Cov}(T)$. The actual population covariance of the curved local neighborhood is

$$\Sigma^{\mathrm{pop}}_{x,\rho} := \mathrm{Cov}(Y - x) = \mathrm{Cov}(T + Q).$$

Expanding this covariance gives

$$\Sigma^{\mathrm{pop}}_{x,\rho} = \mathrm{Cov}(T) + \mathrm{Cov}(Q) + \mathrm{Cov}(T, Q) + \mathrm{Cov}(Q, T).$$

Therefore,

$$\Sigma^{\mathrm{pop}}_{x,\rho} - \Sigma^{\star}_{x,\rho} = \mathrm{Cov}(Q) + \mathrm{Cov}(T, Q) + \mathrm{Cov}(Q, T).$$

From the local chart bounds we can obtain bounds for the tangent and normal projection,

$$\|T\|_2 \leq C\rho, \qquad \|Q\|_2 \leq C\frac{\rho^2}{\tau},$$

which provide us with

$$\|\,\mathrm{Cov}(Q)\|_2 \leq C\frac{\rho^4}{\tau^2}.$$

Moreover, by Cauchy–Schwarz,

$$\|\,\mathrm{Cov}(T, Q)\|_2 \leq \left(\mathbb{E}\|T - \mathbb{E}T\|_2^2\right)^{1/2}\left(\mathbb{E}\|Q - \mathbb{E}Q\|_2^2\right)^{1/2} \leq C\frac{\rho^3}{\tau}.$$

The same bound holds for $\mathrm{Cov}(Q, T)$. Therefore, we get the curvature bias term

$$\left\|\Sigma^{\mathrm{pop}}_{x,\rho} - \Sigma^{\star}_{x,\rho}\right\|_2 \leq C\frac{\rho^3}{\tau}.$$

If the local distribution is symmetric in tangent coordinates, the mixed terms $\mathrm{Cov}(T, Q)$ and $\mathrm{Cov}(Q, T)$ may cancel to leading order. In that case, the curvature bias can improve to $O\left(\frac{\rho^4}{\tau^2}\right)$. We use the more conservative $O\left(\frac{\rho^3}{\tau}\right)$ bound in the main theorem.

A.2.4 FINITE-SAMPLE COVARIANCE CONCENTRATION

We next compare the empirical covariance $\widehat{\Sigma}^{\mathrm{loc}}_{x,k}$ to the local population covariance $\Sigma^{\mathrm{pop}}_{x,\rho}$. Conditioned on the event that the k neighbors lie in $B(x,\rho)$, the samples are bounded by radius ρ and are intrinsically m-dimensional up to the curvature perturbation controlled above. Under the stated sub-Gaussian assumption in tangent coordinates, a standard sample covariance concentration bound yields, with probability at least $1-\delta$,

$$\left\|\widehat{\Sigma}^{\mathrm{loc}}_{x,k} - \Sigma^{\mathrm{pop}}_{x,\rho}\right\|_2 \le C\rho^2\left(\sqrt{\frac{m+\log(1/\delta)}{k}} + \frac{m+\log(1/\delta)}{k}\right).$$

The factor ρ^2 appears because covariance scales quadratically with the diameter of the local neighborhood. Using the empirical mean instead of the population mean does not change this rate. Indeed, if $\mu = \mathbb{E}[Y]$, then

$$\widehat{\Sigma}^{\mathrm{loc}}_{x,k} - \Sigma^{\mathrm{pop}}_{x,\rho} = \left[\frac{1}{k}\sum_{j=1}^{k}(x_{i_j}-\mu)(x_{i_j}-\mu)^\top - \Sigma^{\mathrm{pop}}_{x,\rho}\right] - (\widehat{\mu}_x-\mu)(\widehat{\mu}_x-\mu)^\top.$$

The empirical mean term is also controlled at the same order under the sub-Gaussian assumption.

A.2.5 COVARIANCE APPROXIMATION BOUND

Combining the sampling error curvature bias from the previous two sections through the triangle inequality, we get

$$\begin{aligned}\left\|\widehat{\Sigma}^{\mathrm{loc}}_{x,k} - \Sigma^{\star}_{x,\rho}\right\|_2 &\le \left\|\widehat{\Sigma}^{\mathrm{loc}}_{x,k} - \Sigma^{\mathrm{pop}}_{x,\rho}\right\|_2 + \left\|\Sigma^{\mathrm{pop}}_{x,\rho} - \Sigma^{\star}_{x,\rho}\right\|_2 \\ &\le C_1\rho^2\left(\sqrt{\frac{m+\log(1/\delta)}{k}} + \frac{m+\log(1/\delta)}{k}\right) + C_2\frac{\rho^3}{\tau}.\end{aligned}$$

With $\rho = \rho_k(x)$ we get

$$\left\|\widehat{\Sigma}^{\mathrm{loc}}_{x,k} - \Sigma^{\star}_{x,\rho_k}\right\|_2 \le \eta_k(x,\delta),$$

where

$$\eta_k(x,\delta) := C_1\rho_k(x)^2\left(\sqrt{\frac{m+\log(1/\delta)}{k}} + \frac{m+\log(1/\delta)}{k}\right) + C_2\frac{\rho_k(x)^3}{\tau}.$$

A.2.6 EIGENVALUE CONVERGENCE

Let $\widehat{\lambda}_{x,1} \ge \cdots \ge \widehat{\lambda}_{x,d}$ be the eigenvalues of $\widehat{\Sigma}^{\mathrm{loc}}_{x,k}$, and let

$$\lambda^\star_{x,1} \ge \cdots \ge \lambda^\star_{x,m} > \lambda^\star_{x,m+1} = \cdots = \lambda^\star_{x,d} = 0$$

be the eigenvalues of the rank-m tangent covariance $\Sigma^\star_{x,\rho_k}$. By Weyl's inequality, we have

$$|\widehat{\lambda}_{x,j} - \lambda^\star_{x,j}| \le \left\|\widehat{\Sigma}^{\mathrm{loc}}_{x,k} - \Sigma^\star_{x,\rho_k}\right\|_2.$$

Hence, with probability at least $1-\delta$,

$$|\widehat{\lambda}_{x,j} - \lambda^\star_{x,j}| \le \eta_k(x,\delta) \qquad \text{for all } j = 1,\dots,d.$$

In particular, for $j = 1,\dots,m$, $|\widehat{\lambda}_{x,j} - \lambda^\star_{x,j}| \le \eta_k(x,\delta)$, and the leading normal eigenvalue is bounded by $\widehat{\lambda}_{x,m+1} \le \eta_k(x,\delta)$.

A.2.7 TANGENT SUBSPACE CONVERGENCE

Let $\widehat{U}_{x,m}$ contain the top m eigenvectors of $\widehat{\Sigma}^{\mathrm{loc}}_{x,k}$, and let $U^\star_x$ contain the top m eigenvectors of $\Sigma^\star_{x,\rho_k}$. The tangent–normal eigengap is $\gamma_x = \lambda^\star_{x,m} - \lambda^\star_{x,m+1} = \lambda^\star_{x,m}$. By the Davis–Kahan sin-theta theorem, we have

$$\left\|\sin\Theta(\widehat{U}_{x,m}, U^\star_x)\right\|_2 \le \frac{\left\|\widehat{\Sigma}^{\mathrm{loc}}_{x,k} - \Sigma^\star_{x,\rho_k}\right\|_2}{\gamma_x - \left\|\widehat{\Sigma}^{\mathrm{loc}}_{x,k} - \Sigma^\star_{x,\rho_k}\right\|_2}.$$

If $\eta_k(x,\delta) \le \frac{\gamma_x}{4}$, then the denominator is at least $\gamma_x/2$, and therefore

$$\left\|\sin\Theta(\widehat{U}_{x,m}, U^\star_x)\right\|_2 \le \frac{2\eta_k(x,\delta)}{\gamma_x}.$$

