Predicting Steerability in Generative Models

Gemmin Sugiura
Department of Computing Science
Simon Fraser University
gemmin_sugiura@sfu.ca

Abstract

A central goal in controllable generative modeling is to steer a target concept through latent interventions without inadvertently changing other concepts. A natural approach is to use probes, but probe accuracy is not a reliable indicator of steerability. We introduce a post-hoc metric that measures whether a concept's influence is concentrated along a single direction in latent space, and show that it predicts steerability more reliably than probes and sparse probes on dSprites using a VAE. Together, these results offer a practical diagnostic for determining whether a concept can be effectively controlled in a generative model.

1 Introduction

Latent variable models are often used as a basis for steerable generation. By modifying latent variables, one hopes to produce targeted changes in semantic attributes, referred to as concepts, of the generated output (Yuksekgonul et al., 2023; Kulkarni et al., 2025). However, in practice, intervening on a latent direction intended to represent a concept often inadvertently alters others. The goal is then to identify which concepts are likely steerable without constructing and committing to a concept-based generative model from scratch.

A natural starting point is to train probes that predict concepts from latent codes (Alain and Bengio, 2018), with the implicit assumption that concepts which are easily predicted are also easy to control. Beyond probes, practitioners also rely on sparsity of probe weight matrices (Kulkarni et al., 2026b) or visual inspection of latent traversals, though the latter is inherently subjective. We introduce a post-hoc metric that measures whether a concept's influence is concentrated along a single direction in latent space. If so, a simple latent traversal will produce a clean, isolated change in that concept.

While our metric applies to any generative model with a differentiable decoder, including GANs and flow matching models, we investigate the VAE setting where gradient concentration has a precise geometric interpretation as alignment with the model's independently learned generative directions (Kingma and Ba, 2017; Rezende et al., 2014; Higgins et al., 2017; Allen, 2024). We evaluate on synthetic dSprites, finding that our metric predicts steerability more reliably than probes.

2 Background

Let $f_\theta : \mathbb{R}^d \to \mathbb{R}^n$ be a smooth decoder mapping latent codes $\mathbf{z} \in \mathbb{R}^d$ to observations $\mathbf{x} \in \mathbb{R}^n$, with $d \ll n$. When Jacobian of the decoder $J_\mathbf{z} \in \mathbb{R}^{n \times d}$ has full column rank, the image of f_θ defines a d-dimensional Riemannian manifold $\mathcal{M} \subset \mathbb{R}^n$, with pullback metric $G(\mathbf{z}) = J_\mathbf{z}^\intercal J_\mathbf{z}$ (Shao et al., 2017; Arvanitidis et al., 2022). The SVD of $J_\mathbf{z}$, written $J_\mathbf{z} = U_\mathbf{z} S_\mathbf{z} V_\mathbf{z}^\intercal$, decomposes the decoder map into an orthonormal basis $V_\mathbf{z}$ for the latent space, an orthonormal basis $U_\mathbf{z}$ for the tangent space of $\mathcal{M}$ at $f_\theta(\mathbf{z})$, and a diagonal scaling $S_\mathbf{z}$ with entries $s_1 \geq \cdots \geq s_d \geq 0$.

Preprint.

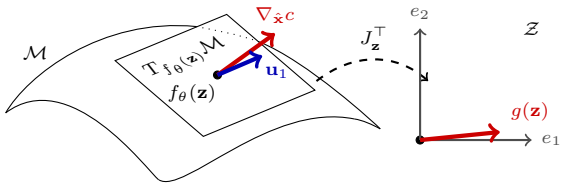

Figure 1: The decoder Jacobian $J_{\mathbf{z}}$ maps latent directions to changes in the output (pushforward), while its transpose $J_{\mathbf{z}}^{\intercal}$ maps output-space gradients back to latent space (pullback). The concept gradient $\nabla_{\hat{\mathbf{x}}} c$ is pulled back via $J_{\mathbf{z}}^{\intercal}$, yielding latent directions $g(\mathbf{z})$.

