Rethinking Intrinsic Dimension Estimation in Neural Representations

Rickmer Schulte
LMU Munich, MCML

David Rügamer
LMU Munich, MCML

Abstract

The analysis of neural representation has become an integral part of research aiming to better understand the inner workings of neural networks. While there are many different approaches to investigate neural representations, an important line of research has focused on doing so through the lens of intrinsic dimensions (IDs). Although this perspective has provided valuable insights and stimulated substantial follow-up research, important limitations of this approach have remained largely unaddressed. In this paper, we highlight a crucial discrepancy between theory and practice of IDs in neural representations, theoretically and empirically showing that common ID estimators are, in fact, not tracking the true underlying ID of the representation. We contrast this negative result with an investigation of the underlying factors that may drive commonly reported ID-related results on neural representation in the literature. Building on these insights, we offer a new perspective on ID estimation in neural representations.

1 INTRODUCTION

Intrinsic dimensions (IDs) play a central role in deep learning and have been the focus of research across a broad range of related studies. IDs are often encountered in the context of the so-called *manifold hypothesis* (Tenenbaum et al., 2000; Fefferman et al., 2016). The hypothesis postulates that many high-dimensional datasets frequently encountered in deep learning, such as image and text data, lie on or near a low-dimensional manifold despite being embedded in high-dimensional ambient spaces of dimension d, e.g., the number of pixels of an image. The hypothesis implies that a small number of dimensions $d_{\mathcal{M}} \ll d$ would theoretically suffice to fully characterize such datasets. This manifold dimension $d_{\mathcal{M}}$ is commonly referred to as the *intrinsic dimension*.

The manifold hypothesis has been explored both empirically and theoretically in numerous studies. Although the validity of the hypothesis remains debated, many researchers attribute at least part of the success of deep learning to this phenomenon. In other words, the fact that deep learning models are able to *learn* in the context of high-dimensional image and text data, and thereby escape the so-called *curse of dimensionality* (Bellman, 1961; Bishop, 2006; Bengio et al., 2013; Goodfellow et al., 2016), is said to be enabled by the presence of low IDs of the data that neural networks can adapt to (Chen et al., 2019; Schmidt-Hieber, 2019; Nakada and Imaizumi, 2020; Kohler et al., 2023; Schulte et al., 2025).

Related Literature In recent years, there has been a rise in research investigating deep neural networks through the lens of IDs. Besides investigating IDs of frequently encountered datasets in deep learning (Pope et al., 2021; Konz and Mazurowski, 2024a), researchers have also aimed to understand the inner workings of these models by examining IDs of neural representations from different layers of the neural network and found consistently occurring patterns over various models (Gong et al., 2019; Ansuini et al., 2019; Cai et al., 2021; Valeriani et al., 2023; Konz and Mazurowski, 2024b; Doimo et al., 2024; Aljaafari et al., 2025; Viswanathan et al., 2025; Cheng et al., 2025).

Problem Statement Investigating IDs of neural representations on various datasets and neural network architectures, ranging from vision to text-based models, all previous studies found ID estimates to vary over different network layers. Most strikingly, almost all of these studies find estimated IDs to increase in the early layers and to decrease in later layers (cf. Fig. 1). Such patterns are then often interpreted as the emergence of abstractions or phase transitions (Cheng et al., 2025).

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

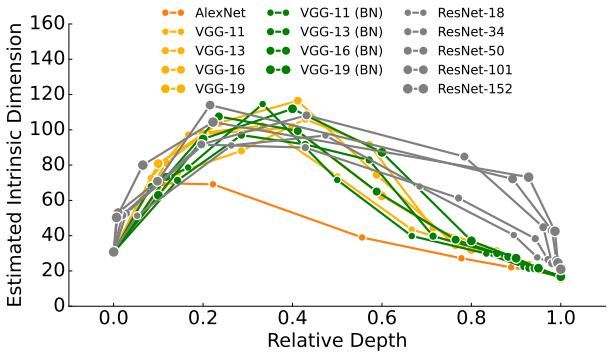

Figure 1: Layer-wise ID patterns for various architectures (adapted from Ansuini et al., 2019). ID estimates are based on the TwoNN estimator. To compare models with varying number of layers, results are depicted over relative depth.

While we do not question these previous empirical results, the question remains whether the observed phenomena can truly be attributed to the ID.

Our Contributions In this work, we investigate this research question and provide the following contributions:

1. We show that commonly used ID estimators are not only heavily biased in high dimensions (Section 2), but are provably not tracking the true underlying IDs of layer-wise neural representations (Section 3.1).
2. We further show that this result also holds in case the data does not lie on a single but a *union of manifolds* (Section 3.2).
3. Following this, we give a precise characterization of manifolds of LLM embeddings and hidden layer representations (Section 3.3).
4. Finally, we uncover driving forces behind layer-wise ID patterns and their connection to other metrics.

2 INTRINSIC DIMENSIONS & ESTIMATORS

As described in the previous section, the core idea of the manifold hypothesis is that many deep learning datasets, represented in high-dimensional ambient spaces of dimension d, may lie on or close to a low-dimensional manifold with ID $d_{\mathcal{M}} \ll d$.

Before diving into the analysis of IDs, we require a formal definition of IDs. While many different definitions are used in related literature, we will work with two of the most common IDs, the *Hausdorff dimension* and the *pointwise dimension*, and discuss corresponding estimators in the following. We provide a brief definition below, but refer the interested reader to the excellent surveys of Camastra and Staiano (2016) and Binnie et al. (2025) that discuss several details, including other notions of IDs and estimators.

Intrinsic Dimensions A common definition of ID is the *Hausdorff dimension* (Hausdorff, 1918), which generalizes the concept of dimension to arbitrary sets in metric spaces. A formal definition can be found in Appendix B.1. A very related notion of ID is the so-called *pointwise dimension* (Young, 1982). Given that it is locally defined for each point (instead of all points as in the Hausdorff dimension), it is also known as the *local Hausdorff dimension*. In case the lower and upper pointwise dimensions (formally defined in Appendix B.2) agree, the pointwise dimension defined at each point x can be written as

$$d_\mu(x) = \lim_{r \downarrow 0} \frac{\log \mu(B(x,r))}{\log r}\,, \tag{1}$$

where $B(x,r)$ corresponds to a ball with radius r that is centered around the point x and μ is a probability measure, such as Borel, or a measure supported on the set under study. In case $d_\mu(x)$ is μ-a.s. the same for all points x, the pointwise dimension is said to be *exact dimensional* and abbreviated by d_μ. Further details can be found in Appendix B.2.

ID Estimators and its Targets While there are generally many different ID estimators, we will focus on the ones that are most commonly applied in the analysis of neural representations, namely the Maximum Likelihood Estimator (MLE) (Levina and Bickel, 2004), TwoNN (Facco et al., 2017), and variants thereof. Both ID estimators build on the ratios of nearest neighbor (NN) distances. For example, the MLE at point x that considers k neighbors, is defined as

$$\widehat{d}_{\mathrm{MLE}}(x) \;=\; \left[\frac{1}{k-1}\sum_{i=1}^{k-1} \log \frac{T_k(x)}{T_i(x)}\right]^{-1}\,, \tag{2}$$

where $T_i(x)$ denotes the distance from x to its i-th nearest neighbor. The final MLE estimate is obtained by averaging over the estimates of all points x, usually using the approach described in MacKay and Ghahramani (2005). The TwoNN estimator is a special case of the MLE, using only the distances to the first two $k = 2$ nearest neighbors (NNs). Although these estimators may differ in finite samples, it is fundamental for our analysis that both target the same underlying notion of ID, namely, the pointwise dimension. A formal derivation of this connection can be found in Appendix B.4.

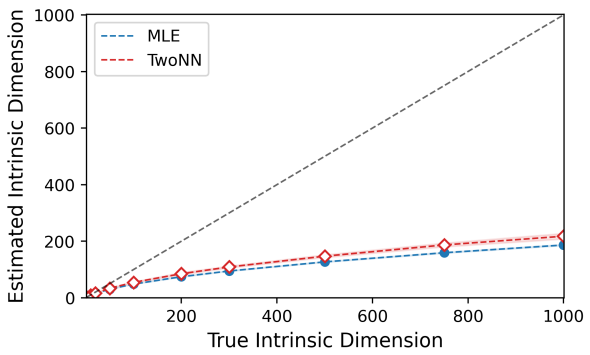

Figure 2: Estimated IDs using TwoNN and MLE vs. true ID of the manifold. Details are provided in Appendix C.

Bias of ID Estimators in High Dimensions Similar to other ID estimators, the MLE and TwoNN estimators are known to be sensitive to their underlying assumptions and to underestimate the true ID in high dimensions. Although the phenomenon of underestimation is well recognized in the literature (Camastra and Staiano, 2016; Binnie et al., 2025), it is usually demonstrated for relatively small numbers of dimensions, revealing relatively small biases compared to ID estimates (Levina and Bickel, 2004). However, we show that this bias grows drastically with increasing true ID (cf. Fig. 2). This is particularly concerning given that dimensions as those shown in Fig. 2 are highly relevant in modern deep learning. For example, the latest DINO embeddings increased from 1,535 in version 2 to 4,096 in version 3 (Siméoni et al., 2025).

Given this underestimation, related papers usually state that ID estimates should be treated as a *lower bound* of the actual ID (Ansuini et al., 2019). However, in order to provide meaningful insights into ID patterns of layer-wise representations of neural networks, the bias of such lower bound estimators must at least be consistent. If this is the case, *relative comparisons* of estimates, e.g., by looking at the estimator's *patterns*, would be rendered meaningful. However, as we will show in the next section, these estimated ID patterns also fail to track a (biased) version of the underlying ID and are therefore not at all indicative of it.

3 INTRINSIC DIMENSIONS OF NEURAL REPRESENTATIONS

In the following, we show that ID estimators do not estimate the true IDs of neural representations and can also not be considered indicative of those. We derive this result for the *pointwise dimension*. Analogous results for the *Hausdorff dimension* are deferred to the Appendix. Before stating our results, we briefly introduce some relevant notation.

Notation Unless otherwise stated, the ambient spaces considered are Euclidean $(\mathbb{R}^d, \|\cdot\|)$. For $f : \mathbb{R}^n \to \mathbb{R}^m$, we say f is *L-Lipschitz* if $\|f(x) - f(y)\| \leq L\|x - y\|$ for all $x, y \in \mathbb{R}^n$, and (L, α)-*Hölder*, $\alpha \in (0, 1]$, if $\|f(x) - f(y)\| \leq L\|x - y\|^\alpha$. We denote the Hausdorff dimension by $\dim_{\mathrm{H}}$. For a probability measure μ on $\mathbb{R}^n$, the pushforward by f is $\nu = f_{\#}\mu$, i.e. $\nu(\cdot) = \mu(f^{-1}(\cdot))$. For a feedforward neural network $(f_\ell)_{\ell=1}^L$ and an input law μ_0, define $\mu_\ell := (f_\ell)_{\#}\mu_{\ell-1}$.

3.1 Intrinsic Dimensions of Layer-wise Neural Representations are Non-increasing

Our main finding rests on the observation that almost all neural network architectures are a composition of layer-wise Lipschitz mappings, and that common notions of IDs cannot increase under Lipschitz mappings.

Neural Networks are Lipschitz Mappings The observation that most neural networks are Lipschitz mappings follows from the fact that they are usually a composition of the following Lipschitz components:

- Standard linear or convolutional layers (e.g., proven in Kim et al., 2021, Cor. 2.1);
- Pointwise activations such as ReLU (see, e.g., Tsuzuku et al., 2018) and Softmax (see, e.g., Gao and Pavel, 2017, Prop. 4);
- Pooling operators and residual additions (derived in, e.g., Tsuzuku et al., 2018; Béthune et al., 2022);
- Normalization layers such as BatchNorm and RMSNorm (e.g., Tsuzuku et al., 2018).

The conclusion that neural networks are Lipschitz mappings then just follows from the fact that compositions of Lipschitz mappings remain Lipschitz. We discuss the above results as well as the special case of self-attention in more detail in Appendix A.

Using this insight, we now show in the following lemma that the pointwise dimension **cannot** increase under Lipschitz mappings.

Lemma 1 (Pointwise dimensions under Lipschitz mappings)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be L-Lipschitz, μ a Borel probability measure on $\mathbb{R}^n$, and $\nu = f_{\#}\mu$. For μ-a.e. x with $y = f(x)$, we get for the upper $(\overline{d}_\nu)$ and lower $(\underline{d}_\nu)$ pointwise dimension that*

$$\overline{d}_\nu(y) \leq \overline{d}_\mu(x), \qquad \underline{d}_\nu(y) \leq \underline{d}_\mu(x).$$

In particular, if the pointwise dimension d_ν exists $(\overline{d}_\nu = \underline{d}_\nu)$, then

$$d_\nu(y) \leq d_\mu(x).$$

The corresponding proof can be found in Appendix B. A related result to Lemma 1, studying the pointwise dimension under linear maps, can be found in Hochman (2014, Lemma 4.5). Lemma 1 is more general in the sense that any linear map is Lipschitz, but not vice versa. In the following, we will omit a separate treatment of $\overline{d}_\nu$ and $\underline{d}_\nu$, assuming that d_μ exists.