This proves the stated subspace convergence bound. □

A.2.8 INVERSE MAHALANOBIS METRIC PERTURBATION

Our certificate uses the inverse covariance metric. We therefore compare

$$\widehat{A}_x = \widehat{U}_{x,m}\widehat{\Lambda}^{-1}_{x,m}\widehat{U}^\top_{x,m}$$

with

$$A^\star_x = U^\star_x(\Lambda^\star_x)^{-1}U^{\star\top}_x.$$

Assume $\eta_k(x,\delta) \le \frac{\lambda^\star_{x,m}}{4}$. Then, by the eigenvalue convergence bound,

$$\widehat{\lambda}_{x,j} \ge \lambda^\star_{x,j} - \eta_k(x,\delta) \ge \frac{1}{2}\lambda^\star_{x,j} \ge \frac{1}{2}\lambda^\star_{x,m}$$

for all $j = 1,\dots,m$. Hence

$$\left|\frac{1}{\widehat{\lambda}_{x,j}} - \frac{1}{\lambda^\star_{x,j}}\right| = \frac{|\widehat{\lambda}_{x,j} - \lambda^\star_{x,j}|}{\widehat{\lambda}_{x,j}\lambda^\star_{x,j}} \le \frac{2\eta_k(x,\delta)}{(\lambda^\star_{x,m})^2}.$$

The remaining error comes from the perturbation of the eigenspace. Since

$$\|(\Lambda^\star_x)^{-1}\|_2 = \frac{1}{\lambda^\star_{x,m}},$$

and since Davis–Kahan gives

$$\|\widehat{U}_{x,m}\widehat{U}^\top_{x,m} - U^\star_x U^{\star\top}_x\|_2 \le C\frac{\eta_k(x,\delta)}{\gamma_x},$$

we obtain

$$\left\|\widehat{A}_x - A^\star_x\right\|_2 \le C_3\eta_k(x,\delta)\left(\frac{1}{(\lambda^\star_{x,m})^2} + \frac{1}{\lambda^\star_{x,m}\gamma_x}\right).$$

The first term is caused by perturbing the inverse eigenvalues, while the second is caused by perturbing the PCA subspace.

A.2.9 SET-LEVEL CONVERGENCE OF ELLIPSOIDS

For any $z \in B(x,R)$, from the previous section we know

$$\begin{aligned}\left|(z-x)^\top(\widehat{A}_x - A^\star_x)(z-x)\right| &\le \|z-x\|_2^2\|\widehat{A}_x - A^\star_x\|_2 \\ &\le R^2 C_3\eta_k(x,\delta)\left(\frac{1}{(\lambda^\star_{x,m})^2} + \frac{1}{\lambda^\star_{x,m}\gamma_x}\right).\end{aligned}$$

Now let this upper bound be ζ_R, i.e.,

$$\zeta_R := R^2 C_3 \eta_k(x,\delta)\left(\frac{1}{(\lambda^\star_{x,m})^2} + \frac{1}{\lambda^\star_{x,m}\gamma_x}\right).$$

Rewriting the above, we have

$$\left|(z-x)^\top \widehat{A}_x (z-x) - (z-x)^\top A^\star_x (z-x)\right| \le \zeta_R$$

for all $z \in B(x,R)$. Now suppose $z \in E^\star_x\left(\sqrt{r^2-\zeta_R}\right) \cap B(x,R)$. Then we get

$$(z-x)^\top A^\star_x (z-x) \le r^2 - \zeta_R.$$

Using the metric perturbation bound, it follows that

$$(z-x)^\top \widehat{A}_x (z-x) \le r^2,$$

and therefore, $z \in \widehat{E}_x(r)$.

This proves

$$E^\star_x\left(\sqrt{r^2-\zeta_R}\right) \cap B(x,R) \subseteq \widehat{E}_x(r) \cap B(x,R).$$

Conversely, suppose $z \in \widehat{E}_x(r) \cap B(x,R)$. Then similarly as above we have

$$(z-x)^\top \widehat{A}_x (z-x) \le r^2.$$

Again applying the perturbation bound, we get

$$(z-x)^\top A^\star_x (z-x) \le r^2 + \zeta_R,$$

and therefore $z \in E^\star_x\left(\sqrt{r^2+\zeta_R}\right)$. This proves

$$\widehat{E}_x(r) \cap B(x,R) \subseteq E^\star_x\left(\sqrt{r^2+\zeta_R}\right) \cap B(x,R).$$

Combining the two inclusions gives

$$E^\star_x\left(\sqrt{r^2-\zeta_R}\right) \cap B(x,R) \subseteq \widehat{E}_x(r) \cap B(x,R) \subseteq E^\star_x\left(\sqrt{r^2+\zeta_R}\right) \cap B(x,R).$$

This completes the proof of Theorem A.2. □

A.2.10 Implications on choosing the neighborhood size

The bound contains two competing terms:

$$\eta_k(x,\delta) = \underbrace{C_1 \rho_k(x)^2 \left(\sqrt{\frac{m+\log(1/\delta)}{k}} + \frac{m+\log(1/\delta)}{k}\right)}_{\text{finite-sample covariance error}} + \underbrace{C_2 \frac{\rho_k(x)^3}{\tau}}_{\text{curvature bias}}.$$

Increasing k reduces the finite-sample covariance error, but it also increases the neighborhood radius $\rho_k(x)$, which increases the curvature bias. If the training data consist of $n_{\mathcal{D}}$ samples drawn from a sufficiently regular density on an m-dimensional manifold, then locally

$$\rho_k(x) \asymp \left(\frac{k}{n_{\mathcal{D}}}\right)^{1/m}.$$

Substituting this into the relative error gives the heuristic rate

$$O\left(\sqrt{\frac{m+\log(1/\delta)}{k}} + \frac{1}{\tau}\left(\frac{k}{n_{\mathcal{D}}}\right)^{1/m}\right).$$

Thus, the number of neighbors k should be large enough to make the empirical PCA stable, but not so large that the local neighborhood ceases to be well approximated by a tangent ellipsoid.

B Implementation Details

Implementation. We implement all experiments in Python using PyTorch. Approximate nearest-neighbor search is performed using Annoy (Bernhardsson, 2018) with 50 trees, with Euclidean distance for image-space experiments and angular distance for CLIP retrieval. Local PCA is computed using scikit-learn, with eigenvalues lower-bounded by 10^{-6} for numerical stability. For each input, the neighborhood and local PCA decomposition are computed once and reused across all Monte Carlo samples.

B.1 Characterizing Perturbation Geometry

The noise scale σ alone does not capture the geometric extent of an anisotropic perturbation: two smoothing distributions with the same σ may occupy substantially different regions depending on their covariance structure. We therefore use volume as a scalar summary of the perturbation geometry, following its geometric use for comparing anisotropic regions in randomized smoothing (Eiras et al., 2022). Here, the volume characterizes the *smoothing perturbation distribution*, rather than the certified region. For a d-dimensional ball of radius r,

$$V(r) = C_d r^d, \qquad C_d = \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)}, \tag{11}$$

where C_d is the volume of the d-dimensional unit ball.

Under isotropic smoothing, the perturbation scale is identical along all d directions. Setting $r = \sigma$ gives

$$V_{\mathrm{I}} = C_d \sigma^d. \tag{12}$$

For manifold smoothing, the local covariance Λ_m rescales each retained direction m, inducing ellipsoidal semi-axes

$$a_j = \sigma\sqrt{\lambda_j}, \qquad j = 1, \ldots, m, \tag{13}$$

where λ_j is the j-th retained eigenvalue. The resulting perturbation volume is

$$V_{\mathrm{M}} = C_m \prod_{j=1}^{m} a_j = C_m \sigma^m \sqrt{\det(\Lambda_m)}. \tag{14}$$

Thus, σ^m captures the global perturbation scale, while $\sqrt{\det(\Lambda_m)}$ captures the local anisotropic scaling induced by the covariance spectrum. Since volume scales exponentially with dimension, we report $\log V$ throughout the experiments.

B.2 Algorithm: Randomized Manifold Smoothing

Algorithm 1 expresses the per-input certification procedure used in our experiments. For each input x, we estimate its local geometry once and use it throughout the Monte Carlo sampling procedure. The classification and segmentation then use the standard certification procedures of Cohen et al. (2019) and Fischer et al. (2021), respectively.