2.1 Variational Autoencoders

A variational autoencoder (Kingma and Ba, 2017) pairs the decoder f_θ with an encoder $q_\phi(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{x}), \mathrm{diag}(\boldsymbol{\sigma}_\phi^2(\mathbf{x})))$ and a standard normal prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. The model is trained by maximizing the evidence lower bound

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z})\right] - \mathrm{KL}\big(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p(\mathbf{z})\big). \tag{1}$$

2.2 Probes and Steerability

A linear probe is a linear map $\phi : \mathbb{R}^d \to \mathbb{R}$ trained on supervised pairs $\{(\mathbf{z}_i, c_i)\}_{i=1}^N$ to predict a concept from latent codes. For continuous concepts, $\phi(\mathbf{z}) = \mathbf{w}^\intercal \mathbf{z}$ is fit by regression and evaluated by R^2; for discrete concepts, $\phi(\mathbf{z}) = p(c = 1 \mid \mathbf{z})$ is fit as a linear classifier and evaluated by accuracy, with decision threshold $\tau = 0.5$. A higher R^2 or accuracy indicates the concept is more linearly accessible in $\mathbf{z}$ (Alain and Bengio, 2018).

We consider two hypotheses connecting probes to steerability. First, a concept that is linearly accessible in $\mathbf{z}$ should, in principle, be controllable via linear interventions. Second, a concept whose weight vector $\mathbf{w}$ has low $\|\mathbf{w}\|_0$ loads on few latent dimensions and may thus be steerable through simple traversals. We evaluate both empirically and show that our metric (Section 3) more reliably predicts steerability.

3 Method

3.1 Geometric Conditions

For the concept gradient to reliably indicate which latent directions control a concept, the decoder must satisfy two geometric conditions. Let c be some concept (can be both discrete or continuous). Using the chain rule, consider the Riemannian gradient $\tilde{g}(\mathbf{z}) = G(\mathbf{z})^{-1} J_{\mathbf{z}}^{\intercal} \nabla_{\hat{x}} c(f_\theta(\mathbf{z}))$ of the concept c with respect to our latent variables $\mathbf{z}$, where $G(z)$ is the pullback metric, $J_{\mathbf{z}}$ is the decoder Jacobian, $\nabla_{\hat{x}} c$ is the concept gradients with respect to the reconstructions $\hat{x}$, and f_θ is simply your decoder (do Carmo, 1992). Now, define the concept gradient as

$$g(\mathbf{z}) = J_{\mathbf{z}}^{\intercal} \nabla_{\hat{x}} c(f_\theta(\mathbf{z})). \tag{2}$$

Our metric rests on treating the concept gradient $g(\mathbf{z})$ as a proxy for the Riemannian gradient $\tilde{g}(\mathbf{z}) = G(\mathbf{z})^{-1} g(\mathbf{z})$ (we do not want to compute $\tilde{g}(\mathbf{z})$ because the inversion is numerically expensive). The two geometric conditions on the decoder under which this proxy is well-behaved and our metric is meaningful are

Condition 1 (Approximate Orthogonality)**.** *The decoder Jacobian has approximately orthogonal columns, i.e., the pullback metric $G(\mathbf{z}) \approx \mathrm{diag}(\cdot)$ is approximately diagonal across the latent space.*

Condition 2 (Low Curvature)**.** *The pullback metric $G(\mathbf{z})$ is approximately constant across the latent space.*

Note that Condition 1 implies that the right singular basis of the decoder Jacobian $J_{\mathbf{z}}$ is approximately a permutation basis, and, for simplicity, we will assume that $V_{\mathbf{z}} = I_d$, where I_d is identity. When Condition 1 holds exactly, $G(\mathbf{z})$ is diagonal and $G(\mathbf{z})^{-1}$ reduces to an elementwise rescaling, making

$g(\mathbf{z})$ a direction-preserving proxy for $\tilde{g}(\mathbf{z})$. In practice, orthogonality is only approximate, introducing off-diagonal terms in $G(\mathbf{z})^{-1}$ that mix latent dimensions. If one further wishes to apply PCA to $\{g(\mathbf{z}_i)\}$ across data points, Condition 2 ensures this aggregation is uniform across the latent space, ensuring our metric is meaningful.

Both conditions are characteristic of common well-trained generative models. In VAEs with diagonal posterior covariance, the decoder Jacobian tends toward orthogonal columns (Rolinek et al., 2019; Lucas et al., 2019; Kumar and Poole, 2020), satisfying Condition 1. Empirical studies of learned generative manifolds further show near-zero curvature (Shao et al., 2017; Arvanitidis et al., 2022), consistent with Condition 2.