Using the previous lemma, we can now show our first main result, namely the layer-wise monotonicity of the pointwise dimension.

Theorem 1 (Layer-wise monotonicity of pointwise dimensions)**.** *Let $f_1, \dots, f_L$ be Lipschitz maps and set $\mu_\ell = (f_\ell)_{\#}\mu_{\ell-1}$. Then for each $\ell \in \{1, \dots, L\}$ and $\mu_{\ell-1}$-a.e. x with $y = f_\ell(x)$, we have that*

$$d_{\mu_\ell}(y) \;\leq\; d_{\mu_{\ell-1}}(x).$$

In other words, the theorem states that pointwise dimension cannot increase over the layers of any Lipschitz neural network.

Remark 1. *In case every μ_ℓ is exact dimensional (i.e., $d_{\mu_\ell}(x)$ exists and equals a constant d_ℓ for μ_ℓ-a.e. x), then $d_\ell \leq d_{\ell-1}$ for all ℓ.*

Proof. Apply Lemma 1 to each layer. In the exact-dimensional case, the pointwise dimensions are a.e. constant, so the a.e. inequality becomes $d_\ell \leq d_{\ell-1}$. □

In Appendix B.1, we present an analogous result for the *Hausdorff dimension* (Lemma 2 and Theorem 2), namely that the Hausdorff dimension cannot increase under Lipschitz maps. These results have important implications for the ID estimation of neural representation as they uncover an important contradiction between theory and commonly found empirical ID-related results.

Remark 2. *A special case of Lemma 1 and 2 (Appendix) arises for bi-Lipschitz mappings. In that case, the results in the two lemmas hold with equality. However, as most neural networks cannot be guaranteed to be compositions of bi-Lipschitz maps, this result might only be of minor relevance in our context. We therefore deferred the discussion to Appendix B.3.*

A Contradiction Theorem 1 and Theorem 2 (Appendix) imply that the ID cannot increase over the layers of any Lipschitz neural network. This theoretical result stands in stark contrast to the increasing patterns of estimated IDs that are commonly observed across empirical studies investigating layer-wise ID in neural networks (e.g., ID patterns observed in Fig. 1). Given that the actual layer-wise ID cannot increase, the layer-wise ID patterns found by these studies cannot correspond to the true IDs of the neural representations, not even in a relative sense (the bias cannot be consistent). Hence, estimated neural ID patterns are not only strongly biased but also *not at all indicative* of the underlying IDs of neural representations.

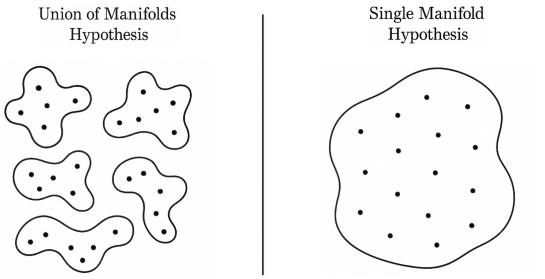

Figure 3: Union of Manifolds vs. Single Manifold Hypothesis: Dots illustrate samples living on multiple disconnected manifolds (left) vs. a single manifold (right).

3.2 Extensions to Union of Manifolds

So far, we have introduced the classic version of the manifold hypothesis, in which data is assumed to lie on a single low-dimensional manifold. However, the hypothesis can be generalized to consider data that lies on a union of disconnected manifolds. A schematic visualization of the single and union of manifolds hypotheses can be found in Fig. 3.

3.2.1 Single vs. Union of Manifolds Hypothesis

While the idea of a union of manifold hypothesis has been prominent in the clustering literature for some time (Vidal, 2011; Elhamifar and Vidal, 2011), Brown et al. (2023) more recently provided empirical evidence that many commonly used image datasets are more likely to live on a manifold union rather than a single manifold. Simplified, the key observation of their work is that images from the same classes (i.e., classes in the MNIST dataset) seem to share a common support, while images from different classes have disconnected supports. Moreover, as the union of manifold hypothesis allows images from different categories or classes to lie on disconnected manifolds, each of these class-specific manifolds may have a different ID. In line with this, Brown et al. (2023) find estimated IDs to vary between classes for several image datasets, providing further evidence for the union of manifold hypothesis.

3.2.2 IDs Under Unions of Manifolds

Given the empirical evidence for the union of manifolds hypothesis (at least for image datasets), we investigate how our results from Section 3.1 extend to this generalized version of the manifold hypothesis.

Pointwise Dimension Extending Lemma 1 to a union of manifolds is straightforward. Given that the dimension is only locally defined, the pointwise dimension is the same for points on the same manifold, but potentially different between points from disconnected manifolds. Lemma 1 can then be applied to each disconnected manifold of the union separately, yielding exactly the same conclusion as in Lemma 1 and Theorem 1. That is, the pointwise dimension cannot increase under Lipschitz mappings, even if the data lies on a union of manifolds.

Hausdorff Dimension Similarly, the result of Theorem 2 (Appendix) also extends to a union of manifolds, given that the considered set is a union of sets. However, as the Hausdorff dimension is defined globally for the entire set instead of locally for single points, the dimension can be made more precise in this case. For a (finite) union $Z = \bigcup_{i=1}^{p} M_i \subset \mathbb{R}^n$ of p manifolds M_i, each with Hausdorff dimension $\dim_{\mathrm{H}}(M_i) = d_{\mathcal{M}_i}$, we get that the Hausdorff dimension of Z is $\dim_{\mathrm{H}}(Z) = \max_i d_{\mathcal{M}_i}$ (Hochman, 2014, Prop. 2.12). Combining this with Lemma 2 (Appendix), for a union of manifolds Z under any Lipschitz map $f : \mathbb{R}^n \to \mathbb{R}^m$, we get that

$$\dim_{\mathrm{H}}(f(Z)) \leq \dim_{\mathrm{H}}(Z) = \max_i d_{\mathcal{M}_i}.$$

Class-specific ID Estimates The previous two paragraphs show that the result from Theorem 1 and Theorem 2 (Appendix) also hold under the union of manifold hypothesis. In addition, we can also verify that the mismatch between theory and practice of ID estimates remains. For this, we further refine the analysis of Fig. 1 by estimating the class-specific IDs of layer-wise representations, depicted in Fig. 4 for the ResNet-34 model. That is, we now estimate IDs of layer-wise representations for different classes of the ImageNet dataset separately. Results show that even class-specific ID estimates increase over the layers of the model. Hence, also these class-specific ID estimates cannot be considered indicative of the underlying ID of class-specific representations. Similar results for other models can be found in Appendix C.5.

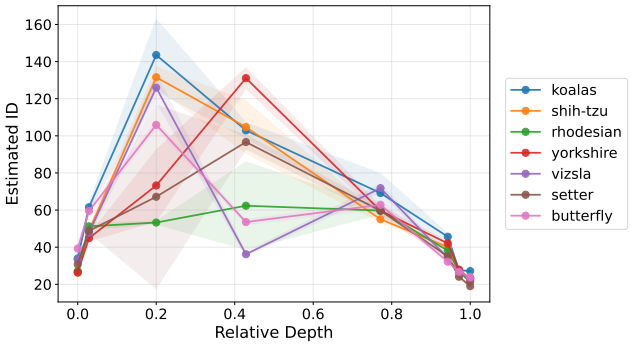

Figure 4: Estimated IDs of layer-wise representations from a ResNet-34 model for images with different categories (colors) from ImageNet. The x-axis shows the relative depth of model layers, and shaded areas show the estimated standard errors.

3.3 IDs & Manifolds of LLM Representations

The previous sections discussed ID estimation of neural representations in general, covering various types of model architectures. While the experiments mainly deal with vision models, the theoretical results derived in Section 3.1 also hold for other types of model architectures such as transformers. However, due to the increasing relevance and growing usage of large language models (LLMs) in recent years, we believe that providing a dedicated discussion and targeted results for LLMs is essential.

IDs in LLMs The first question that arises in estimating ID in LLMs is what notion of ID one aims to estimate. This is relatively straightforward in the case of vision models, as each image gets mapped to a corresponding layer-wise representation, and IDs are then estimated layer-wise over the entire image dataset. However, in the case of LLMs, datasets at inference time are usually prompts, each corresponding to a single or multiple sentences. Each token in a sentence then gets mapped to a specific representation in the embedding layer of the LLM (cf. Fig. 5). In an autoregressive decoder, the hidden state of the current layer at a position t is computed from the previous layer's states at positions $\leq t$. The hidden state at the final position of a prefix thus encodes information about the entire input seen so far, giving it a special role in the prediction of the next layer and output.

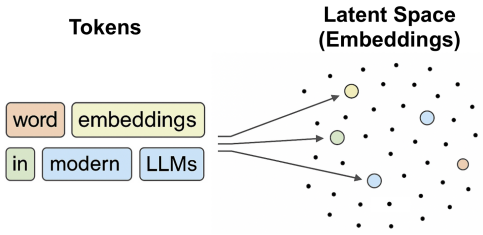

Figure 5: Visualization of word embeddings in LLMs: Each token in the prompt gets mapped to a unique point in the latent space of the embedding.

For this reason, the last hidden state can be used as a representative candidate in each layer for ID estimation (see, e.g., Cheng et al., 2025). Before discussing theoretical properties specific to LLMs, we will empirically analyze their ID patterns and relate findings to those of previous sections.

3.3.1 ID Patterns of LLMs

For ID estimation in LLMs, we use 10k prompts from the popular wikitext (Merity et al., 2017) dataset. For each prompt, we extract layer-wise representations from the pretrained Llama-3.1-8B (Grattafiori et al., 2024), Mistral-7B-v0.3 (Jiang et al., 2023), and Pythia-6.9B (Biderman et al., 2023). Further details can be found in Appendix C.

Gride ID estimation in LLMs is often based on the Gride ID estimator (Denti et al., 2022). It can be seen as a variant of the TwoNN and MLE estimator. The only difference is that Gride uses distance ratios of non-consecutive NN pairs (e.g. 1^{st} vs. 2^{nd}, 2^{nd} vs. 4^{th} etc.) rather than comparing only distances of the first nearest neighbors. Although the other two estimators could in principle also be used, Gride is sometimes preferred in the context of LLMs, as it can capture the local geometry beyond only the next neighbors, which can be advantageous in the high-dimensional representation spaces of LLMs. Important for our analysis is that Gride also targets the pointwise dimension analogous to TwoNN and MLE. Hence, the results from Section 3.1 also apply.

ID Patterns in LLMs The results of our LLM-based ID analysis are depicted in Fig. 6. The figure shows the averages (thick line) of six different scalings of Gride (thin lines). Each scaling compares one of the $2^{nd}/1^{st}, \ldots, 64^{th}/32^{nd}$ NN distance ratios. Dotted lines correspond to the first scaling, which is equivalent to the TwoNN estimator. Further details are provided in Appendix C. Estimated ID patterns are relatively consistent among the considered LLMs and exhibit an increase, especially in the early layers. As before, this is in violation of the theoretical results about layer-wise IDs in neural networks, raising the same contradiction as in Section 3.1.

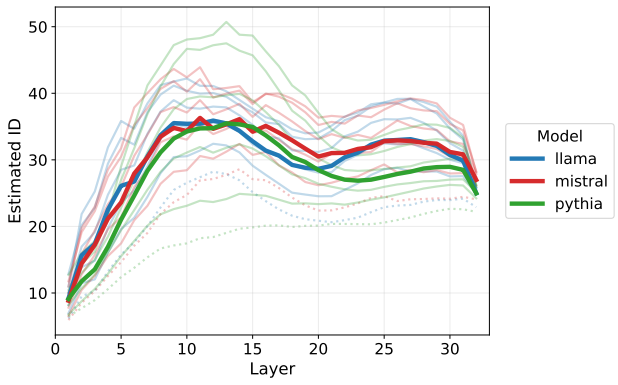

Figure 6: Estimated IDs of the layer-wise representations of various LLMs. Averaged ID estimates (thick lines) of six Gride scalings (thin lines), including the TwoNN (dotted).

3.3.2 Exact IDs of LLMs representation

As shown in Fig. 5, LLMs map each token in a sequence to a unique embedding. While the number of unique points in this latent space depends on the vocabulary size of the model, they are always *finite*. This has a simple but important implication. As ID definitions, such as the Hausdorff (or pointwise) dimension, are zero for point sets (or the distribution over those), the ID of such LLM embeddings must be zero. For finite sequences of tokens, this reasoning can be extended to hidden layers, given that they are continuous functions, which can map a finite point set only to a countable set of points. The latter has again a Hausdorff and pointwise dimension of zero. A formal result of this is given in Lemma 3 (Appendix B.5). We emphasize, however, that this result is specific to token-based inputs and therefore to LLMs. In Appendix B.5.1, we discuss why it does not transfer to image representations, and hence does not apply to vision models such as CNNs and vision transformers.