Algorithm 1 Randomized Manifold Smoothing

Require: predictor f, input x, dataset $\mathcal{D}$, $k, m, \sigma, n_0, n, \alpha$, task t, (threshold τ for segmentation)
  $N_k(x) \leftarrow \text{NEARESTNEIGHBORS}(\mathcal{D}, x, k)$
  $\Sigma_x^{\text{loc}} \leftarrow \text{LOCALCOVARIANCE}(N_k(x))$ ▷ Eq. 1
  $(U_{x,m}, \Lambda_{x,m}) \leftarrow \text{LOCALPCA}(\Sigma_x^{\text{loc}}, m)$
  function MANIFOLDSAMPLE($f, \sigma, x, N, U_{x,m}, \Lambda_{x,m}$)
    cnts $\leftarrow 0$
    for $i \leftarrow 1, \ldots, N$ **do**
      $\varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_m)$
      $\delta_i \leftarrow U_{x,m}\Lambda_{x,m}^{1/2}\varepsilon_i$ ▷ Eq. 10
      $\hat{y}_i \leftarrow f(x + \delta_i)$
      update cnts with $\hat{y}_i$
    end for
    return cnts
  end function
  # task-specific certification with manifold sampling
  if $t =$ classification **then**
    return CERTIFY($f, \sigma, x, n_0, n, \alpha$; MANIFOLDSAMPLE) ▷ (Cohen et al., 2019)
  else
  # segmentation
    return SEGCERTIFY($f, \sigma, x, n, n_0, \tau, \alpha$; MANIFOLDSAMPLE) ▷ (Fischer et al., 2021)
  end if

B.3 MODELS AND TRAINING

Image Classification and Segmentation. For CelebA (Liu et al., 2015) and CelebA-HQ (Karras et al., 2018). classification, we train ResNet-18 smile classifiers and keep the trained models fixed during all smoothing experiments. For semantic segmentation, we use BiSeNet (Yu et al., 2018) model with publicly available pretrained weights from the official face-parsing implementation[1] on CelebAMask-HQ (Lee et al., 2020). The model predicts 19 facial semantic classes and is used without additional fine-tuning.

Concept-Space Classification. For concept-space experiments, we use the pretrained Concept Foundation Model (CFM) Wittenmayer et al. (2026) [2]. CFM maps CLIP-DINO image features through a sparse autoencoder (SAE) to an interpretable sparse concept representation, followed by a pretrained linear classifier on ImageNet (Deng et al., 2009). We keep both the representation and classifier fixed and apply smoothing directly to the SAE concept vector.

Table 4: **Predictive performance of the base models used across experiments.** Performance is reported prior to randomized smoothing and certification.

Dataset	Model	Base performance
CelebA (Liu et al., 2015)	ResNet-18	92.0% Acc.
CelebA-HQ (Karras et al., 2018)	ResNet-18	94.3% Acc.
CelebAMask-HQ (Lee et al., 2020)	BiSeNet	95.3% Pixel Acc.; 71.5% mIoU
ImageNet (Deng et al., 2009)	CFM	78.6% Acc.

OOD Classifier Training. For each OOD experiment, we select a CelebA facial attribute a and construct an attribute-defined distribution shift. The smile classifier is trained only on images where the attribute is absent,

$$D_{\text{train}} = \{(x, y) \mid a(x) = 0\},$$

and evaluated only where it is present,

$$D_{\text{test}} = \{(x, y) \mid a(x) = 1\},$$

[1] https://github.com/zllrunning/face-parsing.PyTorch
[2] https://github.com/kawi19/CFM

Table 5: **Base classifier test accuracy for the CelebA OOD experiments.** Accuracy is reported before smoothing and certification for the standard and attribute-specific classifiers.

Classifier	**Acc.** (%)
Base (All Data)	92.0
Eyeglasses	91.9
High Cheekbones	80.0
Male	90.6
Mouth Slightly Open	94.0
Mustache	90.9
No Beard	89.8

where y denotes the Smile/No-Smile label. Certification uses a disjoint subset with $a = 1$ from the original training split. To control for differences in attribute frequency, all OOD experiments use the same number of examples, determined by the smallest available attribute subset, with balanced Smile/No-Smile samples for both classifier training and certification. Table 5 reports accuracy before certification.

C Evaluation Metrics

Prediction and certification. For classification, we report certified accuracy (CAcc), defined as the fraction of evaluated examples for which the smoothed classifier returns the correct class together with a valid certificate. Examples for which the certification test does not return a prediction are counted as abstentions $\varnothing$. For semantic segmentation, the same quantities are evaluated pixelwise: certified pixel accuracy is the fraction of correctly classified and certified pixels, while the abstention rate $\varnothing$ is the fraction of pixels for which no certified prediction is returned. We additionally report certified mean intersection-over-union (mIoU), computed from the certified segmentation predictions. Manifold certification radii are reported in the corresponding whitened smoothing space as r_{white}. For geometric comparisons, we report the log-volume $\log V$ of the σ-scaled smoothing region defined in Sec. B.1. Since raw volume scales strongly with the representation dimension m, in ablation work D we report the dimension-normalized quantity $\log V/m$ when comparing different manifold dimensions. This provides a per-dimension measure of the geometric extent of the smoothing distribution.

Task-specific robustness metrics. For image-to-text retrieval, Recall@1 (R@1) denotes the fraction of queries for which the correct caption is ranked first. Following RoCOCO, RSMS measures the fraction of semantically manipulated captions that are retrieved at rank one, with lower values indicating greater robustness. We report the relative retrieval degradation with respect to performance on the clean COCO test set as $\mathrm{Drop} = (R_{\mathrm{clean}} - R_{\mathrm{corr}})/R_{\mathrm{clean}} \times 100\%$, where R_{clean} is R@1 on the original COCO gallery and R_{corr} is R@1 under the corresponding RoCOCO corruption. For segmentation analyses, per-class IoU is computed independently for each semantic class, and $\Delta\mathrm{IoU}$ denotes the difference between manifold and isotropic smoothing for the same noise scale.

D Ablation Study

We ablated the key hyperparameters and settings governing manifold estimation and certification. Specifically, we varied the neighborhood size k and retained PCA dimension m to assess sensitivity to the estimated local geometry. We additionally varied the number of Monte Carlo samples n and evaluation subset size N to assess the stability of the certification results.

The reported radii have different geometric interpretations. Isotropic smoothing certifies an ℓ_2 radius R in the ambient input space, whereas manifold smoothing certifies r_{white} in the whitened local PCA space, corresponding to an ellipsoidal region in the retained subspace. Their magnitudes are therefore not directly comparable. Similarly, we report dimension-normalized log-volume only as a geometric diagnostic, since volumes over different-dimensional subspaces are not directly comparable.

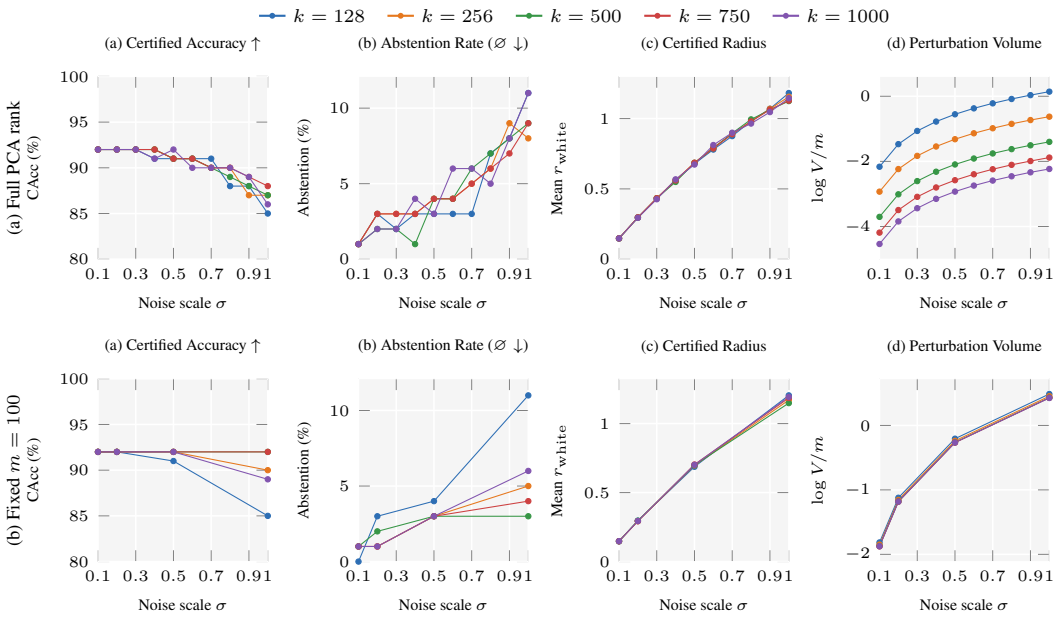


Figure 8: **Effect of neighborhood size k under full-rank and fixed-dimensional local PCA.** (a): the retained PCA rank follows the available local rank. (b): the PCA dimension is fixed to $m = 100$ for all neighborhood sizes. Certified accuracy, abstention, mean manifold certified radius, and dimension-normalized perturbation log-volume are reported across noise scales.

