A Geometric Framework for Understanding Memorization in Generative Models

Brendan Leigh Ross [1] **Hamidreza Kamkari** [1] **Zhaoyan Liu** [1] **Tongzi Wu** [1]
George Stein [1] **Gabriel Loaiza-Ganem** [1] **Jesse C. Cresswell** [1]

Abstract

As deep generative models have progressed, recent work has shown that they are capable of memorizing and reproducing training datapoints when deployed. These findings call into question the usability of generative models, especially in light of the legal and privacy risks brought about by memorization. To better understand this phenomenon, we propose a geometric framework which leverages the manifold hypothesis into a clear language in which to reason about memorization. We propose to analyze memorization in terms of the relationship between the dimensionalities of (i) the ground truth data manifold and (ii) the manifold learned by the model. In preliminary tests on toy examples and Stable Diffusion (Rombach et al., 2022), we show that our theoretical framework accurately describes reality. Furthermore, by analyzing prior work in the context of our geometric framework, we explain and unify assorted observations in the literature and illuminate promising directions for future research on memorization.

1. Introduction

Suppose $\{x_i\}_{i=1}^n$ is a dataset in $\mathbb{R}^d$ drawn independently from a ground truth probability distribution $p_*(x)$. A deep generative model (DGM) is a tractable probability distribution $p_\theta(x)$ designed to capture $p_*(x)$ only from knowledge of the dataset $\{x_i\}_{i=1}^n$. DGMs, and most famously, diffusion models (DMs; Sohl-Dickstein et al., 2015; Ho et al., 2020), have featured in the "generative AI" boom with their ability to generate realistic and diverse images from text prompts (Karras et al., 2019; Rombach et al., 2022). They have also been applied successfully in other domains such as tabular data and language (Li et al., 2022; Zhang et al., 2023). DMs are thus likely to be deployed in an increasing number of public-facing or safety-critical applications.

However, when sufficiently powerful, DGMs are known to memorize their training data. Memorization occurs at various degrees of specificity, including identities of brands, layouts of specific scenes, or exact copies of images (Webster et al., 2021; Somepalli et al., 2023a; Carlini et al., 2023).

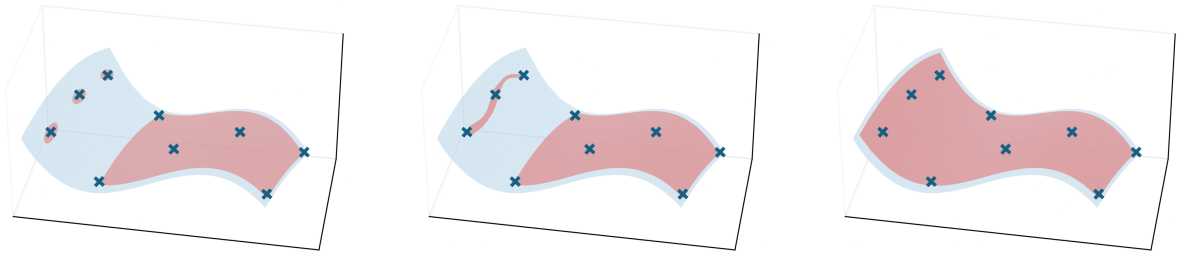
Memorization is undesirable for myriad reasons. Simply put, a model that reproduces its training data is no more useful than the training data itself. Memorization is a modelling failure under the DGM definition provided above; if the underlying ground truth $p_*(x)$ does not place positive probability mass on individual datapoints, then a $p_\theta(x)$ that memorizes must be failing to generalize (Yoon et al., 2023). But memorization's risks go beyond mere utility. Training datasets may contain private information which, if memorized, might be exposed in downstream applications. Reproduced training samples can also open up model builders or users to legal liability; for instance, the recent legal decision of Orrick (2023) hinged on whether generated images were "substantially similar" to training data.

The increasing dependence of society on generative models and resulting risks call for work to better understand memorization. Recent empirical work has identified mechanistic causes of memorization: data complexity, duplication of training points, and highly specific labels (Somepalli et al., 2023b; Gu et al., 2023). We group these insights under the umbrella of "memorization phenomena", a catch-all term for the various interesting memorization-related observations we would like to understand better. Though useful in practice, these memorization phenomena have yet to be unified and interpreted under a single theoretical framework. Meanwhile, formal treatments of memorization have led to isolated usecases such as detecting (Meehan et al., 2020) and preventing (Vyas et al., 2023) memorization on a model level, but provided little explanatory power for these memorization phenomena. In addition to providing theoretical insights, a unifying framework could yield more capabilities such as identifying whether a training image has been memorized, altering the sampling process to reduce memorization, and detecting memorized generations post hoc.