Remark 3. *While other concerns about the manifold hypothesis for LLM embeddings have been raised (Robinson et al., 2025), we believe that this simple yet insightful result above may be of interest on its own.*

4 FORCES BEHIND LAYER-WISE ID PATTERNS

In the previous section, we saw that increasing layer-wise ID patterns of neural representations are common to all discussed neural architectures. While the exact shape may differ between vision and language models (cf. Fig. 1 and Fig. 6), all exhibit increasing estimated IDs in the early layers. However, our theoretical investigation in Section 3.1 showed that ID cannot increase over layers of such neural networks, implying that their estimates do not track the underlying true ID.

Nonetheless, the consistency of ID patterns across pretrained models with different architectures suggests an important phenomenon that merits attention. We therefore investigate the forces underlying these layer-wise ID patterns, aiming to better understand what ID estimates may capture. While the experiments discussed here mainly focus on LLMs, analogous results for various vision models, including different CNNs and vision transformers (ViTs), are provided in Appendices C.4 to C.9.

4.1 NN Distances

Each of the considered ID estimators (MLE, TwoNN, Gride) is based on nearest-neighbor (NN) distances. Therefore, we investigate these distances in more detail in the following.

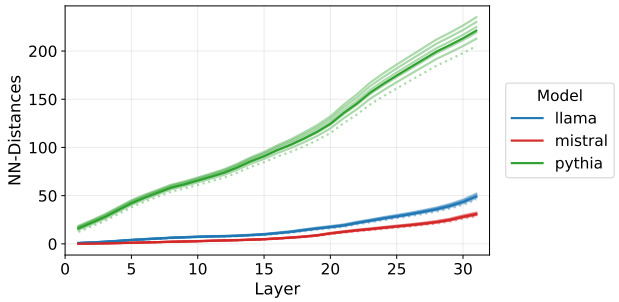

Figure 7: NN distances of layer-wise LLM representations (last layer excluded). For each model, each line (top to bottom) corresponds to the averages of 64^{th} & $32^{nd}, \ldots, 2^{nd}$ & 1^{st} NN distances. Solid lines denote the average over all 6 lines, with the TwoNN highlighted as a dotted line.

In Fig. 7, we plot the layer-wise NN distances that give rise to ID estimators in Fig. 6. Results clearly show that NN distances are growing over the layers of all models. However, the 1^{st} and 2^{nd} NN distances grow by a similar amount as the 32^{nd} and 64^{th} NN distances (at least in the early layers). This implies that the last hidden state representations in each layer in LLMs move away from each other in the latent space, creating similar distances between all representations.

This observation underlies the observed ID patterns. Considering again the construction of the ID estimators, each of them involves ratios of different pairs of NN distances. As the distances of all NNs grow by similar amounts (in early layers), their ratios shrink towards one. This is what drives estimated IDs to increase. In later layers, farther neighbors seem to grow slightly faster than closer NNs, which leads to a slight decrease in estimated IDs.

Remark 4. *We omit the last layer from Fig. 7 and most other plots in this section, since its representations are drastically altered by a LayerNorm transformation applied only at the final layer. As this effect is not central to our analysis, we exclude the last layer for readability and defer the full results to the Appendix.*

The following sections explore different factors that could drive the increasing separation of representations in latent space.

4.2 Ambient Space Dimension

A natural candidate to explain the increasing separation of representations is the *ambient space dimension*, i.e., the dimension of the representations in each layer. This dimension may be a driving factor because, in high-dimensional spaces, points become increasingly separated, causing the distance to the nearest neighbor to approach that to the farthest neighbor (Beyer et al., 1999; Aggarwal et al., 2001).

Although the described mechanism seems plausible, we find strong evidence against the ambient space dimension being the key driving factor. The reasoning for the language models is simple. For all considered language models, the size of hidden representations remains constant over all layers, in our case 4096. Hence, the ambient space dimension cannot drive observed ID patterns. While the layer-wise ambient dimension can change for the considered vision models, Fig. 13 in the Appendix also indicates that the layer-wise ambient space dimensions and ID-estimates seem rather unrelated.

4.3 Cosine Similarity

Another important factor could be the *cosine similarity*. The rationale behind this is that the last hidden states from different prompts should incorporate different contexts. Therefore, their representations may become (nearly) orthogonal over the layers of the neural network. However, as depicted in Fig. 8, the average cosine similarity between different last hidden state representations does not vary significantly over the model layers. While it is generally low (around 0.2) for the llama and mistral model, it is generally high (around 0.8) for the pythia model. Hence, cosine similarity does not seem to be a driving factor for increasing pairwise distances.

Remark 5. *In contrast to the above results, Viswanathan et al. (2025) find the pairwise cosine similarity to increase over the layers of llama. However, they consider the cosine similarity between representations of the same prompt. In this case, an increasing cosine similarity is expected, given that the representations are updated jointly and impacted by the same tokens.*

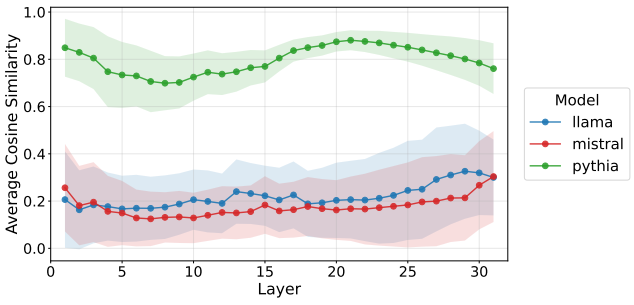

Figure 8: Average L2 similarity between layer-wise representations for llama, mistral, and pythia (last layer excluded). The shaded area band represents twice the standard error.

4.4 Size of Representations

While the previous two candidates cannot sufficiently explain increasing NN distances, we consider the size

of representations next. The size is naturally measured by the distance of each representation to the origin in latent space. As shown in Fig. 9, these distances grow across the hidden layers of the considered LLMs and exhibit patterns similar to the NN distances in Fig. 7. Similar layer-wise growth in the L_2 norm of hidden representations has also been observed across several other LLMs (Heimersheim and Turner, 2023; Gupta et al., 2024; Lawson et al., 2025).

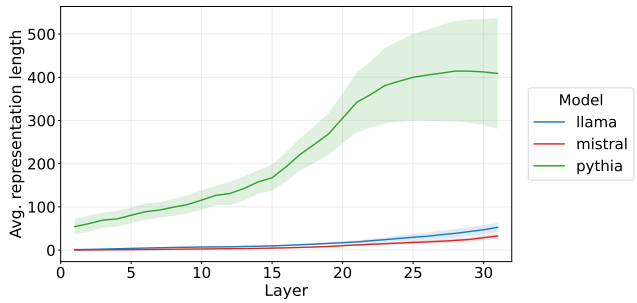

Figure 9: Average length of layer-wise representations for llama, mistral and pythia (last layer excluded). The shaded area band represents twice the standard error.

Expansion in Latent Space The layer-wise growth of the L_2 norms of last hidden state representations corresponds to an expansion in latent space over the hidden layers. We believe that there is an intuitive explanation for this phenomenon: For an accurate next-token prediction, LLMs need to enrich last-token representations with the specific context of each prompt, thereby distinguishing it from other prompts. LLMs may achieve this over their hidden layers by sequentially moving the last hidden state representations into different areas of their latent space. This yields the observed expansion.

We have seen that ID estimates are affected by layer-wise expansions in latent space. However, it remains unclear whether this expansion occurs uniformly or primarily along certain directions. To shed light on this question, we investigate how representations are distributed in latent space and how this distribution changes across hidden layers, using entropy-based metrics in the next section.

4.5 Entropy

Apart from IDs, various other metrics have been used to analyze the layer-wise geometry of neural representations. Recently, Skean et al. (2025) studied *entropy*-based metrics and also found distinct layer-wise patterns across different architectures. While they consider different entropy-based metrics, all of them essentially measure the entropy of the distribution of eigenvalues of layer-wise representations. We provide a formal definition of these metrics in Appendix C.2.

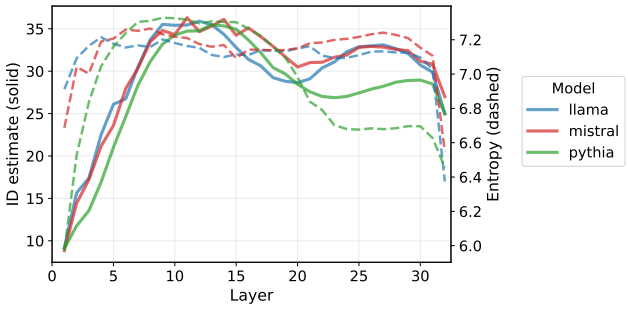

Figure 10: Estimated IDs (Gride) and entropy of layer-wise representation of various LLMs. Details in Appendix C.

Inspired by these findings, we study layer-wise von Neumann entropy estimates and and find that they follow a pattern strikingly similar to that of ID estimates (cf. Fig. 10), rising strongly in early layers and falling again in later ones. The strong connection between layer-wise ID and entropy patterns can also be found across various convolutional architectures and vision transformers (cf. App. Fig. 22 and Fig. 23).

Comparing IDs & Entropy Estimates Compared to ID estimators, layer-wise increases in entropy are well in line with theory and have a natural interpretation. As entropy estimates measure the spread of the eigenvalue distribution, they increase when representations are spread across many eigendirections and decrease when they are concentrated in a few. Hence, the patterns in Fig. 10 suggest that variance becomes more broadly distributed across linear directions in early layers and again more concentrated in later ones.

Small deviations between ID and entropy estimates might come from the information used. ID estimates only use local information from nearest neighbors, while entropy considers the entire distribution of all data points. Moreover, entropy captures the extent to which variance is distributed across linear directions in latent space. This not only gives estimators a clear interpretation, but is also in line with empirical evidence in modern LLMs, suggesting information is encoded linearly in the representations (Marks and Tegmark, 2024; Jiang et al., 2024; Park et al., 2024, 2025). The strong connection between the two metrics in Fig. 10 may therefore indicate that the observed ID patterns are driven by layer-wise changes in the spread of variance across *linear* directions in the latent space.

Theoretical Connection To the best of our knowledge, this is the first work to identify a striking resemblance between ID and entropy estimates for neural representations. While prior work has explored the theoretical connection between the two (Costa and Hero III, 2006; Bailey et al., 2021, 2022), it focuses on

entropy on the manifold, whereas our analysis concerns entropy in the ambient space.

5 CONCLUSION AND FUTURE OUTLOOK

5.1 New Perspective & Impact

A New Perspective on ID Estimates of Neural Representations Our theoretical results, in combination with our experiments, show that common ID estimates of neural representations do not recover the true IDs. However, the fact that the estimated layer-wise ID patterns are consistent across independently trained neural models with varying architectures (both for vision and text models) indicates that they capture an important geometric phenomenon inherent to neural representations more generally. Instead of interpreting these as IDs of layer-wise neural representation manifolds, our analysis suggests that it seems more appropriate to view them as reflecting a distinct geometric characteristic of layer-wise neural representations that can also be captured with entropy-based metrics. Namely, how variance is distributed across linear directions in latent space and how this distribution changes over the hidden layers.

Impact Studying the IDs of layer-wise neural representations, especially in the context of transformer-based models, has become increasingly popular in recent years. Studies have considered layer-wise ID estimation of neural representations in a variety of settings and across a broad range of models, including *decoder-only* models (Doimo et al., 2024; Viswanathan et al., 2025; Cheng et al., 2025), *encoder-decoder* models (Valeriani et al., 2023), and *vision models* including different convolutional architectures and ViTs (Ansuini et al., 2019; Kvinge et al., 2023; Wang et al., 2024; Konz and Mazurowski, 2024a; Shah and Yamins, 2025; Wang and Ma, 2025; Roschmann et al., 2025). Given that increasing ID patterns are commonly observed in these studies, our results suggest that such patterns should be interpreted with care. More broadly, they indicate that conclusions drawn from layer-wise ID estimation may need to be revisited across a wide range of contexts and model classes. Our analysis further suggests that complementing ID-based results with entropy-based measures may offer a more complete picture of the structure of neural representations.

5.2 Future Work

Entropy & ID While there might be many factors driving the found phenomena in neural representations, we provide empirical evidence that suggests ID estimates are connected to entropy. Based on current evidence, a deeper investigation and formalization of this connection is a promising direction for future research.

ID Estimators Another future avenue is the design of a new estimator that provides more reliable ID estimates of neural representations. While developing such an estimator is beyond the scope of the present work, we believe our theoretical results offer a useful insight for benchmarking future methods. In particular, layer-wise ID estimates should not increase across the layers of (Lipschitz) neural networks as a minimal requirement to be consistent with theory. This condition is necessary but not sufficient for reliability, since a non-increasing pattern can still differ from the true IDs, but it already rules out several commonly used estimators, as shown in our analysis. We believe this provides a valuable starting point for principled evaluation of ID estimators on neural representations.

References

Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420–434. Springer, 2001.