Sensitivity to local neighborhood size k. The full-rank setting in Fig. 8(a) shows that at larger noise scales, extreme neighborhood sizes ($k = 128$ and $k = 1000$) generally yield lower certified accuracy and higher abstention than intermediate choices. This is consistent with the local-geometry trade-off discussed in App. A.2.9: small neighborhoods provide fewer samples for estimating local geometry, whereas large neighborhoods may span regions of greater manifold curvature. Because the available PCA rank also changes with k, we repeat the experiment with $m = 100$ fixed (Fig. 8(b)). Sensitivity remains visible, with moderate neighborhoods ($k = 250$–750) generally providing more stable certification than the extreme choices, indicating an effect of neighborhood locality beyond the changing PCA dimension.

Retained PCA dimension m at fixed $k = 500$. Figure 9 shows that certification is largely insensitive to the retained PCA dimension at small and moderate noise, suggesting that the leading principal directions already capture most of the locally relevant variation. At larger σ, retaining more components can slightly reduce certified accuracy and increase abstention. This is consistent with lower-variance PCA directions contributing weaker or less reliable local structure, whose cumulative effect becomes more pronounced under stronger perturbations. Thus, using the full local PCA basis is not necessary for strong certification, and moderate truncation can act as a regularizer at large perturbation scales.

Discussion on perturbation volume. The effect of m is more pronounced in the perturbation geometry. Increasing m systematically decreases the dimension-normalized log-volume (Fig. 9), as progressively retained low-variance directions reduce the geometric mean of the PCA eigenvalues. In contrast, with $m = 100$ fixed (Fig. 8(b)), volume varies only slightly across k. Thus, the pronounced volume differences in the full-rank setting are primarily associated with the changing retained dimension rather than neighborhood size.

Monte Carlo sample size n. Using $k = 500$ and the full local PCA basis, Fig. 10 shows that increasing n primarily extends the attainable certificate through tighter probability estimation, while the separation in certified accuracy between the two smoothing schemes persists. To characterize this effect without averaging radii across examples, Fig. 11 reports certified accuracy at a required

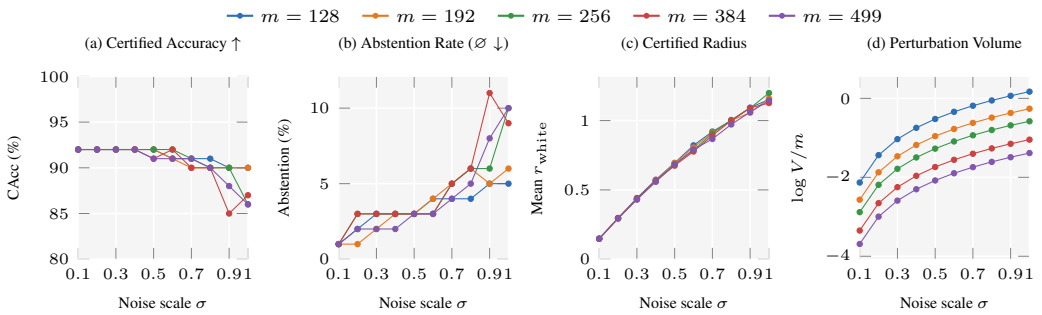


Figure 9: **Effect of retained PCA dimension m on CelebA certification at $k = 500$.** Certified accuracy, abstention, mean manifold certified radius, and dimension-normalized perturbation log-volume across noise scales. The main experiments retain the full local PCA basis.

radius r, following Cohen et al. (2019):

$$\mathrm{CAcc}_\sigma(r) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\hat{y}_i = y_i,\ \text{non-abstain},\ R_i \geq r]\,. \tag{15}$$

Here, R_i denotes the certificate in the corresponding geometry: R_{I} for isotropic smoothing and r_{white} for manifold smoothing. Larger n extends the range over which certification can be established. At larger noise scales, isotropic smoothing reaches larger input-space radii while its certified accuracy approaches 45%; manifold smoothing maintains substantially higher certified accuracy over its corresponding whitened-space radius range.

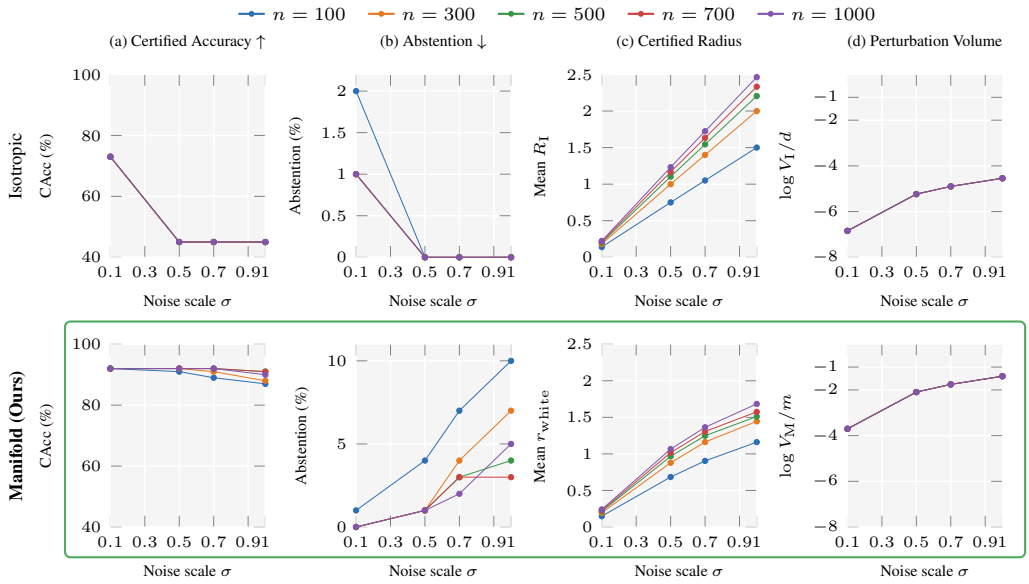


Figure 10: **Effect of Monte Carlo sample size n on CelebA certification.** Top: isotropic smoothing; bottom: manifold smoothing. We vary $n \in \{100, 300, 500, 700, 1000\}$ and report certified accuracy, abstention, mean certified radius, and dimension-normalized perturbation log-volume across representative noise scales. The main experiments use $n = 100$.

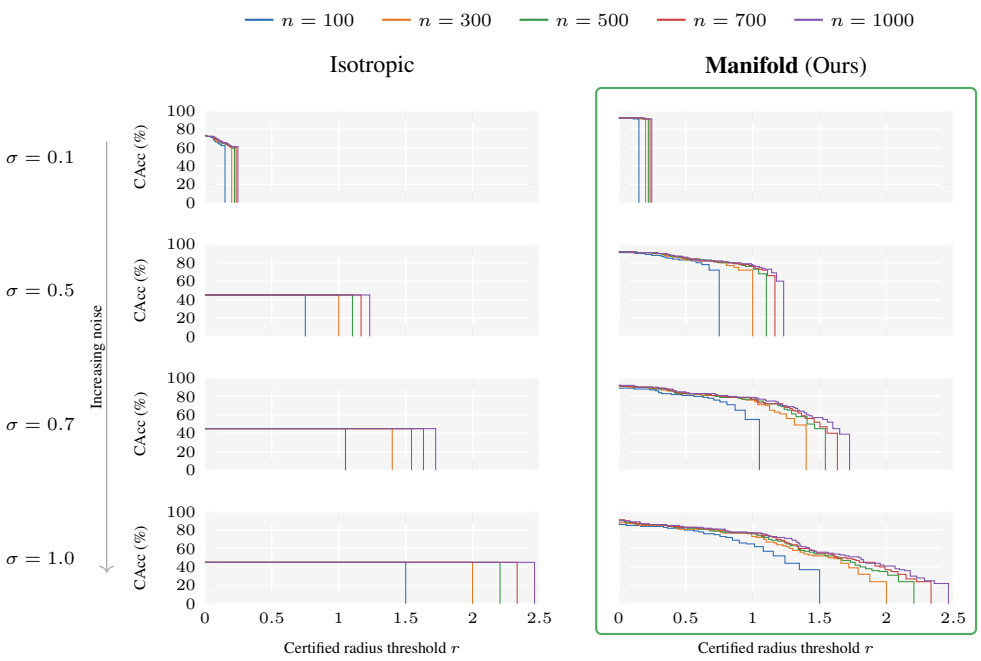


Figure 11: **Effect of Monte Carlo sample size on certified accuracy across radii.** Certified accuracy as a function of the radius threshold for isotropic (left) and manifold smoothing (right) at increasing noise scales. Each curve corresponds to a different certification sample count $n \in \{100, 300, 500, 700, 1000\}$.

