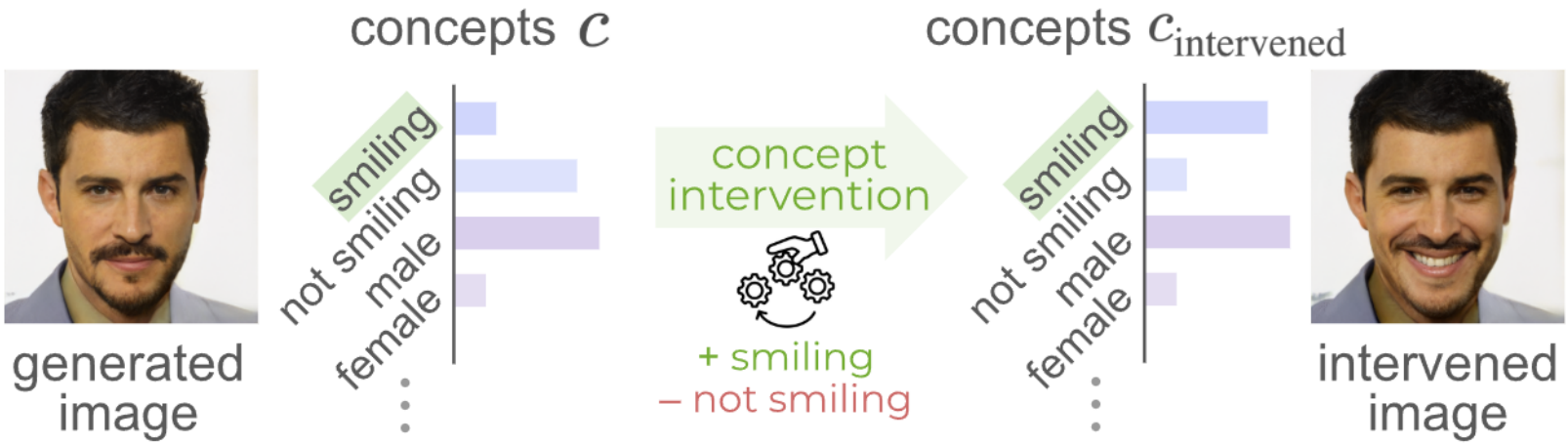
When Are Concepts Steerable?

Motivation

We focus on generative models where a common goal is to steer a specific concept in the output. Ex: in face image generation, one might want to control the "smile" attribute. However, steering concepts is known to be difficult (8; 1) for two reasons:

1. The target concept is hard to control directly, i.e., intervening on the latent supposed to encode it does not reliably change it.
2. Intervening on one concept unintentionally changes others. Ex: hair color sometimes changes during a smile intervention.



Notation

Let $x \in \mathbb{R}^n$ be a data point (e.g., an image with n pixels) and $z \in \mathbb{R}^d$ be the latent vector, where typically $d \ll n$. Let $f : \mathbb{R}^d \to \mathbb{R}^n$ be the generative decoder, and since f is differentiable, let $J_z = \frac{\partial x}{\partial z}$ denote its Jacobian with SVD $J_z = U\Sigma V^\top$.

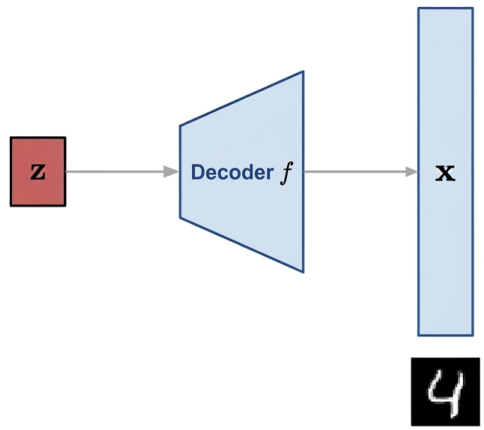


Figure: Notice that in the formulation, we actually ignore the encoder.

For a concept of interest, let $c : \mathbb{R}^n \to \mathbb{R}$ be a concept function that assigns a scalar score to a semantic attribute (discrete or continuous), and let $\nabla_x c$ denote the gradient of c with respect to the image.

Intuition (Informal)

Assuming J_z has orthogonal columns, we can project the concept gradient into latent space via the chain rule:

$$g(z) = \nabla_z c(f(z)) = J_z^\intercal \nabla_x c \tag{1}$$

If $g(z)$ aligns with a single singular vector of J_z, then the concept has a single latent axis to steer along and is therefore steerable.