Nura Aljaafari, Danilo S Carvalho, and André Freitas. TRACE for Tracking the Emergence of Semantic Representations in Transformers. *arXiv preprint arXiv:2505.17998*, 2025.

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Jonathan Bac, Evgeny M Mirkes, Alexander N Gorban, Ivan Tyukin, and Andrei Zinovyev. Scikit-dimension: a python package for intrinsic dimension estimation. *Entropy*, 23(10):1368, 2021.

James Bailey, Michael E Houle, and Xingjun Ma. Relationships between local intrinsic dimensionality and tail entropy. In *International Conference on Similarity Search and Applications*, pages 186–200. Springer, 2021.

James Bailey, Michael E Houle, and Xingjun Ma. Local intrinsic dimensionality, entropy and statistical divergences. *Entropy*, 24(9):1220, 2022.

Richard E. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Princeton, 1961.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspec-

tives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Louis Béthune, Thibaut Boissin, Mathieu Serrurier, Franck Mamalet, Corentin Friedrich, and Alberto Gonzalez Sanz. Pay attention to your loss: understanding misconceptions about lipschitz neural networks. *Advances in Neural Information Processing Systems*, 35:20077–20091, 2022.

Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *International conference on database theory*, pages 217–235. Springer, 1999.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR, 2023.

James AD Binnie, John Harvey, Jakub Malinowski, Ka Man Yim, et al. A survey of dimension estimation methods. *arXiv preprint arXiv:2507.13887*, 2025.

Christopher M Bishop. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Bradley C. A. Brown, Anthony L. Caterini, Brendan Leigh Ross, Jesse C. Cresswell, and Gabriel Loaiza-Ganem. Verifying the Union of Manifolds Hypothesis for Image Data. In *International Conference on Learning Representations*, 2023.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations*, 2021.

Francesco Camastra and Antonino Staiano. Intrinsic dimension estimation: Advances and open problems. *Information Sciences*, 328:26–41, 2016.

Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Efficient approximation of deep relu networks for functions on low dimensional manifolds. *Advances in Neural Information Processing Systems*, 32, 2019.

Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Lei Yu, Alessandro Laio, and Marco Baroni. Emergence of a High-Dimensional Abstraction Phase in Language Transformers. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jose A Costa and Alfred O Hero III. Determining intrinsic dimension and entropy of high-dimensional shape spaces. In *Statistics and analysis of shapes*, pages 231–252. Springer, 2006.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Francesco Denti, Diego Doimo, Alessandro Laio, and Antonietta Mira. The generalized ratios intrinsic dimension estimator. *Scientific Reports*, 12(1):20005, 2022.

Diego Doimo, Alessandro Serra, Alessio Ansuini, and Alberto Cazzaniga. The representation landscape of few-shot learning and fine-tuning in large language models. *Advances in Neural Information Processing Systems*, 37:18122–18165, 2024.

Ehsan Elhamifar and René Vidal. Sparse manifold clustering and embedding. *Advances in Neural Information Processing Systems*, 24, 2011.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):12140, 2017.

Kenneth Falconer. *Fractal geometry: mathematical foundations and applications*. John Wiley & Sons, 2013.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4): 983–1049, 2016.

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

Luis Gonzalo Sanchez Giraldo, Murali Rao, and Jose C Principe. Measures of entropy from data using infinitely divisible kernels. *IEEE Transactions on Information Theory*, 61(1):535–548, 2014.

Aldo Glielmo, Iuri Macocco, Diego Doimo, Matteo Carli, Claudio Zeni, Romina Wild, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. DADApy: Distance-based analysis of data-manifolds in Python. *Patterns*, page 100589, 2022. ISSN 2666-3899.

Sixue Gong, Vishnu Naresh Boddeti, and Anil K Jain. On the intrinsic dimensionality of image representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3987–3996, 2019.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*, volume 1. MIT press Cambridge, 2016.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Akshat Gupta, Atahan Ozdemir, and Gopala Anumanchipalli. Geometric Interpretation of Layer Normalization and a Comparative Analysis with RMSNorm. *arXiv preprint arXiv:2409.12951*, 2024.

Felix Hausdorff. Dimension und äußeres Maß. *Mathematische Annalen*, 79(1):157–179, 1918.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Stefan Heimersheim and Alex Turner. Residual stream norms grow exponentially over the forward pass. In *AI Alignment Forum*, page 23, 2023.

Shohei Hidaka and Neeraj Kashyap. On the estimation of pointwise dimension. *arXiv preprint arXiv:1312.2298*, 2013.

Michael Hochman. Lectures on dynamics, fractal geometry, and metric number theory. *J. Mod. Dyn*, 8 (3-4):437–497, 2014.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, 2023.

Yibo Jiang, Goutham Rajendran, Pradeep Kumar Ravikumar, Bryon Aragam, and Victor Veitch. On the Origins of Linear Representations in Large Language Models. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 21879–21911. PMLR, 2024.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention. In *International Conference on Machine Learning*, pages 5562–5571. PMLR, 2021.

Michael Kohler, Sophie Langer, and Ulrich Reif. Estimation of a regression function on a manifold by fully connected deep neural networks. *Journal of Statistical Planning and Inference*, 222:160–181, 2023. ISSN 0378-3758. doi: https://doi.org/10.1016/j.jspi.2022.05.008.

Nicholas Konz and Maciej A Mazurowski. The effect of intrinsic dataset properties on generalization: Unraveling learning differences between natural and medical images. In *International Conference on Learning Representations*, 2024a.

Nicholas Konz and Maciej A Mazurowski. Preprocessing and Compression: Understanding Hidden Representation Refinement Across Imaging Domains via Intrinsic Dimension. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*, 2024b.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

Henry Kvinge, Grayson Jorgenson, Davis Brown, Charles Godfrey, and Tegan Emerson. Internal Representations of Vision Models Through the Lens of Frames on Data Manifolds. In *NeurIPS 2023 Workshop on Symmetry and Geometry in Neural Representations*, 2023.

Tim Lawson, Lucy Farnik, Conor Houghton, and Laurence Aitchison. Residual Stream Analysis with Multi-Layer SAEs. In *The Thirteenth International Conference on Learning Representations*, 2025.

Elizaveta Levina and Peter Bickel. Maximum likelihood estimation of intrinsic dimension. *Advances in Neural Information Processing Systems*, 17, 2004.

David J.C. MacKay and Zoubin Ghahramani. Comments on 'Maximum Likelihood Estimation of Intrinsic Dimension' by E. Levina and P. Bickel (2004). 2005.

Samuel Marks and Max Tegmark. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. In *First Conference on Language Modeling*, 2024.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer Sentinel Mixture Models. In *International Conference on Learning Representations*, 2017.

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174):1–38, 2020.

Kiho Park, Yo Joong Choe, and Victor Veitch. The Linear Representation Hypothesis and the Geometry of Large Language Models. In *International Conference on Machine Learning*, 2024.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The Geometry of Categorical and Hierarchical Concepts in Large Language Models. In *International Conference on Learning Representations*, 2025.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

Mathew D. Penrose and J. E. Yukich. Limit theory for point processes in manifolds. *The Annals of Applied Probability*, 23(6):2161–2211, 2013.

Mathew D Penrose and Joseph E Yukich. Weak laws of large numbers in geometric probability. *The Annals of Applied Probability*, 13(1):277–303, 2003.

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*, 2021.

Michael Robinson, Sourya Dey, and Tony Chiang. Token Embeddings Violate the Manifold Hypothesis. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Simon Roschmann, Quentin Bouniot, and Zeynep Akata. Time Series Representations for Classification Lie Hidden in Pretrained Vision Transformers. In *Recent Advances in Time Series Foundation Models Have We Reached the 'BERT Moment'?*, 2025.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pages 606–610. IEEE, 2007.

Johannes Schmidt-Hieber. Deep ReLu network approximation of functions on a manifold. *arXiv preprint arXiv:1908.00695*, 2019.

Rickmer Schulte, David Rügamer, and Thomas Nagler. Adjustment for confounding using pre-trained representations. In *Forty-second International Conference on Machine Learning*, 2025.

Yash Shah and Daniel LK Yamins. Topographic Vision Transformers. 2025.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by Layer: Uncovering Hidden Representations in Language Models. In *Forty-second International Conference on Machine Learning*, 2025.

Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. *Advances in Neural Information Processing Systems*, 36:51234–51252, 2023.

René Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2011.

Karthik Viswanathan, Yuri Gardinazzi, Giada Panerai, Alberto Cazzaniga, and Matteo Biagetti. The geometry of tokens in internal representations of large language models. *arXiv preprint arXiv:2501.10573*, 2025.

Hanzhang Wang and Qingyuan Ma. Textural or Textual: How Vision-Language Models Read Text in Images. In *Forty-second International Conference on Machine Learning*, 2025.

Hanzhang Wang, Jiawen Zhang, and Qingyuan Ma. Exploring Intrinsic Dimension for Vision-Language Model Pruning. In *Forty-first International Conference on Machine Learning*, 2024.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual Transformers: Token-based Image Representation and Processing for Computer Vision. *arXiv preprint arXiv:2006.03677*, 2020.

Lai-Sang Young. Dimension, entropy and lyapunov exponents. *Ergodic theory and dynamical systems*, 2 (1):109–124, 1982.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] Clear descriptions are provided both in the main text and the Supplementary Material.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] The code is made available in the Supplementary Material.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] A discussion of the full set of assumptions is provided along each theoretical result.
   (b) Complete proofs of all theoretical results. [Yes] Complete proofs are provided for all our theoretical results.
   (c) Clear explanations of any assumptions. [Yes] We aimed to clearly explain all the mentioned assumptions.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] Code, data, and instructions reproduce all experimental results.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] Details are provided in the Supplementary Material.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] Clear definitions are provided.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] Details are provided in the Supplementary Material.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials: Rethinking Intrinsic Dimension Estimation in Neural Representations

A Neural Networks are Lipschitz Mappings

In this section, we discuss in more detail the Lipschitz assumption used in our main results. We consider deterministic neural network layers at inference time and assume all weights and scalars are finite. Then, the following holds:

- Standard linear or convolutional layers are Lipschitz mappings (Kim et al., 2021, Cor. 2.1);
- Pointwise activations such as ReLU, leaky-ReLU, tanh, sigmoid, softplus, GELU / SiLU are Lipschitz (Tsuzuku et al., 2018);
- Pooling operators and residual additions preserve Lipschitzness of the composition (Tsuzuku et al., 2018; Béthune et al., 2022);
- Softmax is Lipschitz on $\mathbb{R}^d$ (Gao and Pavel, 2017, Prop. 4);
- Normalization layers such as LayerNorm, BatchNorm, and RMSNorm are Lipschitz (Tsuzuku et al., 2018).

Given that most neural networks are compositions of Lipschitz mappings (compositions of the above components), and given that compositions of Lipschitz mappings remain Lipschitz, Theorem 2 and Theorem 1 apply for such networks.

Not Globally Lipschitz Mappings Operations that are discontinuous or not globally Lipschitz are, for example, hard quantization or sign/argmax/top-k gating. Another subtle exception is self-attention. While there are Lipschitz variants such as L_2 self-attention, standard (scaled) dot-product self-attention is not globally Lipschitz on unbounded domains (Kim et al., 2021). Clearly, the same holds for multi-head attention given that it is just a linear map of single self-attention outputs. Nevertheless, if the input space is compact (e.g. for bounded inputs), self-attention is Lipschitz on that set (Kim et al., 2021). Hence, in this and the other special cases, one could alternatively state the results of Theorem 2 and Theorem 1 on a compact subset of the data domain on which each layer is Lipschitz. The conclusions then hold relative to that subset.

B Omitted Proofs and Derivations

B.1 Hausdorff Dimension

Definition 1 (Hausdorff measure and dimension (Hausdorff, 1918))**.** *Let $s \geq 0$ and $\delta > 0$ and $E \subset \mathbb{R}^d$. Define*

$$\mathcal{H}^s_\delta(E) \;:=\; \inf\Big\{ \sum_{i=1}^{\infty} (\operatorname{diam} U_i)^s : \; E \subset \bigcup_{i=1}^{\infty} U_i, \; \operatorname{diam} U_i \leq \delta \Big\},$$

where the infimum is considered with respect to all countable δ-covers $\{U_i\}$ of E, and $\operatorname{diam} U := \sup\{\|x - y\| : x, y \in U\}$. Further, the s-dimensional Hausdorff measure is defined by

$$\mathcal{H}^s(E) \;:=\; \lim_{\delta \downarrow 0} \mathcal{H}^s_\delta(E) \;=\; \sup_{\delta > 0} \mathcal{H}^s_\delta(E).$$

The Hausdorff dimension of the set E is then defined by

$$\dim_H(E) \;:=\; \inf\{s : \; \mathcal{H}^s(E) = 0\} \;=\; \sup\{s : \; \mathcal{H}^s(E) = \infty\}.$$

B.1.1 Monotonicity of the Hausdorff dimension under Lipschitz Maps

The following Lemma 2 is a classic result from fractal geometry and demonstrates that the Hausdorff dimension cannot grow under Lipschitz mappings (see, e.g., Falconer, 2013, Prop. 2.3 & Cor. 2.4 (a)). We provide a proof for completeness. We start by showing the result for the more general case of Hölder smooth maps (f is (L,α)-*Hölder*, $\alpha \in (0,1]$, if $\|f(x)-f(y)\| \leq L\|x-y\|^\alpha$) and follow the result for Lipschitz maps ($\alpha = 1$) from it.