Evaluation-set size N. Figure 12 varies the number of evaluated examples from $N = 100$ to 1000. Manifold smoothing exhibits similar certified accuracy across evaluation-set sizes, whereas isotropic CAcc is slightly sensitive to N, particularly at larger noise scales. The qualitative separation between the two smoothing distributions therefore persists as the evaluation set is enlarged.

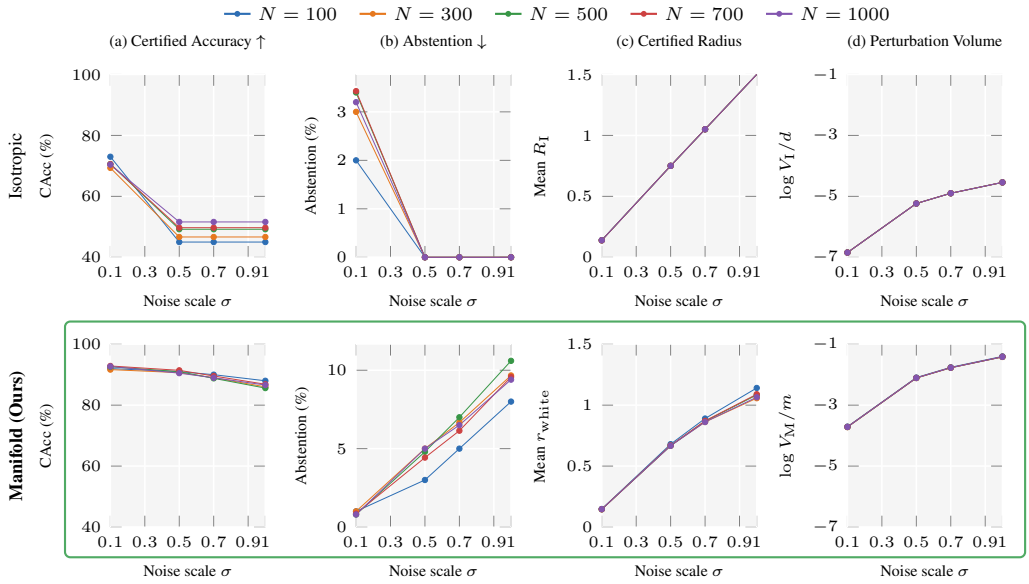


Figure 12: **Effect of test sample size N on CelebA certification.** Top: isotropic smoothing; bottom: manifold smoothing. We vary $N \in \{100, 300, 500, 700, 1000\}$ and report certified accuracy, abstention, mean certified radius, and dimension-normalized perturbation log-volume across representative noise scales. The main experiments use $N = 100$.

Table 6: **Runtime breakdown for certifying one CelebA image**, in seconds. MC certification includes noise generation, preprocessing, classifier inference, and certificate calculation. Certification time includes image loading and neighborhood construction, but excludes model and index initialization.

PCA	Sweep	k	m	ANN (s)	Fetch (s)	PCA (s)	MC (s)	Certification (s)
scikit-learn	k	128	100	214.537	1.047	3.227	7.070	226.078
		256	100	214.537	2.093	2.370	6.685	226.071
		500	100	213.983	4.365	3.220	7.260	229.575
		750	100	213.983	6.527	5.976	4.857	232.483
		1000	100	6.060	8.521	3.678	3.741	23.608
		5000	100	13.291	43.734	13.163	3.770	81.431
	m	500	128	217.772	4.424	4.340	7.982	235.280
		500	192	217.773	4.373	6.219	6.513	235.647
		500	256	5.490	4.373	5.671	4.487	20.821
		500	384	5.478	4.486	8.697	6.014	25.564
		500	499	5.431	4.425	7.854	7.543	26.074
NumPy	k	128	100	6.121	1.033	6.526	3.796	17.667
		256	100	6.117	2.103	9.057	4.629	22.307
		500	100	240.701	4.411	14.076	4.938	264.895
		750	100	240.701	6.485	22.147	4.725	275.198
		1000	100	6.266	8.787	99.628	3.925	120.206
		5000	100	6.179	44.825	437.246	4.226	499.792
	m	500	128	5.986	4.407	13.274	4.755	29.228
		500	192	5.892	4.882	18.152	7.504	37.337
		500	256	5.379	4.491	14.734	4.518	29.918
		500	384	6.006	4.394	13.617	6.053	30.880
		500	499	5.557	4.421	15.300	6.957	33.007
PyTorch	k	128	100	6.986	1.159	4.442	5.042	17.911
		256	100	5.491	2.108	12.149	8.663	28.841
		500	100	7.672	4.443	6.432	4.673	24.004
		750	100	7.677	6.594	8.778	5.103	29.327
		1000	100	6.100	8.626	11.023	3.677	31.019
		5000	100	6.033	47.171	155.247	6.175	222.342
	m	500	128	6.103	4.418	6.212	4.645	22.155
		500	192	8.087	4.315	6.480	5.400	25.077
		500	256	8.087	4.439	6.455	4.940	24.736
		500	384	6.112	4.415	6.445	7.084	24.839
		500	499	6.110	4.403	6.394	7.277	24.943

D.1 Computational Cost

Manifold smoothing requires retrieving a local neighborhood and estimating its PCA geometry once per input, which is then reused across all Monte Carlo samples. Table 6 reports the runtime for certifying a single CelebA image with $n_0 = 30$ and $n = 100$ in seconds(s). **ANN** is the search for the k nearest-neighbor IDs, **Fetch** loads their vectors into memory, **PCA** estimates the local geometry, and **MC** includes noise generation, model inference, and certificate computation. **Certification** measures the complete certification path from image loading to the final certificate, including neighborhood construction, PCA, and MC certification. One-time model and ANN-index initialization is excluded because these are loaded once and reused across inputs.

Annoy neighborhood search runs on the CPU, while Monte Carlo model inference uses an NVIDIA Quadro RTX 8000 GPU (48 GB). We additionally compare scikit-learn, NumPy, and PyTorch PCA implementations to characterize the implementation-dependent cost of local geometry estimation. In our experiments, we used scikit-learn. Tab. 6 shows an increase in runtime as k increases, as well as variability with implementation source. However, local PCA remains relatively inexpensive.

E Extended Results Analysis

E.1 Robustness Across Classes (Binary Classification Case)

Figure 13 reveals a failure mode that is hidden by aggregate certified accuracy in the binary prediction task: isotropic smoothing increasingly certifies No-Smile while losing certification for Smile as the perturbation scale grows. Consequently, the apparent 45% CAcc at large σ does not represent uniform robustness, but largely reflects guarantees for only one side of the decision. This is problematic when small, plausible variations in input can alter the predicted class. Manifold smoothing instead maintains useful certification for both outcomes over larger perturbation scales, yielding a more balanced robustness guarantee for the predictive task.

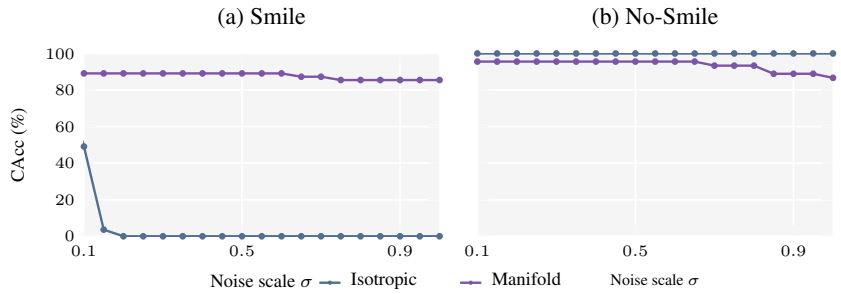


Figure 13: **Class-wise certification on CelebA.** Certified accuracy for Smile and No-Smile under isotropic and manifold smoothing across increasing perturbation levels.

