# A Geometric View of Data Complexity:

## Efficient Local Intrinsic Dimension Estimation with Diffusion Models

**Hamidreza Kamkari**, Brendan Ross, Rasa Hosseinzadeh, Jesse Cresswell, Gabriel Loaiza-Ganem

Layer 6 AI, Toronto, Canada

**Spotlight** presentation

*arXiv:2406.03537*

NEURAL INFORMATION PROCESSING SYSTEMS

layer6

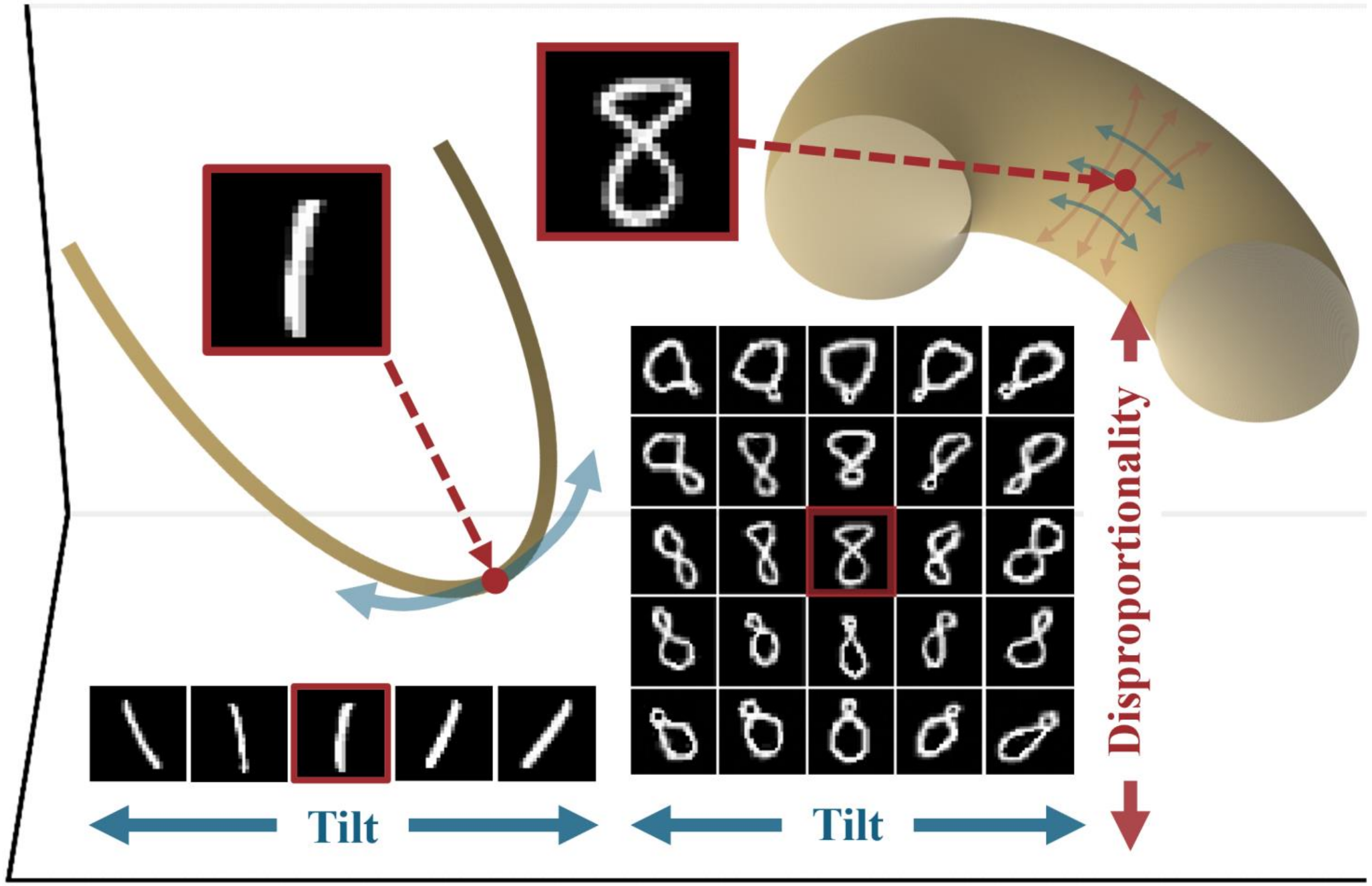
# The Manifold Hypothesis

*Data used in machine learning often lies on **low-dimensional** submanifolds of their ambient space* [1]

***(Definition)* Local Intrinsic Dimension** $\mathbf{LID}(\boldsymbol{x})$

*The dimension of the submanifold datapoint* $x$ *belongs to*