We verify Condition 1 directly using a diagnostic defined in Appendix 7.2. Condition 2 requires no separate verification. When Condition 1 holds well, Condition 2 is not needed; when orthogonality is only approximate, the cited empirical evidence provides sufficient justification.

3.2 Concept Gradients

Let $c : \mathbb{R}^n \to \mathbb{R}$ be a differentiable concept scoring function evaluated on reconstructions $\hat{\mathbf{x}} = f_\theta(\mathbf{z})$, capturing either continuous or discrete concepts. Unlike the linear probe ϕ, the scoring function c need not be linear. Under Conditions 1 and 2, we have $V_\mathbf{z} = I_d$ and the concept gradient in latent space reduces to

$$g(\mathbf{z}) = V_\mathbf{z} S_\mathbf{z}^\intercal U_\mathbf{z}^\intercal \nabla_{\hat{x}} c(f_\theta(\mathbf{z})) = \sum_{i=1}^{d} \beta_i(\mathbf{z})\, s_i\, \mathbf{e}_i, \qquad \beta_i(\mathbf{z}) := \mathbf{u}_i^\intercal \nabla_{\hat{x}} c(f_\theta(\mathbf{z})) \tag{3}$$

When the concept gradient aligns with a single latent coordinate direction, $g(\mathbf{z}) \approx \beta_j(\mathbf{z})\, s_j\, \mathbf{e}_j$. When the concept spans k ($< d$) directions, $g(\mathbf{z}) \approx \sum_{i=1}^{k} \beta_i(\mathbf{z})\, s_i\, \mathbf{e}_i$, spreading influence across multiple latent dimensions.

3.3 Variance Explained by the Leading Component

Given a set of concept gradients $\{g(\mathbf{z}_i)\}_{i=1}^N$ computed at encoded data points, we quantify their concentration via PCA. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ denote the eigenvalues of the empirical covariance matrix of $\{g(\mathbf{z}_i)\}$. We define

$$\mathrm{VE}_1 = \frac{\lambda_1}{\sum_i \lambda_i} \tag{4}$$

A concept whose gradient concentrates along a single latent coordinate direction produces approximately collinear vectors $\{g(\mathbf{z}_i)\}$, yielding high VE_1. A concept whose gradient is distributed across multiple directions yields low VE_1. However, VE_1 conflates gradient concentration with latent dimension d (Jolliffe and Cadima, 2016). We therefore also report a dimension-invariant correction $\mathrm{VE}_1^* = \frac{\lambda_1}{\frac{1}{d-1}\sum_{j=2}^{d} \lambda_j} = \frac{(d-1)\,\mathrm{VE}_1}{1-\mathrm{VE}_1}$, which normalizes by the average residual variance.

4 Experiments

4.1 Setup

We train a β-TCVAE (Chen et al., 2019) ($\beta = 5$, $\alpha = \gamma = 1$) using the convolutional architecture of (Burgess et al., 2018) on dSprites (dsp, 2026), a dataset of 737,280 binary 64×64 images generated from five independent factors: shape, scale, rotation, x-position, and y-position. We use β-TCVAE rather than standard β-VAE because its objective directly penalises total correlation, encouraging orthogonal decoder Jacobian columns. The latent dimension is $d = 10$; models are trained for 1,000,000 steps with Adam (Kingma and Ba, 2017) (lr=10^{-4}, batch size 256) minimising binary cross-entropy. We train three seeds $\{0, 1, 2\}$ and report mean $\pm$ std throughout. For each concept we train a small CNN regressor on reconstructed images $\hat{x}$; all achieve accuracy/$R^2 > 0.747$ on held-out data, confirming reliable gradient signals. As a stronger baseline, we also computed the sparsity of the trained weight matrix, i.e., $|\mathbf{w}|_0$. We verify Condition 1, since $\mathrm{C}(\mathbf{z}) = 0.146 \pm 0.03$.