(Formally, $g(z)$ is a proxy for the Riemannian gradient $G(z)^{-1}g(z)$, where $G(z) = J_z^\intercal J_z$ is the pullback metric. Under perfect column orthogonality, $G(z)$ is diagonal, so $G(z)^{-1}g(z)$ differs from $g(z)$ only by a coordinatewise scaling — preserving direction. Since perfect orthogonality is rare in practice, we additionally require the low-curvature assumption that $G(z)$ varies smoothly across the latent space (5; 6).)

From Concepts to Latent Directions

Idea: A concept induces a direction in data space; we project it into latent space via the Jacobian.

$$g(z) = J_z^\top \nabla_x c(x) = V \Sigma U^\top \nabla_x c(x) = \sum_i \lambda_i v_i \tag{2}$$

Interpretation:

- If $\nabla_x c(x)$ aligns with a single direction $\Rightarrow g(z) \approx \lambda_i v_i \Rightarrow$ one latent direction $\Rightarrow$ **steerable**
- If $\nabla_x c(x)$ spreads across multiple directions $\Rightarrow g(z) = \sum_i \lambda_i v_i \Rightarrow$ multiple latent directions $\Rightarrow$ **entangled, not steerable**

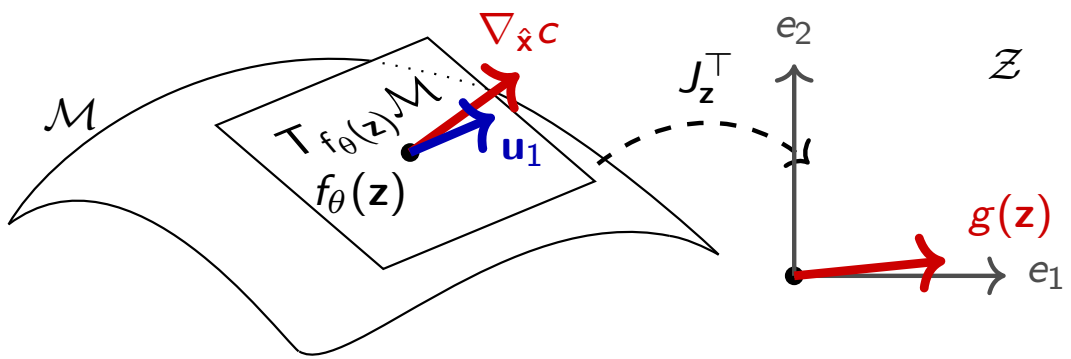


Figure: The concept gradient $\nabla_{\mathbf{x}} c$ is projected to the latent space via $J_{\mathbf{z}}^\top$ to latent space.

Measuring Alignment: VE_1

The goal is to quantify if $\{g(z)\}$ aligns along a single direction. We apply **PCA** on $\{g(z)\}$ and let $\lambda_1 \geq \lambda_2 \geq \ldots$ be its variance along principal directions.

$$VE_1 = \frac{\lambda_1}{\sum_j \lambda_j} \tag{3}$$

Interpretation:

- $VE_1 \approx 1 \quad \Rightarrow$ one dominant direction (clean control)
- $VE_1 \ll 1 \quad \Rightarrow$ concept controlled by multiple directions