Lemma 2 (Hausdorff dimension under Hölder/Lipschitz mappings)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be (L,α)-Hölder and $E \subset \mathbb{R}^n$. Then*

$$\dim_{\mathrm{H}}\big(f(E)\big) \;\leq\; \frac{1}{\alpha}\,\dim_{\mathrm{H}}(E).$$

In particular, if f is Lipschitz ($\alpha = 1$), then

$$\dim_{\mathrm{H}}(f(E)) \leq \dim_{\mathrm{H}}(E).$$

Proof. Fix $s > \dim_H(E)$. Then by definition $\mathcal{H}^s(E) = 0$, hence for each $\eta > 0$ there exists $\delta > 0$ with $\mathcal{H}^s_\delta(E) < \eta$. This means there is a cover $E \subset \bigcup_i U_i$ with $\operatorname{diam}(U_i) \leq \delta$ and $\sum_i (\operatorname{diam} U_i)^s < \eta$. Set $t > s/\alpha$ and fix an arbitrary $\delta' > 0$. Then choose $\delta \leq (\delta'/L)^{1/\alpha}$. For the cover above, we then have by Hölder continuity, $\operatorname{diam}\big(f(U_i)\big) \;\leq\; L\,\operatorname{diam}(U_i)^\alpha \;\leq\; \delta'$, so $\{f(U_i)\}_i$ is a δ'–cover of $f(E)$. Hence

$$\mathcal{H}^t_{\delta'}\big(f(E)\big) \;\leq\; \sum_i \big(\operatorname{diam} f(U_i)\big)^t \;\leq\; L^t \sum_i \big(\operatorname{diam} U_i\big)^{\alpha t} \;\leq\; L^t \sum_i \big(\operatorname{diam} U_i\big)^s \;<\; L^t\,\eta,$$

where we used $\alpha t > s$ in the second-to-last inequality. Since $\eta > 0$ was arbitrary, $\mathcal{H}^t_{\delta'}(f(E)) = 0$ for every $\delta' > 0$, and therefore $\mathcal{H}^t\big(f(E)\big) \;=\; \lim_{\delta' \downarrow 0} \mathcal{H}^t_{\delta'}\big(f(E)\big) \;=\; 0$. As this holds for all $t > s/\alpha$, we obtain $\dim_H(f(E)) \leq s/\alpha$. Letting $s \downarrow \dim_{\mathrm{H}}(E)$ concludes the proof. □

A similar result for the so-called *Minkowski dimension* can be found in Hochman (2014, Prop. 2.4). Lemma 2 can be used to show the layer-wise monotonicity of the Hausdorff dimension. Hence, $\dim_{\mathrm{H}}$ cannot increase over the layers of any Lipschitz neural network.

Theorem 2 (Layer-wise monotonicity of Hausdorff dimension)**.** *Let $f_1, \ldots, f_L$ be Lipschitz maps and set $\mu_\ell = (f_\ell)_{\#}\mu_{\ell-1}$. Then, for each $\ell \in \{1,\ldots,L\}$,*

$$\dim_{\mathrm{H}}\big(\operatorname{supp}\mu_\ell\big) \;\leq\; \dim_{\mathrm{H}}\big(\operatorname{supp}\mu_{\ell-1}\big).$$

If f_ℓ is only (L_ℓ,α_ℓ)-Hölder, then

$$\dim_{\mathrm{H}}(\operatorname{supp}\mu_\ell) \leq \alpha_\ell^{-1}\,\dim_{\mathrm{H}}(\operatorname{supp}\mu_{\ell-1}).$$

Proof. Consider Lemma 2 with $E = \operatorname{supp}\mu_{\ell-1}$ and $f = f_\ell$, and note $f_\ell(\operatorname{supp}\mu_{\ell-1}) = \operatorname{supp}\mu_\ell$. □

B.2 Pointwise Dimension

We begin by defining the concept of *pointwise dimension* that was first introduced by Young (1982) more formally. For a measure μ, its *upper* and *lower pointwise dimension* at point x are

$$\overline{d}_\mu(x) = \limsup_{r\downarrow 0} \frac{\log \mu(B(x,r))}{\log r}, \quad \underline{d}_\mu(x) = \liminf_{r\downarrow 0} \frac{\log \mu(B(x,r))}{\log r},$$

where $B(x,r)$ corresponds to a ball with radius r that is centered around the point x. When the upper and lower limit agree, i.e. $\overline{d}_\mu(x) = \underline{d}_\mu(x)$, it is also called *pointwise* (or *local Hausdorff*) *dimension* and is denoted by $d_\mu(x)$. Note that the pointwise dimension is defined for a single point instead of the entire dataset, which is why it is sometimes considered a local instead of a global dimension. Nonetheless, it can be considered at multiple points x. In particular, $d_\mu(x)$ is said to be *exact dimensional* in case it exists and is μ-a.s. independent of the point x (hence, $d_\mu(x)$ equals the same constant for all x μ-a.s.). In this case, it is sometimes denoted by d_μ (Hochman, 2014, Def. 3.9).

B.2.1 Proof of Lemma 1

Proof. For small $\rho > 0$, (L, α)-Hölder gives $f(B(x,\rho)) \subseteq B(y, L\rho^\alpha)$. Setting $r = L\rho^\alpha$,

$$\nu(B(y,r)) = \mu\big(f^{-1}(B(y,r))\big) \;\geq\; \mu\big(B(x,\rho)\big).$$

Taking logs and dividing by $\log r < 0$ (fulfilled for a sufficiently small r) reverses the inequality:

$$\frac{\log \nu(B(y,r))}{\log r} \;\leq\; \frac{\log \mu(B(x,\rho))}{\log r} = \frac{\log \mu(B(x,\rho))}{\alpha \log \rho + \log L}.$$

As $r \downarrow 0$, we have $\rho \downarrow 0$ and $\log \rho \to -\infty$, so the additive constant $\log L$ is negligible in lim sup/lim inf. Hence, for the upper pointwise dimension (same logic applies to the lower pointwise dimension), we get

$$\overline{d}_\nu(y) = \limsup_{r\downarrow 0} \frac{\log \nu(B(y,r))}{\log r} \;\leq\; \limsup_{\rho\downarrow 0} \frac{\log \mu(B(x,\rho))}{\alpha \log \rho + \log L} = \frac{1}{\alpha} \limsup_{\rho\downarrow 0} \frac{\log \mu(B(x,\rho))}{\log \rho} = \frac{1}{\alpha}\, \overline{d}_\mu(x).$$

For Lipschitz maps ($\alpha = 1$), we obtain $\overline{d}_\nu(y) \leq \overline{d}_\mu(x)$ and $\underline{d}_\nu(y) \leq \underline{d}_\mu(x)$, which concludes the proof. □

B.3 Invariance of Hausdorff and Pointwise Dimensions Under Bi-Lipschitz Mappings

The first part of the following result is another classic result from fractal geometry that can be found in Falconer (2013, Cor. 2.4 (b)). The second part that is about the (upper/lower) pointwise dimension is discussed in Hidaka and Kashyap (2013).

Proposition 1 (Invariance under Bi-Lipschitz mappings)**.** *If f is bi-Lipschitz on $E \subset \mathbb{R}^n$, then* $\dim_{\mathrm{H}}(f(E)) = \dim_{\mathrm{H}}(E)$*. If moreover $\nu = f_{\#}\mu$ and f is bi-Lipschitz on* $\operatorname{supp}\mu$*, then $\overline{d}_\nu(f(x)) = \overline{d}_\mu(x)$ and $\underline{d}_\nu(f(x)) = \underline{d}_\mu(x)$ for μ-a.e. x.*

Proof. Applying Lemma 2 to the Lipschitz function $f : E \to \mathbb{R}^m$ yields $\dim_{\mathrm{H}}(f(E)) \leq \dim_{\mathrm{H}}(E)$. Due to the bi-Lipschitzness, $f^{-1} : f(E) \to E$ is also Lipschitz. Hence, by Lemma 2 we get that $\dim_{\mathrm{H}}(E) \leq \dim_{\mathrm{H}}(f(E))$. This proofs $\dim_{\mathrm{H}}(E) = \dim_{\mathrm{H}}(f(E))$ for bi-Lipschitz f. For the (upper/lower) pointwise dimensions, apply Lemma 1 to both f and f^{-1} using the same logic from above. □

B.4 MLE and TwoNN target the pointwise dimension

Let M be a d-dimensional manifold and let $Y_1, \ldots, Y_n \in M$ be i.i.d. with probability measure μ. For simplicity, we use a slightly different notation in the following derivation (compared to other sections), with d and D ($d \ll D$) denoting the dimension of the manifold and the ambient space, respectively. The observed sample $\{X_j\}_{j=1}^n$ is its (smooth) embedding $X_j := g(Y_j) \in \mathbb{R}^D$, with a continuous and sufficiently smooth mapping g as in Levina and Bickel (2004).

Local Model (Homogeneous PPP) Fix a point $x \in M$ for which the pointwise dimension exists, $d_\mu(x) = d$. Assume that in a neighborhood of x, μ has a density κ with respect to the Riemannian volume on M, with κ continuous at x and $0 < \kappa(x) < \infty$. Under these conditions, the standard Binomial-to-Poisson coupling (Penrose and Yukich, 2003, 2013) implies that, at sufficiently small scales around x, the sample $\{X_j\}_{j=1}^n$ can be approximated by a *homogeneous Poisson point process* (PPP) in the tangent space $\mathbb{R}^d$ with *intensity* $\lambda_n = n\,\kappa(x)$, meaning the expected number of points in a set equals λ_n times its d-dimensional volume. In particular, for small $r > 0$, the count $N(r,x) := \sum_{j=1}^n \mathbf{1}\{X_j \in B(x,r)\}$ is well-approximated by $N(r,x) \sim \mathrm{Poisson}\big(\lambda_n\, \omega_d\, r^d\big)$, where ω_d is the volume of the unit ball in $\mathbb{R}^d$.

Levina-Bickel MLE Let $T_i(x)$ be the distance from x to its i-th nearest neighbor. Under the local PPP model, conditional on the distance $T_k(x)$ to the k-th neighbor, the ratios $U_i = T_i(x)/T_k(x)$ for $i = 1, \ldots, k-1$ are the order statistics of $k-1$ i.i.d. random variables drawn from a distribution with CDF $\mathfrak{F}(u) = u^d$ and corresponding PDF $\mathfrak{f}(u|d) = du^{d-1}$ for $u \in [0,1]$. The joint log-likelihood of these order statistics is $\ell(d) = C + \sum_{i=1}^{k-1} \log(d{u_i}^{d-1})$, where C is a constant independent of d. Maximizing $\ell(d)$ w.r.t. d, yields the MLE from Levina and Bickel (2004):

$$\widehat{d}_{\mathrm{MLE}}(x) \;=\; \left[\frac{1}{k-1}\sum_{i=1}^{k-1} \log \frac{T_k(x)}{T_i(x)}\right]^{-1}.$$

As $n \to \infty$, $k \to \infty$ and $k/n \to 0$ so that $r = T_k(x) \downarrow 0$, the PPP small-scale approximation becomes exact in the limit. Because $\widehat{d}_{\mathrm{MLE}}(x)$ is asymptotically unbiased and its variance scales as $O(1/k)$ (Levina and Bickel, 2004), it follows that $\widehat{d}_{\mathrm{MLE}}(x) \to d$ in probability. Note that by defining M as a d-dimensional manifold, we assumed exact dimensionality ($d_\mu(x) = d$ for almost all $x \in M$). However, even in more general spaces with varying pointwise dimension, as the local PPP approximation becomes exact, we obtain the localized result that $\widehat{d}_{\mathrm{MLE}}(x) \to d_\mu(x)$ in probability.