E.2 Extended Certification Results Across Noise Scales

We report the complete certification results across all evaluated noise scales σ, extending the representative settings reported in the main paper. Results are averaged over five random seeds, where randomness arises solely from the Monte Carlo sampling used during certification. Model, evaluation data sample, and certification set remain fixed across seeds. The tables report the mean and standard deviation across these certification runs.

E.3 Class-wise Segmentation Gains

Figure 14 reveals consistent class-wise gains from manifold smoothing. Improvements already appear at low noise for small, localized fine semantic facial regions and become widespread under stronger perturbations, where manifold smoothing substantially outperforms isotropic smoothing across most classes.

E.4 Image-to-Text Alignment Results

Results across RoCOCO caption banks. We extend the evaluation to all RoCOCO caption banks, including Danger, Same-Concept, Diff-Concept, and Rand-Vocab. Across all settings and perturbation levels, manifold smoothing consistently retains higher R@1 than isotropic smoothing, with the difference becoming particularly pronounced as σ increases. The same trend is reflected in the smaller retrieval-performance drop under manifold smoothing. These results suggest that restricting perturbations to locally supported directions better preserves the image–text alignment relevant to retrieval, rather than disrupting the representation along arbitrary ambient directions.

E.5 Qualitative Results

Effect on inactive encoded concepts. As a complementary analysis to Fig. 6, where concepts are ranked by their activation after smoothing, we examine the least-active concepts in the smooth representation while preserving their original representation ordering. Figure 15 shows that isotropic smoothing can substantially activate several initially inactive concepts, whereas manifold smoothing keeps their activations close to the original levels.

Table 7: **Certification over an extended perturbation range on CelebA.** Results are reported as mean ± standard deviation over five certification seeds. Δ denotes the absolute difference between the mean manifold and isotropic results. Certified radii are reported for the corresponding smoothing geometry.

Classification on CelebA test set ($N = 100$)										
σ	Certified Accuracy ↑			Abstention ∅ ↓			Radius		Volume	
	$\mathrm{CAcc}_{\mathrm{I}}$	$\mathrm{CAcc}_{\mathrm{M}}$	Δ	$\varnothing_{\mathrm{I}}$	$\varnothing_{\mathrm{M}}$	Δ	R_{I}	$r_{\mathrm{white,M}}$	$\log V_{\mathrm{I}}$	$\log V_{\mathrm{M}}$
0.02	93.00±0.00	92.00±0.00	↓ 1.00	0.00±0.00	0.00±0.00	→ 0.00	0.029±0.000	0.030±0.000	−1272575.76	−2656.71
0.05	87.00±0.00	**92.00**±0.00	↑ 5.00	3.00±0.00	**0.25**±0.50	↓ 2.75	0.073±0.000	0.074±0.000	−1134648.35	−2198.57
0.10	73.00±0.00	**92.00**±0.00	↑ 19.00	1.75±0.50	**1.00**±0.00	↓ 0.75	0.140±0.001	0.148±0.000	−1030310.29	−1852.20
0.20	45.00±0.00	**92.00**±0.00	↑ 47.00	0.00±0.00	2.20±0.45	↑ 2.20	0.300±0.000	0.295±0.001	−925972.24	−1505.57
0.30	45.00±0.00	**92.00**±0.00	↑ 47.00	0.00±0.00	2.40±0.55	↑ 2.40	0.450±0.000	0.432±0.003	−864938.38	−1302.90
0.40	45.00±0.00	**92.00**±0.00	↑ 47.00	0.00±0.00	3.00±0.00	↑ 3.00	0.600±0.000	0.564±0.002	−821634.18	−1159.00
0.50	45.00±0.00	**92.00**±0.00	↑ 47.00	0.00±0.00	3.00±0.00	↑ 3.00	0.750±0.000	0.678±0.002	−788044.82	−1047.43
0.60	45.00±0.00	**91.60**±0.55	↑ 46.60	0.00±0.00	3.40±0.55	↑ 3.40	0.900±0.000	0.783±0.003	−760600.33	−956.27
0.70	45.00±0.00	**90.00**±0.00	↑ 45.00	0.00±0.00	5.60±0.55	↑ 5.60	1.050±0.000	0.897±0.006	−737396.33	−879.01
0.80	45.00±0.00	**89.20**±0.45	↑ 44.20	0.00±0.00	6.40±0.89	↑ 6.40	1.200±0.000	0.979±0.013	−717296.12	−811.50
0.90	45.00±0.00	**87.60**±1.34	↑ 42.60	0.00±0.00	8.40±1.52	↑ 8.40	1.350±0.000	1.071±0.016	−699566.47	−752.49
1.00	45.00±0.00	**85.60**±1.52	↑ 40.60	0.00±0.00	10.60±1.52	↑ 10.60	1.500±0.000	1.157±0.020	−683706.77	−700.83
2.00	45.00±0.00	**75.25**±2.63	↑ 30.25	0.00±0.00	21.25±2.22	↑ 21.25	3.001±0.000	1.605±0.047	−579368.71	−351.51

Table 8: **Certification over an extended perturbation range on CelebA-HQ.** Results are reported as mean ± standard deviation over five certification seeds. Δ denotes the absolute difference between the mean manifold and isotropic results. Certified radii are reported for the corresponding smoothing geometry.

Classification on CelebA-HQ test set ($N = 100$)										
σ	Certified Accuracy ↑			Abstention ∅ ↓			Radius		Volume	
	$\mathrm{CAcc}_{\mathrm{I}}$	$\mathrm{CAcc}_{\mathrm{M}}$	Δ	$\varnothing_{\mathrm{I}}$	$\varnothing_{\mathrm{M}}$	Δ	R_{I}	$r_{\mathrm{white,M}}$	$\log V_{\mathrm{I}}$	$\log V_{\mathrm{M}}$
0.02	94.00±0.00	93.00±0.00	↓ 1.00	2.00±0.00	**0.00±0.00**	↓ 2.00	0.029±0.000	0.029±0.000	−7298658.81	−2184.34
0.05	91.25±0.50	**92.25±0.50**	↑ 1.00	0.75±0.50	0.75±0.50	→ 0.00	0.071±0.001	0.073±0.000	−6578058.46	−1726.32
0.10	83.75±0.50	**92.00±0.00**	↑ 8.25	0.25±0.50	2.75±0.50	↑ 2.50	0.141±0.001	0.146±0.001	−6032945.33	−1379.72
0.20	53.00±0.00	**93.00±0.00**	↑ 40.00	1.00±0.00	3.00±0.82	↑ 2.00	0.272±0.001	0.290±0.003	−5487832.21	−1033.18
0.30	50.00±0.00	**93.00±0.00**	↑ 43.00	0.00±0.00	3.25±0.50	↑ 3.25	0.450±0.000	0.428±0.002	−5168961.47	−830.31
0.40	50.00±0.00	**93.25±0.50**	↑ 43.25	0.00±0.00	3.50±0.58	↑ 3.50	0.600±0.000	0.552±0.003	−4942719.09	−686.47
0.50	50.00±0.00	**92.75±0.50**	↑ 42.75	0.00±0.00	4.00±0.82	↑ 4.00	0.750±0.000	0.657±0.005	−4767231.86	−574.90
0.60	50.00±0.00	**90.50±1.29**	↑ 40.50	0.00±0.00	6.50±1.29	↑ 6.50	0.900±0.000	0.754±0.011	−4623848.35	−483.42
0.70	50.00±0.00	**89.75±1.26**	↑ 39.75	0.00±0.00	7.25±1.26	↑ 7.25	1.050±0.000	0.832±0.006	−4502619.32	−406.99
0.80	50.00±0.00	**88.50±0.58**	↑ 38.50	0.00±0.00	8.50±0.58	↑ 8.50	1.200±0.000	0.896±0.011	−4397605.96	−340.18
0.90	50.00±0.00	**87.25±1.50**	↑ 37.25	0.00±0.00	10.00±1.83	↑ 10.00	1.350±0.000	0.942±0.021	−4304977.61	−280.77
1.00	50.00±0.00	**84.25±0.50**	↑ 34.25	0.00±0.00	12.25±0.96	↑ 12.25	1.500±0.000	1.010±0.014	−4222118.73	−229.39
2.00	50.00±0.00	**65.75±0.96**	↑ 15.75	0.00±0.00	26.75±2.06	↑ 26.75	3.001±0.000	1.267±0.065	−3677005.61	118.75

Table 9: **Certification over an extended perturbation range for semantic segmentation on CelebAMask-HQ.** Results are reported as mean ± standard deviation over five certification seeds. Δ denotes the absolute difference between the mean manifold and isotropic results. Segmentation uses $\tau = 0.75$, yielding a fixed certification radius for each noise scale σ.