# Local Intrinsic Dimension: Relative Complexity

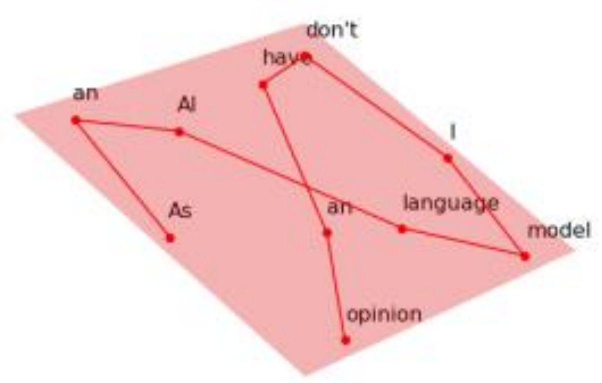




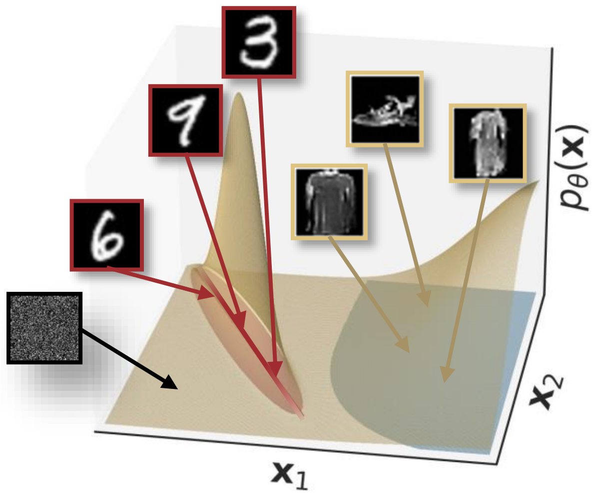

**Spotlight** presentation by: Hamid Kamkari

# Applications

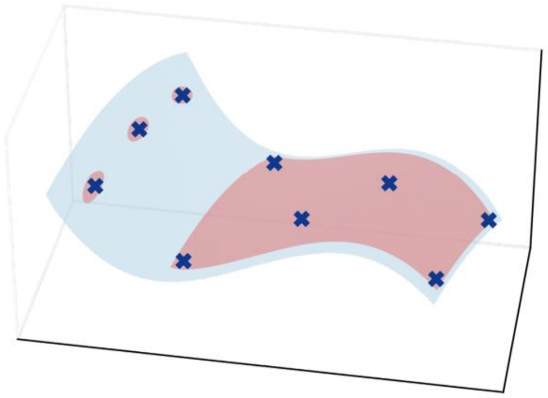
***LID** is useful for:*




***Detecting** *AI-generated* **content [3, 4]***




***Detecting** *out-of-distribution* **samples [2]***



***Detecting** *memorization* **[5]***

*Use a diffusion model trained on the data manifold, to extract the LID.*



**Spotlight** presentation by: Hamid Kamkari

# Fokker-Planck LID

*The forward process:* $t \in [0,1], \qquad \mathrm{d}X_t = \boldsymbol{f}(\boldsymbol{X_t}, \boldsymbol{t})\mathrm{d}t + \mathbf{g}(\boldsymbol{t})\mathrm{d}W_t$