TwoNN In the special case $k = 2$, the ratio $\rho(x) := T_2(x)/T_1(x) \in [1, \infty)$ satisfies $\log \rho(x) \sim \mathrm{Exp}(d)$ and hence ρ is Pareto(d):

$$\mathfrak{f}(\rho \,|\, d) = d\, \rho^{-(d+1)}, \qquad \mathfrak{F}(\rho \,|\, d) = 1 - \rho^{-d} \quad (\rho \geq 1),$$

see Facco et al. (2017, Eqs. (5) & (6)). Treating $\{\rho(x_j)\}$ as approximately independent gives the pseudo-log-likelihood $\ell(d) = n \log d - (d+1) \sum_{j=1}^{n} \log \rho(x_j)$, whose maximizer is

$$\widehat{d}_{\mathrm{TwoNN}} = \left[\frac{1}{n} \sum_{j=1}^{n} \log \frac{T_2(x_j)}{T_1(x_j)} \right]^{-1}.$$

Alternatively, the estimation can be based on linear regression (both approaches are asymptotically equivalent). The regression form used in TwoNN follows from the Pareto distribution-based identity

$$-\log\big(1 - \mathfrak{F}(\rho \,|\, d)\big) = d \log \rho,$$

see Facco et al. (2017, Eqs. (7)). So fitting a straight line (passing through the origin) to $\{\big(\log \rho_j,\ -\log(1 - \widehat{\mathfrak{F}}_n(\rho_j))\big)\}_{j=1}^{n}$ estimates the slope d (Facco et al., 2017). Under exact dimensionality we have that $d_\mu(x) = d$ for μ-a.e. interior x. In that case, as $n \to \infty$ and $r = T_2(x) \downarrow 0$, the PPP small-scale approximation becomes exact, justifying $\log \rho \sim \mathrm{Exp}(d)$. Since $\mathbb{E}[\log \rho] = d^{-1}$, applying the the weak law of large numbers to the sample mean, followed by the continuous mapping theorem, yields $\widehat{d}_{\mathrm{TwoNN}} \to d$ in probability.

B.5 Exact IDs of LLM Embeddings & Representations

As depicted in Fig. 5, each token in a prompt is mapped to a specific point in the latent space of the LLM embeddings. While these embeddings vary between LLMs, and the number of unique points that can be reached in the latent space of the embeddings depends on the size of the vocabulary of the respective LLM, this number is *finite* for all LLMs. Using the fact that the ID of any finite or countable set of points is zero, the lemma below shows that the ID of LLM embeddings are also zero in a strict topological sense. Considering prompts with finite length, the lemma extends this result to the ID of hidden-layer representations in LLMs.

Lemma 3 (ID of LLM representations)**.** *Let $\mathcal{V}$ be a finite vocabulary, and let $\bigcup_{n\geq 1} \mathcal{V}^n$ be the countable set of all prompts, where each prompt x has an arbitrary finite length $|x| = n \in \mathbb{N}$. Let $e : \mathcal{V} \to \mathbb{R}^{d_0}$ be the token embedding map, then the token embedding set $e(\mathcal{V})$ is finite as $\mathcal{V}$ is finite. Further, denote $S_0 = \{(x,t) : x \in \bigcup_{n\geq 1} \mathcal{V}^n, t \in \{1, \ldots, |x|\}\}$ to be the set of all prompts and positions. Consider a deterministic, measurable LLM whose ℓ-th layer representation at position t is given by*

$$F_\ell : \bigcup_{n\geq 1} \mathcal{V}^n \times \mathbb{N} \longrightarrow \mathbb{R}^{d_\ell}, \qquad (x,t) \mapsto F_\ell(x,t),$$

defined for $t \in \{1, \ldots, |x|\}$. For $\ell \geq 1$ define the set of attainable representations in layer ℓ by

$$S_\ell := \big\{F_\ell(x,t) : x \in \textstyle\bigcup_{n\geq 1} \mathcal{V}^n,\ t \in \{1, \ldots, |x|\}\big\}.$$

Then for every $\ell \in \{1, \ldots, L\}$:

1. *(Hausdorff dimension) S_ℓ is countable and $\dim_H(S_\ell) = 0$.*
2. *(Pointwise dimension) If μ_0 is any probability measure supported on S_0 and $\mu_\ell = (F_\ell)_{\#}\mu_0$, then μ_ℓ is purely atomic and, for μ_ℓ-a.e. y,*

$$\underline{d}_{\mu_\ell}(y) = \overline{d}_{\mu_\ell}(y) = 0.$$

Hence each μ_ℓ is exact-dimensional with $d_{\mu_\ell}(y) = 0$ for μ_ℓ-a.e. y.

Proof. The set $\bigcup_{n\geq 1} \mathcal{V}^n$ is countable (countable union of finite sets), and for each x, the index set $\{1, \ldots, |x|\}$ is finite. Thus, the domain $S_0 = \{(x, t)\}$ is countable, and its image S_ℓ under any function is countable. Further, any countable subset of $\mathbb{R}^{d_\ell}$ has a Hausdorff dimension of zero. This proves the first claim.

For the second claim, note that μ_0 is purely atomic as S_0 is countable. The pushforwards of purely atomic measures under any (measurable) map are again purely atomic and supported on S_ℓ, which was shown to be countable above. For any point y (an atom) with mass $c = \mu_\ell(\{y\}) > 0$, we have that $\mu_\ell(B(y, r)) \geq c$ for any arbitrarily small $r > 0$. Combining this with the fact that the (lower/upper) pointwise dimension is bounded below by zero, we get that

$$0 \leq \lim_{r \downarrow 0} \frac{\log \mu_\ell(B(y, r))}{\log r} \leq \lim_{r \downarrow 0} \frac{\log c}{\log r} = 0,$$

which shows that both lower and upper pointwise dimensions are zero at μ_ℓ-a.e. y. □

B.5.1 Conceptual difference between Text and Image Representations

While it might seem natural to extend the above reasoning from text to image representations, the following discussion aims to clarify why Lemma 3 cannot be transferred to image representations. For token-based inputs, the key issue is that one cannot meaningfully and smoothly interpolate between different token embeddings. This is not just due to the finiteness of the input, but to the inherently discrete nature of words and subword units, which yields a finite and separated set of token embeddings.

In contrast, for images it is conceptually natural to think about smooth interpolation between inputs, reflecting that the underlying light spectrum is continuous in reality. Although this continuum is discretized when images are recorded and stored with finite precision, this is typically viewed as a technical approximation rather than a property of the underlying object. Thus, images can be viewed as forming a high-dimensional continuous manifold embedded in a discrete space, where the discreteness arises from digital representation rather than from the underlying object itself, unlike in the token-based case.

Therefore, the difference between images and text with respect to the manifold hypothesis stems from the fundamentally different nature of pixels and token embeddings. Conceptually, smooth interpolation is natural for pixels but not for tokens. Lemma 3 is aimed at making this intuition mathematically precise.

C Additional Experiments and Experimental Details

Computing Environment Experiments involving the extraction of layer-wise representations for the LLMs and ViTs were performed on an NVIDIA Tesla T4 GPU. The CNN experiments can additionally be run using Apple's Metal Performance Shaders (MPS) backend on a MacBook with at least 16 GB of RAM. In our experimental setup, extracting layer-wise representations for a single model and computing each metric required less than one hour. The code is available at `https://github.com/rickmer-schulte/rethinking-neural-id`.

C.1 Layer-wise Analysis

C.1.1 Layer-wise Analysis of LLM Representations

For the layer-wise neural representation analysis of LLMs, we follow along the investigation of Cheng et al. (2025). Similar to their analysis, we base ID-estimation on 10k prompts from the popular wikitext dataset (Merity et al., 2017). For each prompt, we extract layer-wise representations at the final position as described in Section 3.3. We do this for the pretrained models Llama-3.1-8B (Grattafiori et al., 2024), Mistral-7B-v0.3 (Jiang et al., 2023), and Pythia-6.9B (Biderman et al., 2023). Pretrained weights for each of these models are obtained from Hugging Face.

Similar to several other LLM-based analyses, ID estimation is based on the Gride ID estimator (Denti et al., 2022) using the *DADApy* implementation (Glielmo et al., 2022) with default parameter settings, such as range_max = 64. The latter indicates that the 64^{th} NN distance is the maximum NN distance that is involved in the ID estimation.

C.1.2 Layer-wise Analysis of CNN Representations

For the layer-wise neural representation analysis of convolutional neural networks (CNNs), we follow along the investigation of Ansuini et al. (2019). We consider several classic model architectures such as *Alexnet* (Krizhevsky et al., 2012), *VGG* (Simonyan and Zisserman, 2014), *ResNet* (He et al., 2016) with pretrained weights (pretrained on ImageNet (Deng et al., 2009)) obtained from the PyTorch library *torchvision* (Paszke et al., 2019). The addition of "(BN)" for the VGGs of Fig. 1 indicates that models incorporate Batch Normalization layers.

The layer-wise neural representations are obtained by passing a set of images through each pretrained model and extracting the corresponding representations from the layers. Each dataset consists of 500 samples from the seven largest categories of ImageNet. Further details can be found in Ansuini et al. (2019). In Figs. 1, 4, 13, 15, 18, 21 and 23, the point estimates correspond to the mean of the respective estimates on the seven datasets, and the error bars to the corresponding standard deviations. The estimated intrinsic dimensions are obtained via the TwoNN ID estimator (Facco et al., 2017). In each of these plots, the x-axis denotes the relative instead of the absolute depth of each model layer to facilitate visual comparison between models with varying numbers of layers.

The CNN-based layer-wise analysis of representations concerning other metrics than the ID is performed along four representative examples of the CNNs in Fig. 1. In the following sections, we conduct experiments analogous to the ones found for LLMs in the main text, along the example of the four pretrained CNN models *Alexnet*, *VGG-16*, *ResNet-18*, and *ResNet-34*.

C.1.3 Layer-wise Analysis of ViT Representations

For the analysis of vision transformers (ViTs), we extract layer-wise representations from a standard ViT (google/vit-base-patch16-224) (Wu et al., 2020) and two DINOv3 models of different sizes (facebook/dinov3-vitb16-pretrain-lvd1689m; facebook/dinov3-vitl16-pretrain-lvd1689m) (Siméoni et al., 2025) based on 5k images from ImageNet and conducted a layer-wise analysis analogous to our previous investigation. In the spirit of our previous transformer-based analysis, we extract the layer-wise representation of the CLS token for each image, given that all information necessary for prediction tasks is encoded in it. Pretrained weights for each of the ViT models are obtained from Hugging Face.

C.2 Metrics

Von Neumann Entropy In Section 4.5 we considered the *von Neumann entropy* of the distribution of eigenvalues (of the Gram matrix) of layer-wise representations. The entropy can be computed layer-wise. Let $Z \in \mathbb{R}^{n \times d}$ be a latent representation in a hidden layer based on n observations with d-dimensional neural representation. First, define the corresponding Gram matrix $Q = ZZ^\top \in \mathbb{R}^{n \times n}$. Since Q is symmetric positive semidefinite, we can normalize it by its trace to obtain

$$\tilde{Q} = \frac{Q}{\operatorname{tr}(Q)},$$

which satisfies $\operatorname{tr}(\tilde{Q}) = 1$. The eigenvalues $p_1, \ldots, p_n$ of $\tilde{Q}$ are given by

$$p_i = \frac{\lambda_i(Q)}{\operatorname{tr}(Q)}, \qquad i = 1, \ldots, n,$$

where $\lambda_1(Q), \ldots, \lambda_n(Q)$ denote the eigenvalues of Q. Then the *von Neumann entropy* of $\tilde{Q}$ is defined as $S(\tilde{Q}) = -\sum_{i=1}^{n} p_i \log(p_i)$, with the convention that $0 \log(0) = 0$. Because only the nonzero eigenvalues contribute, this can equivalently be written as

$$S(\tilde{Q}) = -\sum_{i=1}^{r} p_i \log(p_i)\,, \tag{3}$$

where r denotes the rank of Q with $r = \operatorname{rank}(Q) \leq \min(n, d)$. Note that this is a special type of so-called *matrix-based entropy* (Giraldo et al., 2014) that was also considered in the analysis of Skean et al. (2025). Considering the normalized singular values instead of the eigenvalues and exponentiating the corresponding results in (3), one obtains the so-called *effective rank* of Z (Roy and Vetterli, 2007). In our experiments, we computed the von Neumann entropy based on mean-centered representations to capture the spread of the representations in latent space independent of translations. In general, point estimates correspond to the mean of the entropy estimates, and the error bars or shaded areas are based on the standard deviations in the entropy-related plots.

Other Metrics Our experiments also cover a layer-wise analysis of the length and pairwise cosine similarity of representations. The length (or size) of the representations is measured by L_2-distances to the origin of the latent space for each representation. The cosine similarity is computed between and averaged over pairs of layer-wise representations, measuring how closely aligned the representations are over several hidden layers of the model. Given the high number of samples and, therefore extremely high number of pairs in the case of the LLM-based analysis, we adapt a block-wise analysis of pairs to be more memory efficient.

C.3 Intrinsic Dimension Estimator Analysis

In Fig. 2, we investigate the accuracy of the two ID estimators, TwoNN and MLE. In order to obtain datasets with known intrinsic dimension, we sample 5k data points uniformly distributed on a $d_{\mathcal{M}}$-hyperball with varying true ID $d_{\mathcal{M}}$. For each true intrinsic dimension, we sample 20 datasets and estimate the IDs via MLE and TwoNN on each of these.

Fig. 2 and Fig. 11 depict both the average over these 20 ID estimates (lines) and the related 95% CI. For the sampling and subsequent ID estimation, we use the *scikit-dimension* library (Bac et al., 2021). Fig. 2 shows a strong negative bias for the two estimators that is growing with increasing true intrinsic dimension. While Fig. 2 uses MLE with $k = 20$, Fig. 11 shows that the negative bias is persistent also for other choices of nearest neighbors k.