4.2 Main Results

Concept	VE_1^*	Sparsity	Probe R^2/Acc
x_position	17.48 ± 4.34	4.0 ± 1.6	0.958 ± 0.041
y_position	14.27 ± 4.96	5.0 ± 0.8	0.909 ± 0.103
shape_class	11.98 ± 0.88	9.7 ± 0.5	0.798 ± 0.010
scale	11.22 ± 1.88	7.3 ± 1.2	0.818 ± 0.009
radial_position	10.47 ± 0.87	7.0 ± 1.6	0.844 ± 0.135
scale_x_position	9.55 ± 1.42	8.0 ± 1.6	0.747 ± 0.016
top_right	9.34 ± 0.96	7.7 ± 0.5	0.807 ± 0.079
perimeter	8.92 ± 1.08	6.3 ± 0.9	0.933 ± 0.000
area	6.29 ± 0.85	4.7 ± 0.9	0.962 ± 0.003

Table 1: VE_1^* separates true generative factors from composite concepts, while probe performance does not. The true generative factors (`x_position`, `y_position`, `shape_class`, etc) rank highest in VE_1^*, whereas composite concepts (`area`, `perimeter`) achieve high probe R^2 but low VE_1^*, suggesting they are predictable but not cleanly steerable. Sparsity tracks neither. Note that `shape_class` reports classification accuracy while all other concepts report R^2.

Probe R^2 is uniformly high across concepts, whereas VE_1^* varies substantially, from 6.29 for `area` to 17.48 for `x_position`. Sparsity, which measures how many latent dimensions the probe weight matrix uses, also fails to track VE_1^*. For example, `area` uses only 4.7 non-zero coefficients yet has the lowest VE_1^*. As another example, the `shape_class` has a VE_1^* score of 11.98 ranking third, but has the highest sparsity of 9.7. True generative factors such as (`x_position`, `y_position`, `shape_class`) rank highest in VE_1^*, while composite concepts (`area`, `perimeter`) achieve comparable probe R^2 but much lower VE_1^*, suggesting their influence is distributed across multiple latent directions.

5 Related Works

Probes are a standard tool for assessing whether representations encode semantic concepts (Alain and Bengio, 2018), but they measure readability rather than controllability. Recent work formalizes steerability as a distinct dimension of generative model performance, showing that a model's ability to produce high-quality outputs does not imply that users can reliably steer it toward desired outcomes (Vafa et al., 2025). Concept bottleneck generative models aim to make steerability inherent by constraining an internal layer to human-understandable concepts, either through training from scratch (Yuksekgonul et al., 2023) or via post-hoc adaptation (Kulkarni et al., 2025). A further approach is the Variational Hard Concept Bottleneck, which improves steerability by reducing concept leakage. (Martínez-García et al., 2026). Sparse autoencoders (SAEs) have been used to reveal steerable features in VLA models (Swann et al., 2026). Another paper combines concept bottleneck models with SAEs to address steerability limitations of traditional SAEs (Kulkarni et al., 2026a). Our work complements these approaches by providing a diagnostic that predicts when such steering is likely to succeed. This builds on geometric analyses of VAEs showing that diagonal posteriors induce a structured decoder Jacobian (Allen, 2024; Rhodes and Lee, 2021), enabling us to measure how a concept's gradient is distributed across independent generative directions. Together, these foundations position our metrics as a practical tool for assessing steerability prior to building concept-based models.

6 Conclusion

We introduced VE_1^*, a post-hoc metric that measures whether a concept's influence is concentrated along a single direction in latent space. A limitation is that we do not directly evaluate steerability via latent interventions, which we leave as future work alongside evaluation on natural image datasets and broader generative architectures.

References

google-deepmind/dsprites-dataset, 2026. URL https://github.com/google-deepmind/dsprites-dataset.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2018. URL http://arxiv.org/abs/1610.01644.

Carl Allen. Unpicking Data at the Seams: VAEs, Disentanglement and Independent Components, 2024. URL http://arxiv.org/abs/2410.22559.

Georgios Arvanitidis, Miguel González-Duque, Alison Pouplin, Dimitris Kalatzis, and Søren Hauberg. Pulling back information geometry, 2022. URL http://arxiv.org/abs/2106.05367.

Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE, 2018. URL http://arxiv.org/abs/1804.03599.

Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating Sources of Disentanglement in Variational Autoencoders, 2019. URL http://arxiv.org/abs/1802.04942.

Manfredo P. do Carmo. *Riemannian Geometry*. Birkhäuser, 1992.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Ian T. Jolliffe and Jorge Cadima. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016. doi: 10.1098/rsta.2015.0202.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, 2017. URL http://arxiv.org/abs/1412.6980.