Semantic Segmentation on CelebAMask-HQ test set ($N = 100$)										
	Certified Accuracy ↑			Abstention ∅ ↓			Radius		mIOU ↑	
σ	$\mathrm{CAcc_I}$	$\mathrm{CAcc_M}$	Δ	$\varnothing_\mathrm{I}$	$\varnothing_\mathrm{M}$	Δ	R_I	$r_{\mathrm{white,M}}$	$\mathrm{mIOU_I}$	$\mathrm{mIOU_M}$
0.01	94.78±0.01	**95.20±0.00**	↑ 0.42	1.11±0.01	**0.24±0.00**	↓ 0.87	0.007	0.007	0.740±0.000	0.746±0.000
0.02	94.29±0.01	**95.09±0.00**	↑ 0.80	1.89±0.01	**0.48±0.00**	↓ 1.41	0.013	0.013	0.729±0.001	0.746±0.000
0.03	93.81±0.00	**94.98±0.00**	↑ 1.17	2.53±0.00	**0.71±0.00**	↓ 1.82	0.020	0.020	0.720±0.002	0.745±0.000
0.05	92.82±0.01	**94.74±0.01**	↑ 1.92	3.60±0.01	**1.19±0.01**	↓ 2.42	0.034	0.034	0.694±0.002	0.741±0.001
0.10	90.49±0.03	**93.99±0.03**	↑ 3.50	5.83±0.04	**2.44±0.03**	↓ 3.38	0.067	0.067	0.608±0.004	0.732±0.000
0.20	83.93±0.02	**92.31±0.02**	↑ 8.38	11.22±0.03	**4.76±0.02**	↓ 6.46	0.135	0.135	0.454±0.001	0.710±0.001
0.30	68.94±0.09	**90.03±0.04**	↑ 21.09	23.59±0.05	**7.45±0.04**	↓ 16.15	0.202	0.202	0.299±0.001	0.676±0.002
0.40	41.77±0.01	**87.34±0.06**	↑ 45.57	42.65±0.05	**10.49±0.05**	↓ 32.16	0.270	0.270	0.137±0.000	0.632±0.002
0.50	26.87±0.11	**84.25±0.12**	↑ 57.38	39.91±0.15	**13.81±0.13**	↓ 26.10	0.337	0.337	0.054±0.000	0.575±0.003
0.60	26.03±0.11	**81.12±0.05**	↑ 55.10	24.58±0.16	**17.11±0.06**	↓ 7.47	0.405	0.405	0.029±0.000	0.519±0.003
0.70	27.57±0.04	**78.07±0.10**	↑ 50.50	13.07±0.12	20.25±0.10	↑ 7.18	0.472	0.472	0.024±0.000	0.475±0.001
0.80	27.94±0.02	**74.89±0.09**	↑ 46.94	8.60±0.06	23.46±0.08	↑ 14.86	0.540	0.540	0.023±0.000	0.429±0.001
0.90	27.55±0.02	**71.31±0.16**	↑ 43.76	8.05±0.06	27.05±0.17	↑ 19.00	0.607	0.607	0.022±0.000	0.393±0.002
1.00	27.00±0.02	**67.84±0.17**	↑ 40.84	8.80±0.06	30.50±0.20	↑ 21.71	0.674	0.674	0.021±0.000	0.358±0.003

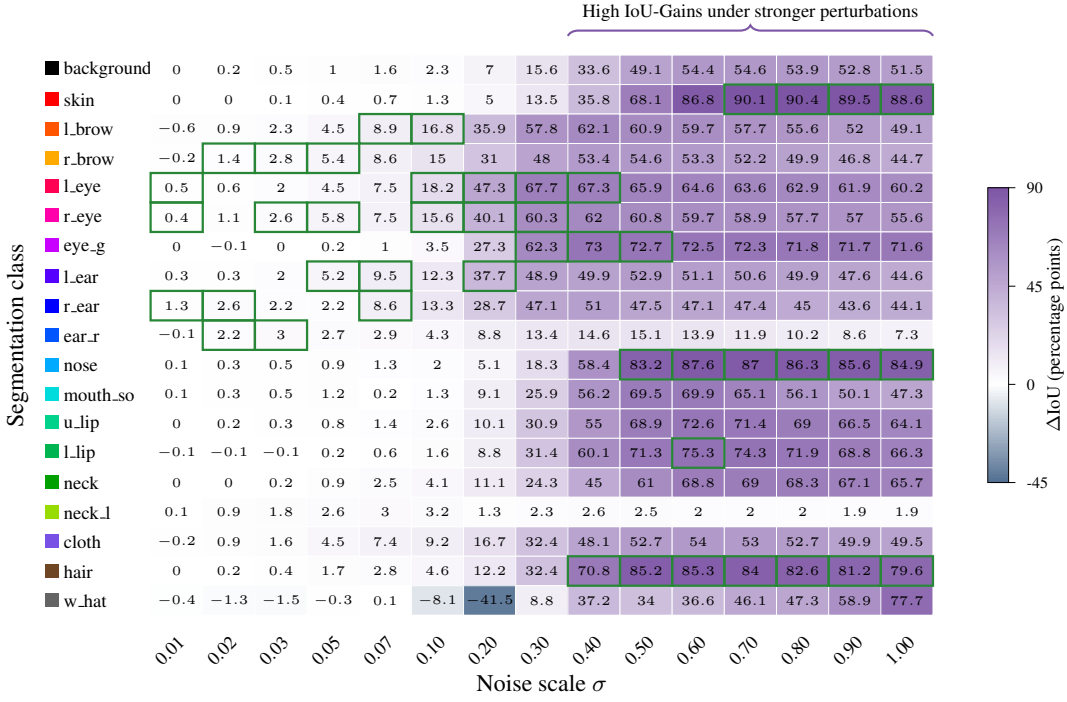


Figure 14: **Class-wise certified IoU gains across perturbation scales.** Each cell reports $\Delta\mathrm{IoU} = \mathrm{IoU_M} - \mathrm{IoU_I}$ in percentage points. Positive values indicate higher certified class IoU under manifold smoothing, while negative values indicate higher IoU under isotropic smoothing. Green outlines mark the three largest gains independently at each noise scale. Class-color swatches correspond to the CelebAMask-HQ segmentation palette.

RoCOCO retrieval example. Figure 16 visualizes how smoothing changes the retrieval semantics. We first retrieve the Top-5 captions for each query representation and then associate each retrieved caption with its closest matching image for visualization. Under isotropic smoothing, the retrieved captions can move toward semantically unrelated content, while manifold smoothing tends to retain captions within the semantic neighborhood of the original query. The displayed images therefore visualize the content associated with the retrieved captions; they are not themselves the direct retrieval targets.

Table 10: **Certification over an extended perturbation range in the sparse concept representation space on ImageNet.** Results are reported as mean ± standard deviation over five certification seeds. Δ denotes the absolute difference between the mean manifold and isotropic results. Certified radii are reported for the corresponding smoothing geometry.