Similar to Fig. 13, we also demonstrate in a controlled study setup (with known ID and datasets sampled as described above) that the ambient space dimension does not seem to have strong effects on the TwoNN and MLE estimates. For a fixed true ID of 50, the two estimators still exhibit a negative bias, but are largely invariant to changes in the size of the ambient space dimension, as can be seen in Fig. 12.

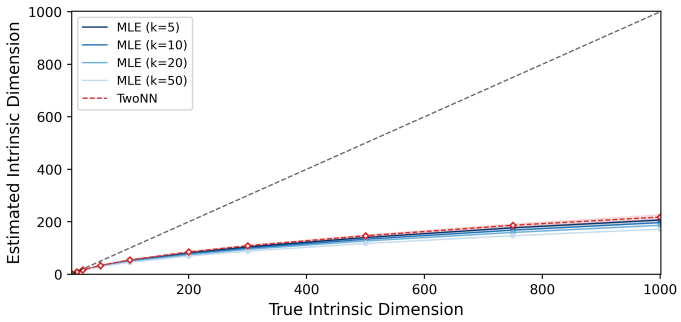

Figure 11: Estimated vs. True ID: Estimated IDs using TwoNN and MLE (different k) of datasets with varying true ID. Each dataset consists of 5k data points uniformly distributed on a $d_{\mathcal{M}}$-hyperball with varying true ID $d_{\mathcal{M}}$. 95% CI are computed based on 20 ID estimates. Both estimators exhibit strong negative bias with increasing $d_{\mathcal{M}}$.

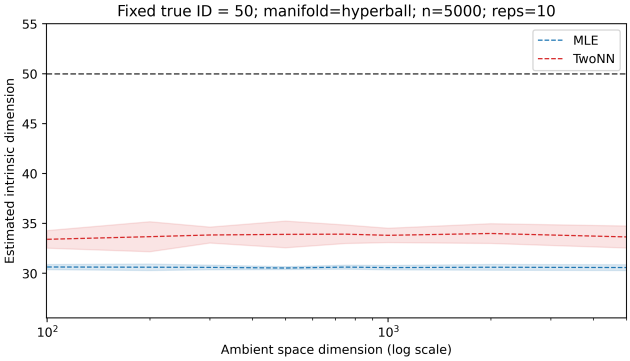

Figure 12: Estimated ID vs. Ambient Space Dim.: Estimated IDs using TwoNN and MLE ($k = 20$) of datasets with varying ambient space dimension d. Each dataset consists of 5k data points uniformly distributed on a $d_{\mathcal{M}}$-hyperball with fixed to $d_{\mathcal{M}} = 50$. 95% CI are computed based on 10 ID estimates. Ambient space dimension does not strongly impact ID estimates.

C.4 ID vs. Ambient Space Dimension

In Section 4.5, we explored what might be underlying factors that drive the commonly found layer-wise ID patterns. Along with this analysis, we also conduct an experiment comparing the layer-wise ID patterns with the layer-wise embedding dimension in different pretrained models. The result in Fig. 13 shows that the two patterns are relatively different, indicating that the ambient space dimension itself does not seem to be the core driving factor for the commonly found ID patterns.

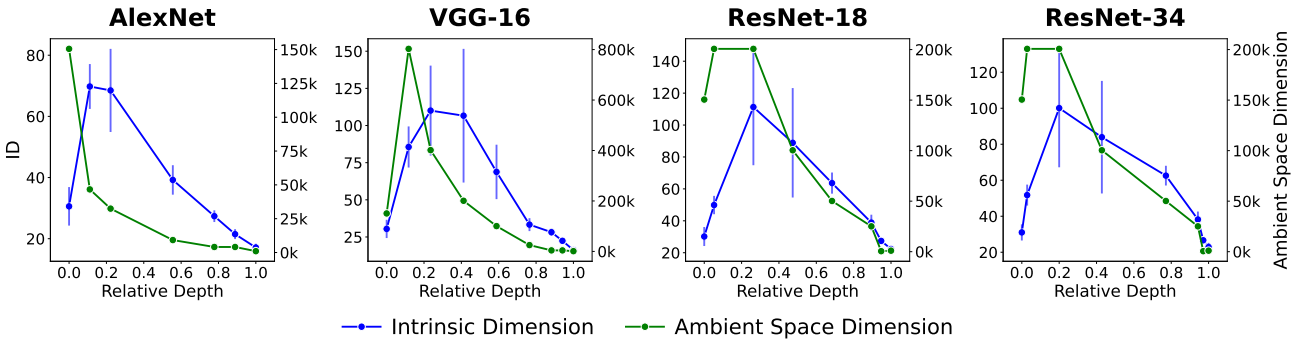

Figure 13: Layer-wise comparison of estimated intrinsic dimensions (left y-axis) vs. ambient space dimensions (right y-axis) of neural representations from different pre-trained convolutional architectures. The x-axis shows the relative depth of the model layers. The ambient space dimension corresponds to the width of each layer.

C.5 Class-specific Intrinsic Dimension Estimates

Analogous to Fig. 4, we also plot the class-specific ID estimates for the other pre-trained convolutional architectures in Fig. 14. The estimated IDs in each plot are computed separately for the layer-wise representations corresponding to images from the seven largest classes of the ImageNet (Deng et al., 2009) dataset. Similar to Fig. 4, also the class-specific ID patterns in case of the other convolutional architectures show an increase in estimated IDs over the layers. ID estimates for layer-wise representations of the vision models are again obtained using the TwoNN estimator.

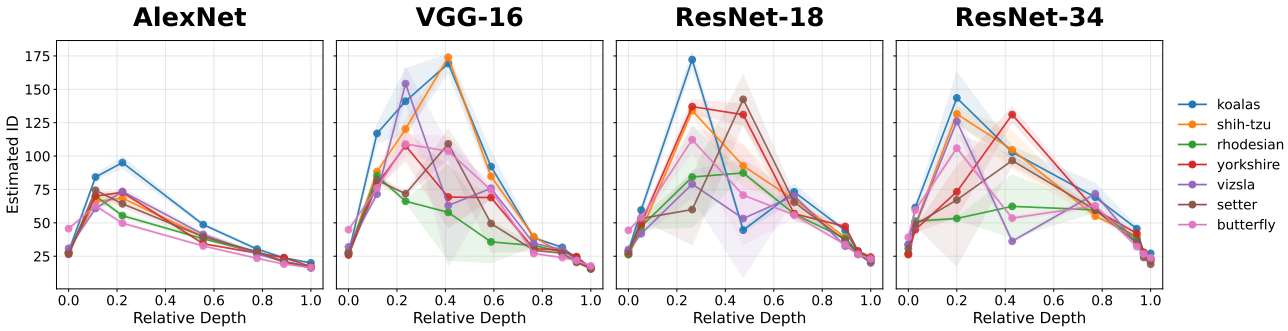

Figure 14: Class-specific estimated IDs of layer-wise representations from various pre-trained convolutional architectures separately for images with different categories (colors) from ImageNet. The x-axis shows the relative depth of model layers, and shaded areas show the estimated standard errors. The first plot is adapted from Ansuini et al. (2019).

C.6 Average k-NN Distance Analysis

We also investigate the average NN distances of layer-wise representations for the pre-trained convolutional architectures. Given that for these types of models, usually the MLE and TwoNN estimators are used for ID estimation, we consider the consecutive NN distances that are involved in their respective calculations. Therefore, Fig. 15 depicts the average 1^{st} to 5^{th}-NN distances ($k = 1, \ldots, 5$) of layer-wise representations from different pretrained vision models.

The layer-wise NN distances differ in both their shapes and overall magnitude between the models. The fact that average NN distances are much higher for the VGG model can, to some extent, be explained by the much larger ambient space dimensions (width of layers) for this model compared to the others. As can be seen in Fig. 13, the representations in intermediate layers of the VGG-16 model are about 400k-800k dimensional (compared to about 50k-200k dimensional intermediate-layer representations in the other models). Thus, k-NN distances grow given that points become increasingly separated in high dimensions (cf. Section 4.2). Interestingly, however, for all models the k-th NN distances in Fig. 15 grow and shrink by similar amounts over the layers. By a similar reasoning as in Section 4.1, this explains the shape of the layer-wise ID patterns for the vision models (cf. Fig. 1).

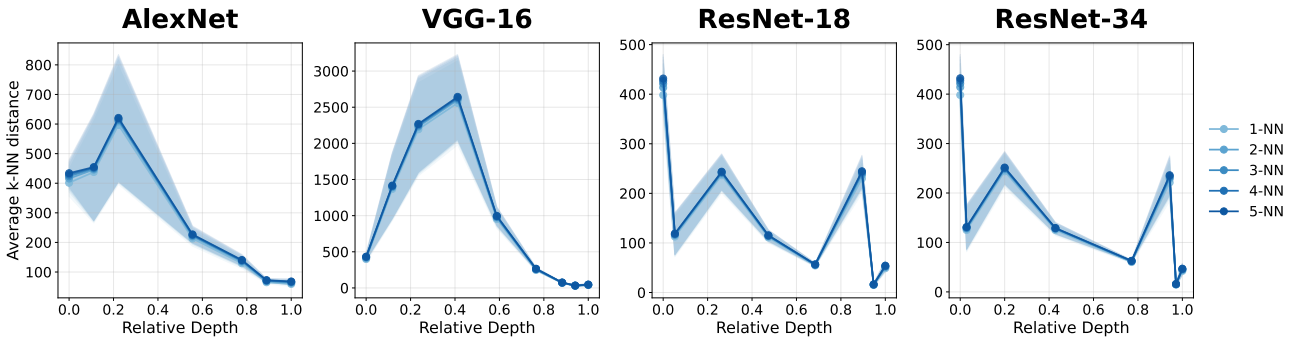

Figure 15: Average k-NN distances of representations over the layers of various pre-trained convolutional architectures. Averages over the 1^{st} to 5^{th}-NN distances are depicted and displayed with a corresponding color. Shaded areas correspond to twice the standard deviation.

C.7 Cosine Similarity Analysis

LLMs The extended results of the analysis of pairwise cosine similarity of layer representations from LLMs can be found in Fig. 16. The latter includes the last layer of the three models. As described in Section 4.1, the last layer representations are subject to a final Layer Norm transformation, which can induce drastic changes in the last layer. The latter is likely the reason for the drop in pairwise cosine similarity in the last layer, which is especially apparent for the pythia model. However, the key insight from Fig. 16 for our analysis is that the layer-wise cosine similarity patterns are very different compared to the layer-wise ID and NN distance patterns depicted in Fig. 6 and Fig. 7, respectively. Accordingly, varying cosine similarities between representations cannot be the key driving force behind observed ID patterns, in line with the conclusion in Section 4.3.

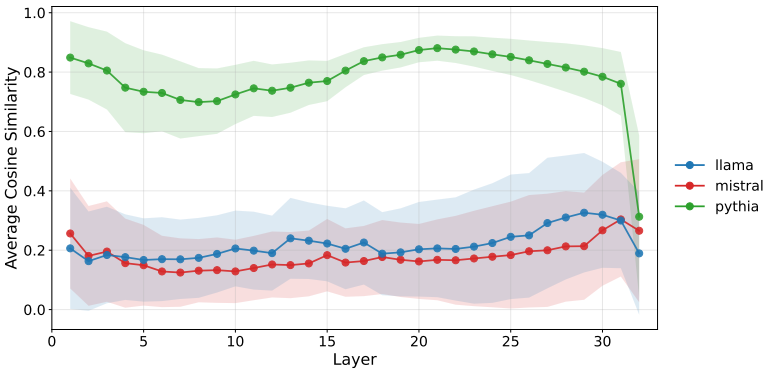

Figure 16: Average cosine similarity between layer-wise representations of the LLM models (llama, mistral, and pythia). The shaded area band represents twice the standard deviation.

ViTs We also investigate the pairwise cosine similarity of layer representations from ViTs. The results can be found in Fig. 17. Similar to the LLM-based analysis, besides a drastic change in the last layer, pairwise cosine

similarity estimates generally do not vary a lot across model layers. Moreover, layer-wise patterns seem very different from the respective layer-wise ID patterns, which are shown in Fig. 22.

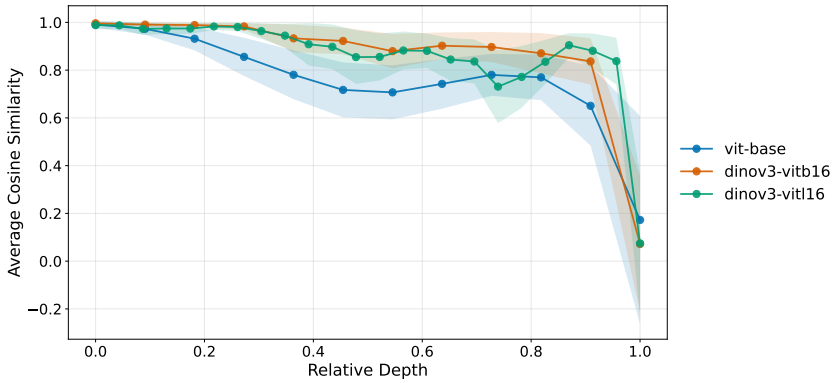

Figure 17: Average cosine similarity between layer-wise representations for different pre-trained ViTs. The shaded area band represents twice the standard deviation.