*The transition kernel:* $\boldsymbol{p}_{X_t|X_0}(\boldsymbol{x_t}|\boldsymbol{x_0}) = \mathcal{N}(\boldsymbol{x_t}; \boldsymbol{\psi}(\boldsymbol{t})\ \boldsymbol{x_0}, \boldsymbol{\sigma}^2(\boldsymbol{t}))$

*Diffusion learns the score* [6]: $\hat{\boldsymbol{s}}_\theta(\cdot, \boldsymbol{t}) \approx \boldsymbol{s}(\cdot, \boldsymbol{t}) \coloneqq \nabla \log \boldsymbol{p}(\cdot, \boldsymbol{t})$

$$\widehat{\mathbf{LID}}(\boldsymbol{x}; \boldsymbol{\delta}) \coloneqq D + \boldsymbol{\sigma}^2(\boldsymbol{t}(\boldsymbol{\delta})) \left( \underbrace{\mathbf{tr}\, \nabla\, \hat{\boldsymbol{s}}_\theta \left( \boldsymbol{\psi}(\boldsymbol{t}(\boldsymbol{\delta}))\boldsymbol{x}, \boldsymbol{t}(\boldsymbol{\delta}) \right)} + \|\hat{\boldsymbol{s}}_\theta \left( \boldsymbol{\psi}(\boldsymbol{t}(\boldsymbol{\delta}))\boldsymbol{x}, \boldsymbol{t}(\boldsymbol{\delta}) \right)\|_2^2 \right)$$

$$\mathbf{LID}(\boldsymbol{x}) = \lim_{\boldsymbol{\delta} \to -\infty} \widehat{\mathbf{LID}}(\boldsymbol{x}; \boldsymbol{\delta})$$

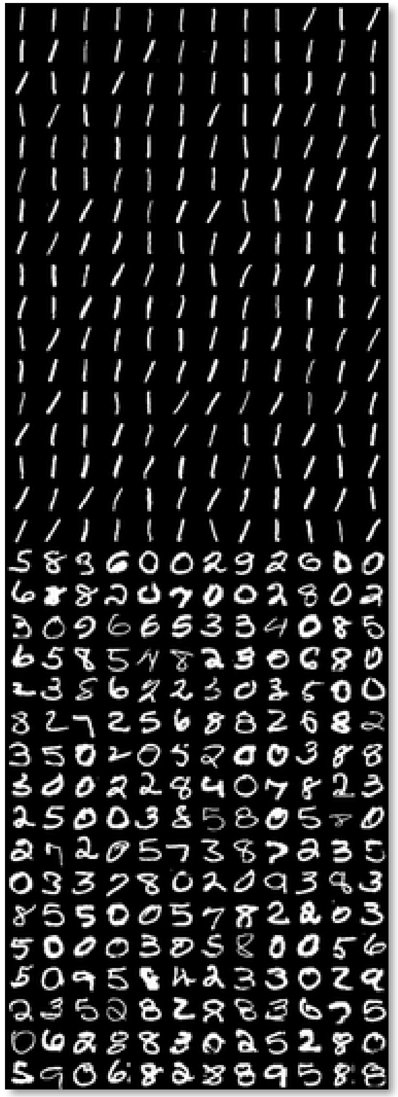
*Use **Hutchinson** with a few JVPs!*

# LID of LAION Images

Smallest LIDs

Largest LIDs

A ***single JVP*** on Stable Diffusion yields data complexity on **~1 million dimensional** images.