Akshay Kulkarni, Ge Yan, Chung-En Sun, Tuomas Oikarinen, and Tsui-Wei Weng. Interpretable Generative Models through Post-hoc Concept Bottlenecks, 2025. URL thttp://arxiv.org/abs/2503.19377.

Akshay Kulkarni, Tsui-Wei Weng, Vivek Narayanaswamy, Shusen Liu, Wesam A. Sakla, and Kowshik Thopalli. Interpretable and steerable concept bottleneck sparse autoencoders, 2026a. URL https://arxiv.org/abs/2512.10805.

Akshay Kulkarni, Tsui-Wei Weng, Vivek Narayanaswamy, Shusen Liu, Wesam A. Sakla, and Kowshik Thopalli. Interpretable and steerable concept bottleneck sparse autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026b.

Abhishek Kumar and Ben Poole. On Implicit Regularization in β-VAEs, 2020. URL http://arxiv.org/abs/2002.00041.

James Lucas, George Tucker, Roger Grosse, and Mohammad Norouzi. Don't Blame the ELBO! A Linear VAE Perspective on Posterior Collapse, 2019. URL http://arxiv.org/abs/1911.02469.

María Martínez-García, Ricardo Vázquez Álvarez, Alejandro Lancho, Pablo M. Olmos, and Isabel Valera. A probabilistic hard concept bottleneck for steerable generative models. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=Kcb6WufAco.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep latent gaussian models. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

Travers Rhodes and Daniel D. Lee. Local Disentanglement in Variational Auto-Encoders Using Jacobian L_1 Regularization, 2021. URL http://arxiv.org/abs/2106.02923.

Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational Autoencoders Pursue PCA Directions (by Accident), 2019. URL http://arxiv.org/abs/1812.06775.

Hang Shao, Abhishek Kumar, and P. Thomas Fletcher. The Riemannian Geometry of Deep Generative Models, 2017. URL http://arxiv.org/abs/1711.08014.

Aiden Swann, Lachlain McGranahan, Hugo Buurmeijer, Monroe Kennedy III, and Mac Schwager. Sparse autoencoders reveal interpretable and steerable features in vla models, 2026. URL https://arxiv.org/abs/2603.19183.

Keyon Vafa, Sarah Bentley, Jon Kleinberg, and Sendhil Mullainathan. What's Producible May Not Be Reachable: Measuring the Steerability of Generative Models, 2025. URL http://arxiv.org/abs/2503.17482.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc Concept Bottleneck Models, 2023. URL http://arxiv.org/abs/2205.15480.

7 Appendix

7.1 Riemannian Manifolds Background (Informal)

We first motivate why we assume the decoder learns a manifold. Image data is very high dimensional: a colored 64×64 pixel image has dimensionality 12288. In contrast, the latent space is typically much smaller. In dSprites, for example, the true dimensionality is just 5, corresponding to the 5 generative factors (`x_position`, `y_position`, `shape_class`, etc.). The decoder f_θ therefore maps a low-dimensional latent space onto a low-dimensional surface, or *manifold*, embedded in this high-dimensional data space.

Now suppose we want to differentiate with respect to the latents, but we measure distances using the Euclidean metric of the high-dimensional data space. This is problematic because distances in the high-dimensional data space do not reflect distances along the manifold. To correct for this, we introduce the *Riemannian metric*, a symmetric positive definite correction term that accounts for how the latent space is embedded in the data space. For a decoder f_θ, this metric is concretely the pullback metric $G(\mathbf{z}) = J_\mathbf{z}^\intercal J_\mathbf{z}$, which pulls the Euclidean geometry of the data space back onto the latent space via the Jacobian. The reason it is called a pullback is that we are pulling back the geometry from the high-dimensional data space to the low-dimensional latent space.

This is also why computing the Riemannian gradient requires specifying $G(\mathbf{z})^{-1}$: by definition, the Riemannian gradient corrects the raw gradient by the inverse metric, accounting for the distortion introduced by the embedding. For formal definitions and further reading, see (Shao et al., 2017; Arvanitidis et al., 2022; do Carmo, 1992).