CNNs We also investigate the pairwise cosine similarity of layer representations from the CNNs. The results for the four pre-trained convolutional architectures can be found in Fig. 18. Interestingly, there seems to be a layer-wise pattern of pairwise cosine similarity that is common to all models, which might be of interest on its own. However, for our analysis, the most important insight of Fig. 18 is that the layer-wise cosine similarity patterns are very different from the layer-wise NN distances and ID patterns, and therefore cannot explain the latter. This insight is in line with the results from LLM-based analysis.

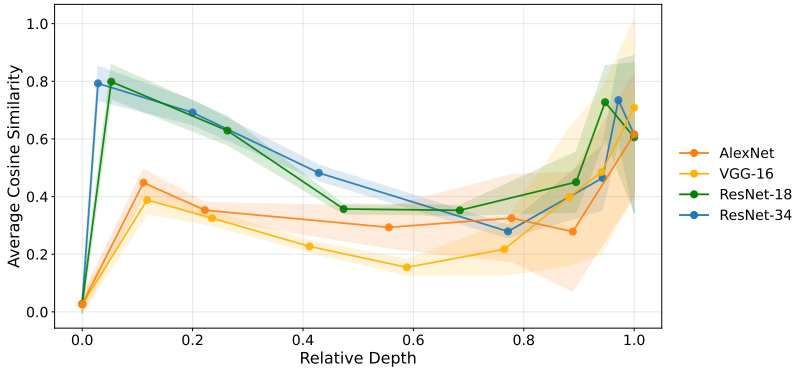

Figure 18: Average cosine similarity between layer-wise representations for different pre-trained convolutional architectures. The shaded area band represents twice the standard deviation.

C.8 Representation Length Analysis

LLMs The results of the analysis of size-based analysis of layer representations from LLMs can be found in Fig. 19. The left plot depicts the average lengths of layer-wise representations from the three LLM models. The length of each representation is measured by the usual L_2 distance to the origin of the latent space. Given that these distances tend to increase over the hidden layers of all models (besides a constant dimension of the layer-wise latent space), this corresponds to an expansion of representation in latent space. This phenomenon was also described in Section 4.4. A notable exception is the last layer. While the size of representations drastically increases in the case of the llama and mistral model, it drastically decreases for the pythia model. While these model-specific differences might be of interest on their own, they are of minor importance for our analysis.

For our purposes, the key insight from Fig. 19 is that the layer-wise L_2 lengths of representations (right plot) very closely match the layer-wise NN distances (left plot). Therefore, L_2 lengths of representations and the expansion

in latent space as very likely to be the driver behind the observed NN distance. As the latter are used by ID estimators, they give rise to observed layer-wise ID patterns.

The left plot in Fig. 19 corresponds to an extension of Fig. 7. The former also includes the NN distances for last-layer representations, which are omitted in Fig. 7. As described above, the drastic change in NN distances in the last layer is induced by the drastic change in the length of representations in the last layer. As this phenomenon is not central to our discussion and makes it very hard to read of differences between NN distances, we omitted the last layer in Fig. 7.

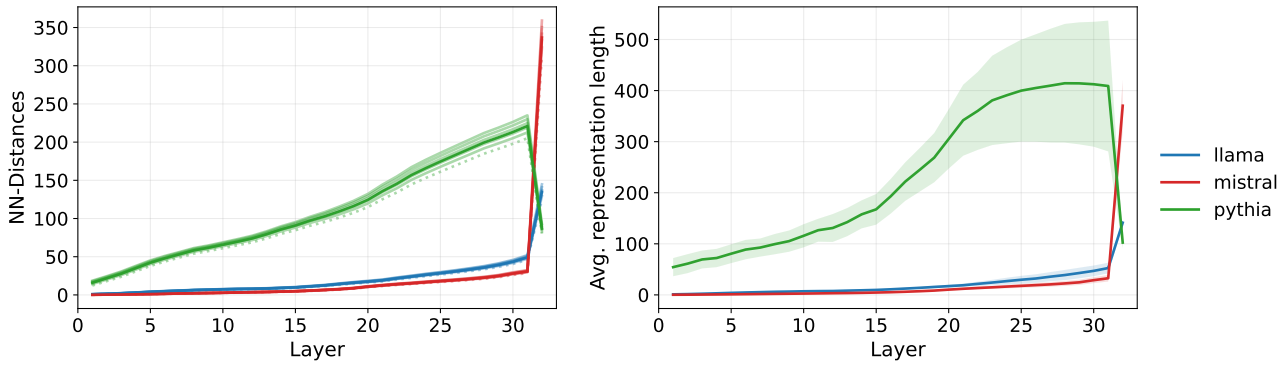

Figure 19: NN distances (left) and average length (measured by L_2-distances to the origin) of layer-wise representations for the llama, mistral, and pythia models. The left plot corresponds to Fig. 7, but additionally includes the last layer. The shaded area band in the right plot represents twice the standard deviation.

ViTs The results for the length-based analysis for layer-wise representations of ViTs are depicted in Fig. 20. Analogous to the LLM-based results, the layer-wise patterns of L_2 length of representations generally grow over the hidden layers and closely resemble the layer-wise NN distance patterns.

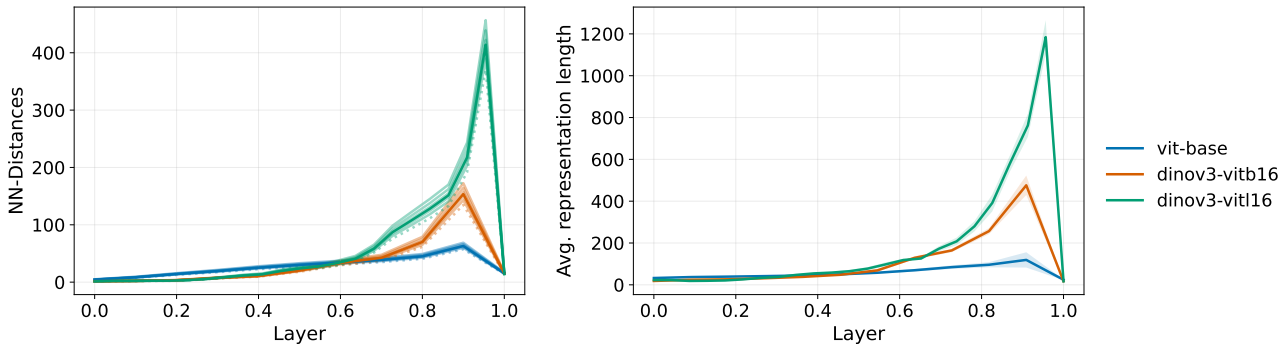

Figure 20: NN distances (left) and average length measured by L_2-distances to the origin (right) of layer-wise representations for different ViTs. In the left plot, for each model, each line (top to bottom) corresponds to the averages of 64^{th} & $32^{nd}, \ldots, 2^{nd}$ & 1^{st} NN distances. Solid lines denote the average over all 6 lines, with the average of the first two NN distances (used in the TwoNN estimator) highlighted as a dotted line. The shaded area band in the right plot represents twice the standard deviation.

CNNs The results for the length-based analysis for layer-wise representations of the CNNs are depicted in Fig. 21. Analogous to the other models, the layer-wise patterns of L_2 length of representations closely resemble the NN distance patterns found in Fig. 15 and are therefore likely the driving force behind the NN distances. However, in contrast to the LLM-based results, the lengths of representations do not exhibit an increasing but rather a decreasing trend over the layers, and therefore no expansion in latent space. It should be noted, however, that the dimension of the layer-wise ambient space of representations is not constant but mostly decreasing over the layers of CNNs (cf. Fig. 13). Hence, a decrease in layer-wise representation lengths is expected for CNNs.

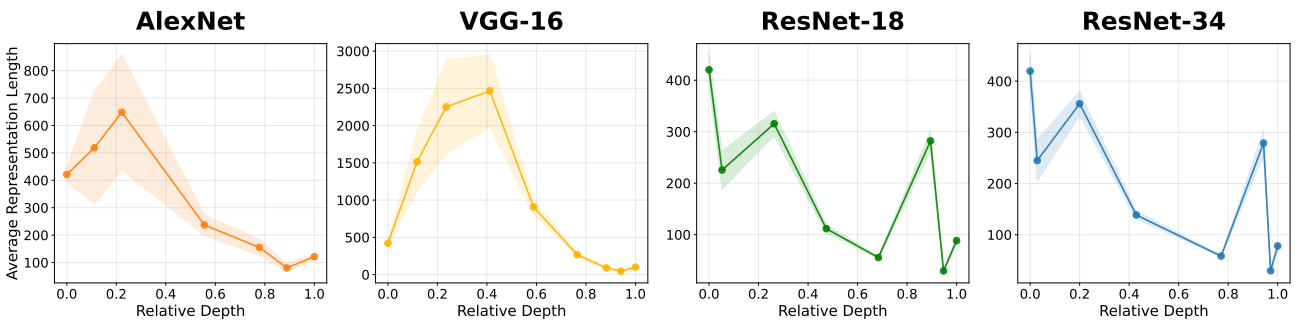

Figure 21: Average length (measured by L_2-distances to the origin) of layer-wise representations for different CNNs. The shaded area band represents twice the standard deviation.

C.9 Entropy vs. ID Analysis

ViTs and CNNs We also extended the analysis comparing layer-wise ID and entropy estimates in Section 4.5 from LLMs to ViTs and CNNs. The results for the ViT and CNN models are depicted in Fig. 22 and Fig. 23, respectively. Analogously to the findings obtained in the LLM-based analysis (cf. Fig. 10), we find a strong connection between the layer-wise patterns of estimated IDs and von Neumann entropy also for these models..

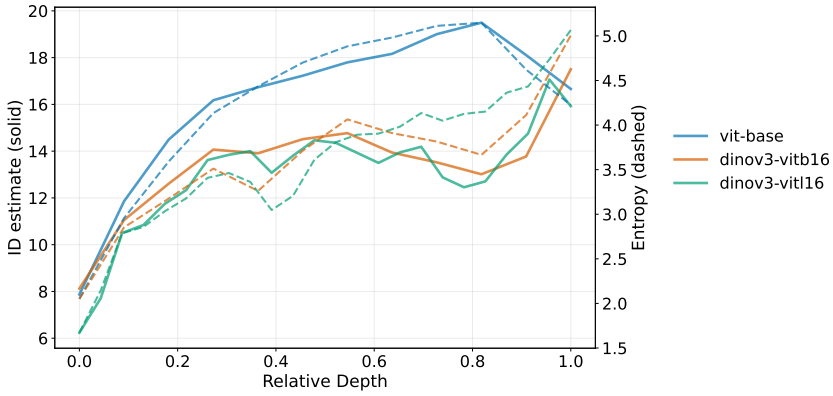

Figure 22: Layer-wise comparison of estimated intrinsic dimensions (left y-axis) vs. von Neumann entropy (right y-axis) of neural representations from different ViTs. Details about the entropy metric can be found in Appendix C.2.

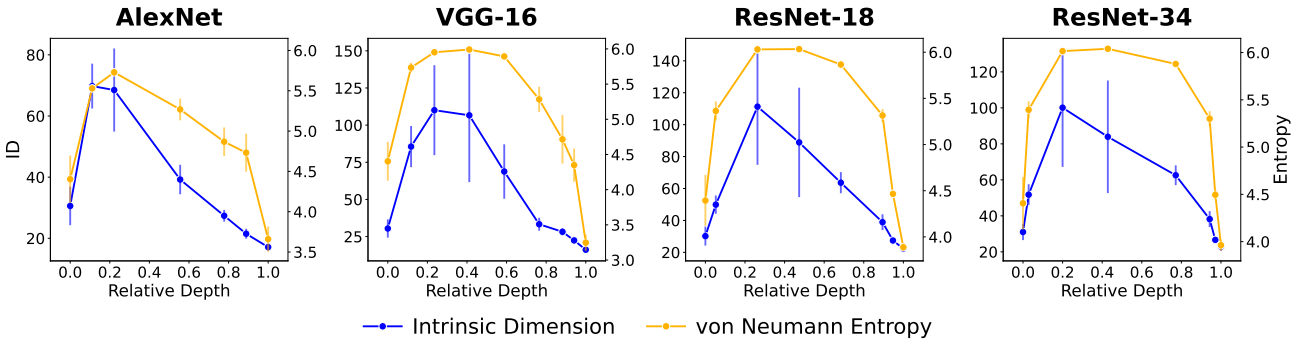

Figure 23: Layer-wise comparison of estimated intrinsic dimensions (left y-axis) vs. von Neumann entropy (right y-axis) of neural representations from different CNNs. Details about the entropy metric can be found in Appendix C.2.