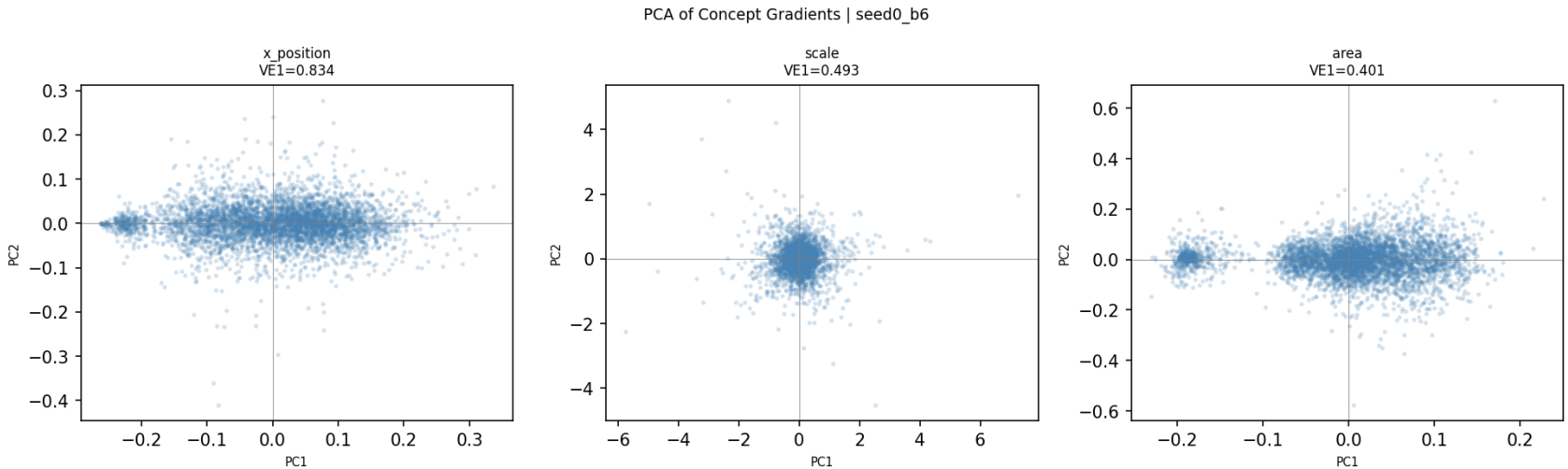


Figure: PCA of the concept gradient field $\{g(z_i)\}$.

dSprites

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: 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.

Condition

Our metric admits a precise geometric interpretation under approximate orthogonality. The decoder Jacobian has approximately orthogonal columns, i.e., $J_{\mathbf{z}}^{\top} J_{\mathbf{z}} \approx \mathrm{diag}(\cdot)$ (2; 3; 4). Ex: $\beta -$ TCVAE enforces orthogonality. In practice, it is not hard to verify empirically.

Latent Traversal

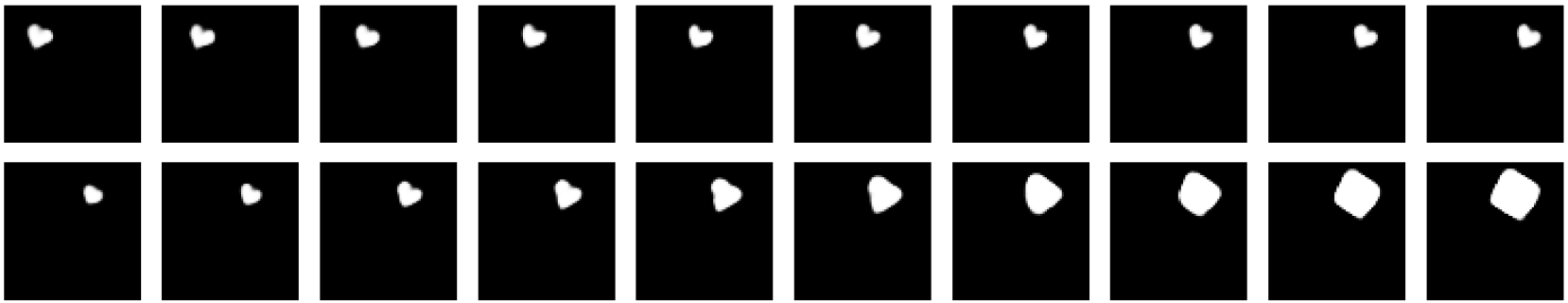


Figure: Top latent traversals show that position is cleanly steerable, while scale and shape are mildly entangled, which is intuitive since scaling an object inherently interacts with its shape.

Conclusion

We provide a practical tool for assessing whether a concept is steerable before committing to training a generative model from scratch. Although our metric is post-hoc, the large ecosystem of pre-trained generative models (e.g., `disentanglement_lib` alone provides $\sim$10,000 models) means practitioners can test steerability against an existing model that satisfies the required conditions.

References I

[1] Kulkarni et al. (2025).
http://arxiv.org/abs/2503.19377

[2] Rolinek et al. (2019).
http://arxiv.org/abs/1812.06775

[3] Lucas et al. (2019).
http://arxiv.org/abs/1911.02469

[4] Kumar & Poole (2020).
http://arxiv.org/abs/2002.00041

[5] Shao et al. (2017).
http://arxiv.org/abs/1711.08014

[6] Arvanitidis et al. (2022).
http://arxiv.org/abs/2106.05367

[7] Chen et al. (2019).
http://arxiv.org/abs/1802.04942

[8] Vafa et al. (2025).
http://arxiv.org/abs/2503.17482