7.2 Verifying Approximate Orthogonality

To verify Condition 1, we define

$$\mathrm{C}(\mathbf{z}) = \frac{\|\text{off-diag}(J_\mathbf{z}^\intercal J_\mathbf{z})\|_F}{\|J_\mathbf{z}^\intercal J_\mathbf{z}\|_F} \tag{5}$$

where $\text{off-diag}(\cdot)$ retains only the off-diagonal entries. Small values indicate that $J_\mathbf{z}^\intercal J_\mathbf{z}$ is approximately diagonal; if not, the spread of gradients across latent dimensions can no longer be attributed to distinct generative factors.

7.3 Counter Example for Intuition

Counterexample (Classification). Let $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, I_2)$, $c(\mathbf{z}) = z_1 e^{\beta z_2}$ with $\beta > 0$ finite, and $\phi(\mathbf{z}) = \text{sign}(\mathbf{w}^\intercal \mathbf{z})$ be a linear classifier. The concept gradient is

$$g(\mathbf{z}) = \nabla_\mathbf{z}\, c(\mathbf{z}) = \begin{pmatrix} e^{\beta z_2} \\ \beta z_1 e^{\beta z_2} \end{pmatrix}. \tag{6}$$

We show that the probe attains perfect accuracy and sparsity for any finite $\beta > 0$, while $\mathrm{VE}_1 < 1$ correctly flags that c is not steerable along a single direction.

Proof. By Stein's lemma, since c is smooth and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, I_2)$, the optimal linear probe weights satisfy $\mathbf{w}^* = \mathbb{E}_{\mathbf{z}}[g(\mathbf{z})]$. Using independence of z_1, z_2 and the MGF ($\mathbb{E}[e^{\beta z_2}] = e^{\beta^2/2}$),

$$\mathbf{w}^* = \begin{pmatrix} \mathbb{E}[e^{\beta z_2}] \\ \beta \mathbb{E}[z_1]\mathbb{E}[e^{\beta z_2}] \end{pmatrix} = \begin{pmatrix} e^{\beta^2/2} \\ 0 \end{pmatrix}, \tag{7}$$

so $\|\mathbf{w}^*\|_0 = 1$. Since $e^{\beta z_2} > 0$ always, $\mathrm{sign}(c(\mathbf{z})) = \mathrm{sign}(z_1)$, and the probe predicts $\phi(\mathbf{z}) = \mathrm{sign}(w_1^* z_1) = \mathrm{sign}(z_1)$, attaining accuracy $= 1$.

Using independence and the MGF ($\mathbb{E}[e^{2\beta z_2}] = e^{2\beta^2}$),

$$\Sigma_g = \mathbb{E}[g(\mathbf{z})g(\mathbf{z})^\intercal] = \begin{pmatrix} \mathbb{E}[e^{2\beta z_2}] & \beta \mathbb{E}[z_1]\mathbb{E}[e^{2\beta z_2}] \\ \beta \mathbb{E}[z_1]\mathbb{E}[e^{2\beta z_2}] & \beta^2 \mathbb{E}[z_1^2]\mathbb{E}[e^{2\beta z_2}] \end{pmatrix} = e^{2\beta^2} \begin{pmatrix} 1 & 0 \\ 0 & \beta^2 \end{pmatrix}, \tag{8}$$

where the off-diagonal vanishes since $\mathbb{E}[z_1] = 0$. The eigenvalues are $e^{2\beta^2}$ and $\beta^2 e^{2\beta^2}$, giving

$$\mathrm{VE}_1 = \frac{\max(1, \beta^2)}{1 + \beta^2} < 1. \tag{9}$$

For instance, at $\beta = 1$, $\mathrm{VE}_1 = \frac{1}{2}$. Thus while probe accuracy and sparsity indicate perfect steerability, $\mathrm{VE}_1 < 1$ correctly identifies that no single latent direction controls c, and that steerability requires coordinated traversal of both z_1 and z_2. □

Remark: The covariance matrix Σ_g computed above is the uncentered second moment $\mathbb{E}[g(\mathbf{z})g(\mathbf{z})^\intercal]$, whereas VE_1 in practice uses the centered empirical covariance. Centering shifts the leading eigenvalue by $\|\mathbb{E}[g(\mathbf{z})]\|^2 = e^{\beta^2}$ but does not affect the conclusion that $\mathrm{VE}_1 < 1$.