**Spotlight** presentation by: Hamid Kamkari

# LID for Synthetic Data

| | Model-based | | | | | | Model-free | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Synthetic Manifold | FLIPD | | NB | | LIDL | | ESS | | LPCA | |
| String within doughnut $\subseteq \mathbb{R}^3$ | **0.06** | 1.00 | 1.48 | 0.48 | 1.10 | 0.99 | 0.02 | 1.00 | **0.00** | 1.00 |
| $\mathcal{L}_5 \subseteq \mathbb{R}^{10}$ | 0.17 | - | 1.00 | - | **0.10** | - | 0.07 | - | **0.00** | - |
| $\mathcal{N}_{90} \subseteq \mathbb{R}^{100}$ | 0.49 | - | **0.18** | - | 0.33 | - | **1.67** | - | 21.9 | - |
| $\mathcal{U}_{10} + \mathcal{U}_{30} + \mathcal{U}_{90} \subseteq \mathbb{R}^{100}$ | **1.30** | 1.00 | 61.6 | 0.34 | 8.46 | 0.74 | 21.9 | 0.74 | **20.1** | **0.86** |
| $\mathcal{N}_{10} + \mathcal{N}_{25} + \mathcal{N}_{50} \subseteq \mathbb{R}^{100}$ | **1.81** | 1.00 | 74.2 | 0.34 | 8.87 | 0.74 | 7.71 | 0.88 | **5.72** | **0.91** |
| $\mathcal{F}_{10} + \mathcal{F}_{25} + \mathcal{F}_{50} \subseteq \mathbb{R}^{100}$ | **3.93** | 1.00 | 74.2 | 0.34 | 18.6 | 0.70 | 9.20 | 0.90 | **6.77** | 1.00 |
| $\mathcal{U}_{10} + \mathcal{U}_{80} + \mathcal{U}_{200} \subseteq \mathbb{R}^{800}$ | **14.3** | 1.00 | 715 | 0.34 | 120 | 0.70 | 1.39 | 1.00 | **0.01** | 1.00 |
| $\mathcal{U}_{900} \subseteq \mathbb{R}^{1000}$ | **12.8** | - | 100 | - | 24.9 | - | **14.5** | - | 219 | - |

# Summary

***Our LID estimator***

- ☑ *qualitatively extracts **data complexity** with a **pre-trained** model*
- ☑ *correlates strongly with **PNG Compression** size*
- ☑ *is theoretically grounded*
- ☑ *achieves **SOTA** on synthetic data*
- ☑ *is the first to **scale** to extremely high dimensions*
- ☑ *is **differentiable** by design.*

# References

**[1]** Bengio, Courville, Vincent. "**Representation learning: A review and new perspectives.**" IEEE Transactions on Pattern Analysis and Machine Intelligence, 35 (8):1798–1828, 2013.

**[2]** Kamkari, Ross, Cresswell, Caterini, Krishnan, Loaiza-Ganem. **"A geometric explanation of the likelihood OOD detection paradox."** International Conference on Machine Learning 2024

**[3]** Tulchinskii, Kuznetsov, Kushnareva, Cherni- avskii, Nikolenko, Burnaev, Barannikov, Piontkovskaya. **"Intrinsic dimension estimation for robust detection of AI-generated texts"**. Neural Information Processing Systems, 2023

**[4]** Ma, Li, Wang, Erfani, Wijewickrema, Schoenebeck, Song, Houle, Bailey. **"Characterizing adversarial subspaces using local intrinsic dimensionality."** In International Conference on Learning Representations, 2018.

**[5]** Ross, Kamkari, Wu, Hosseinzadeh, Liu, Stein, Cresswell, Loaiza-Ganem. "**A geometric framework for understanding memorization in generative models**." *arXiv preprint arXiv:2411.00113* (2024).

**[6]** Song, Sohl-Dickstein, Kingma, Kumar, Ermon, Poole. **Score-based generative modeling through stochastic differential equations**. In International Conference on Learning Representations, 2021.

# Thank You!

*Paper on* ***arXiv***

***Manifolds website***