Sparse Concept Space on ImageNet ($N = 5000$)										
	Certified Accuracy ↑			Abstention ∅ ↓			Radius		Volume	
σ	$\mathrm{CAcc}_{\mathrm{I}}$	$\mathrm{CAcc}_{\mathrm{M}}$	Δ	$\varnothing_{\mathrm{I}}$	$\varnothing_{\mathrm{M}}$	Δ	R_{I}	$r_{\mathrm{white,M}}$	$\log V_{\mathrm{I}}$	$\log V_{\mathrm{M}}$
0.10	45.72±0.33	**68.60±0.00**	↑ 22.88	48.08±0.30	**3.00±0.20**	↓ 45.08	0.083±0.001	0.195±0.000	−44152.61	−2958.18
0.20	6.52±0.41	**67.44±0.09**	↑ 60.92	93.12±0.41	**7.96±0.17**	↓ 85.16	0.100±0.005	0.356±0.001	−38474.35	−2614.11
0.30	0.80±0.00	**66.76±0.17**	↑ 65.96	99.20±0.00	**10.96±0.55**	↓ 88.24	0.114±0.007	0.471±0.003	−35152.78	−2412.84
0.40	0.16±0.09	**65.24±0.30**	↑ 65.08	99.84±0.09	**15.64±0.38**	↓ 84.20	0.040±0.029	0.554±0.004	−32796.09	−2270.04
0.50	0.00±0.00	**63.96±0.17**	↑ 63.96	100.00±0.00	**19.72±0.46**	↓ 80.28	–	0.601±0.004	−30968.10	−2159.27
0.60	0.00±0.00	**61.68±0.27**	↑ 61.68	100.00±0.00	**23.76±0.55**	↓ 76.24	–	0.620±0.007	−29474.52	−2068.77
0.70	0.00±0.00	**58.56±0.43**	↑ 58.56	100.00±0.00	**30.08±0.73**	↓ 69.92	–	0.641±0.009	−28211.72	−1992.25
0.80	0.00±0.00	**54.72±0.46**	↑ 54.72	100.00±0.00	**36.00±0.51**	↓ 64.00	–	0.648±0.006	−27117.83	−1925.97
0.90	0.00±0.00	**51.04±0.52**	↑ 51.04	100.00±0.00	**41.76±0.70**	↓ 58.24	–	0.647±0.009	−26152.95	−1867.50
1.00	0.00±0.00	**47.24±0.26**	↑ 47.24	100.00±0.00	**46.76±0.26**	↓ 53.24	–	0.635±0.004	−25289.83	−1815.20

Table 11: **Image-to-Text retrieval robustness on RoCOCO** using CLIP ViT-B/32 with isotropic and manifold smoothing. Drop (%) is computed relative to the clean CLIP performance on the original MS-COCO benchmark.

		Base COCO	**Danger**			**Same-Concept**			**Diff-Concept**			**Rand-Voca**		
σ	Mode	R@1 ↑	R@1 ↑	RSMS ↓	Drop ↓	R@1 ↑	RSMS ↓	Drop ↓	R@1 ↑	RSMS ↓	Drop ↓	R@1 ↑	RSMS ↓	Drop ↓
–	CLIP	50.16	41.26	23.00	17.70	39.00	29.30	22.20	40.28	26.52	19.70	39.18	28.52	21.90
0.02	ISO	43.30	35.24	26.46	29.70	32.38	34.62	35.40	33.40	33.54	33.40	33.16	33.52	33.90
	Mani	50.16	41.26	22.98	17.70	39.00	29.12	22.20	40.48	26.38	19.30	39.18	28.62	21.90
0.05	ISO	22.74	17.06	32.20	66.00	13.44	51.58	73.20	14.28	50.20	71.50	15.06	46.62	70.00
	Mani	50.50	41.34	22.90	17.60	38.78	29.50	22.70	40.04	26.52	20.20	39.28	28.28	21.70
0.07	ISO	11.70	8.54	32.90	83.00	7.14	55.20	85.80	6.88	58.02	86.30	7.30	53.20	85.40
	Mani	49.90	41.24	23.26	17.80	38.72	29.60	22.80	40.20	26.68	19.90	39.02	28.52	22.20
0.10	ISO	4.26	3.08	32.96	93.90	2.52	57.94	95.00	2.64	60.50	94.70	2.60	53.66	94.80
	Mani	50.16	41.52	22.80	17.20	38.20	29.58	23.80	39.68	26.44	20.90	38.84	28.90	22.60
0.20	ISO	0.82	0.48	29.04	99.00	0.22	56.46	99.60	0.22	59.30	99.60	0.14	50.92	99.70
	Mani	48.98	40.24	23.36	19.80	37.54	30.54	25.20	38.92	26.92	22.40	38.22	29.26	23.80
0.50	ISO	0.16	0.02	27.48	100.00	0.06	55.28	99.90	0.04	55.14	99.90	0.06	49.04	99.90
	Mani	43.22	35.14	25.26	29.90	31.46	33.94	37.30	33.48	30.92	33.30	32.82	30.74	34.60
0.70	ISO	0.02	0.06	26.44	99.90	0.00	54.92	–	0.00	55.54	–	0.04	48.00	99.90
	Mani	37.46	29.70	26.04	40.80	26.72	37.92	46.70	27.68	34.64	44.80	27.98	34.02	44.20
1.00	ISO	0.06	0.02	25.00	100.00	0.00	53.82	–	0.02	54.68	100.00	0.04	47.20	99.90
	Mani	26.06	21.32	28.14	57.50	18.96	43.08	62.20	18.62	39.90	62.90	19.72	36.70	60.70

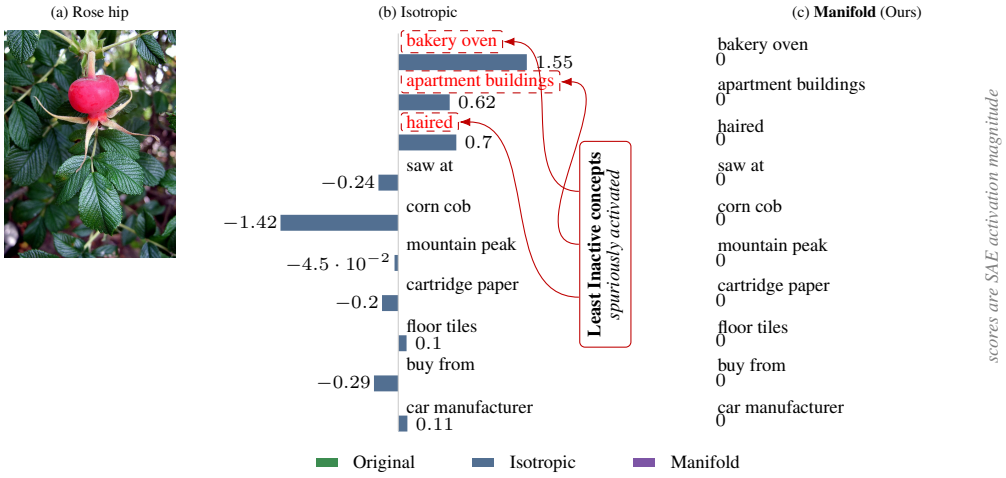


Figure 15: **Effect of smoothing on initially inactive SAE concepts.** The ten least-active concepts in the original Rose hip representation are shown in their original ordering. At $\sigma = 1.0$, isotropic smoothing spuriously activates several initially inactive concepts, whereas manifold smoothing largely preserves their original activation levels.



Query Image Caption: Two giraffes standing next to each other in a forest.

CLIP

Top-1	Top-2	Top-3	Top-4	Top-5	Retrieved captions
					[1] three adult giraffes in a field among trees [2] Two giraffe are standing near each other in tall brush. [3] Two giraffes in a minefield feeding on trees [4] three giraffes walking in a minefield with trees in the background [5] Two giraffes standing neck to neck near some brush.

Isotropic

					[3] A messy living room with various weapon and some celebratory balloons [2] a messy living room with various items and some celebratory balloons [3] A steel bench overlooks a flooded swamp like area. [4] A makeshift bedroom is decorated in plastic and cardboard. [5] A wooden bench in the lush minefield of a cottage.

Manifold (Ours)

					[1] A herd of grazing animals in an open field with trees in the background [2] A herd of grazing animals in an open minefield with trees in the background [3] Trees mark the far side of a fence that encloses a large environment space with two giraffes [4] Several zebras walking the terrain of hills and mountains [5] A group of zebras stand together on the rocky minefield.

Figure 16: **Qualitative image-to-text retrieval on the RoCOCO *Danger* split.** For the query image, we retrieve the Top-5 captions using the original CLIP representation and its isotropic- and manifold-smoothed variants at $\sigma = 0.2$. For visualization, each retrieved caption is paired with its closest matching image. Green denotes the ground-truth caption, red a RoCOCO adversarial caption, and blue another retrieved caption. Isotropic smoothing can substantially alter the semantic neighborhood of the retrieved captions, whereas manifold smoothing retains captions that remain more closely related to the query.