In this work, we advance a geometric framework to explain

[1]Layer 6 AI. Correspondence to: Brendan Leigh Ross, Hamidreza Kamkari, Zhaoyan Liu, Tongzi Wu, George Stein, Gabriel Loaiza-Ganem, Jesse C. Cresswell <{brendan, hamid, zhaoyan, tongzi, george, gabriel, jesse}@layer6.ai>.

Proceedings of the ICML 2024 NextGenAISafety workshop, Vienna, Austria.



(a) Exact memorization with $\text{LID}_\theta(x) = 0$. (b) Memorization with $\text{LID}_\theta(x) = 1$. (c) No memorization with $\text{LID}_\theta(x) = 2$.

Figure 1: An illustrative example of LID values for three models with different degrees of memorization. In these plots, the 2-dimensional ground truth manifold $\mathcal{M}_*$ is depicted in light blue, training samples $\{x_i\}_{i=1}^n \subset \mathcal{M}_*$ are depicted as crosses, and the model manifolds $\mathcal{M}_\theta$ are depicted in red. In Figure 1a, the model assigns 0-dimensional point masses around the three leftmost datapoints, indicating that it will reproduce them directly at test time. The model in Figure 1b still memorizes, but with an extra degree of freedom in the form of a 1-dimensional submanifold containing the three points. Only the model in Figure 1c, which has learned a 2-dimensional manifold through its full support, has generalized well enough to avoid memorization.

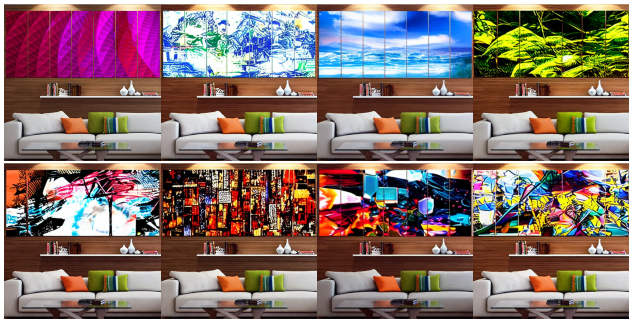


Figure 2: 8 images along a relatively low-dimensional manifold learned by Stable Diffusion v1.5. The first is a real image from LAION (flagged as memorized by Webster (2023)), and the remainder were generated by the model.

memorization. In short, we propose that *memorization occurs at a point $x \in \mathbb{R}^d$ when the manifold learned by the generative model contains x but has too small a dimensionality in its neighbourhood.* As we will see, this understudied perspective is a natural take on memorization that leads to practical insights and effectively explains memorization phenomena like those mentioned above. We mainly focus on DMs, the most notorious memorizers, but our geometric framework applies to any DGM on a continuous data space $\mathbb{R}^d$. Pidstrigach (2022) was the first to show that DMs are capable of learning low-dimensional structure in $\mathbb{R}^d$ and that this manifold learning capability is a driver of memorization; in this sense, our work extends this connection into a general framework, grounds it in empirical findings, and connects it to recent work on memorization.

First in this paper, we lay out our geometric framework for memorization. After defining the key notions of the data manifold and local intrinsic dimension (LID), we describe how LIDs correspond directly to memorization. Second, we provide an empirical proof-of-concept for our framework showing that LID is strongly predictive of memorization by Stable Diffusion (Rombach et al., 2022). Lastly, we describe the memorization phenomena observed in past work and situate them within our framework. We hope these insights will inspire future work on understanding, identifying, and preventing memorization in generative models.

2. Understanding Memorization through LID

Preliminaries Here we presume the manifold hypothesis: that data of interest lies on a manifold $\mathcal{M} \subset \mathbb{R}^d$ (Bengio et al., 2013). In particular, we take a generalized definition of manifold in which $\mathcal{M}$ is allowed to have different dimensionalities in different regions,[1] which is appropriate for realistic, heterogeneous data with varying degrees of structure and complexity. In particular, we assume that both our ground truth distribution $p_*(x)$ and our model $p_\theta(x)$ produce samples on manifolds, which we refer to as $\mathcal{M}_*$ and $\mathcal{M}_\theta$ respectively. We direct readers to Loaiza-Ganem et al. (2024) for a justification and formal mathematical treatment of both of these assumptions, which are especially valid when the data is high-dimensional and the models are high-performing ones such as DMs and GANs (Goodfellow et al., 2014; Karras et al., 2019).

Our framework for understanding memorization revolves around the notion of *local intrinsic dimensionality* (LID). Given a manifold $\mathcal{M}$ and a point $x \in \mathcal{M}$, we define the LID of x ($\text{LID}(x)$) with respect to $\mathcal{M}$ as the dimensionality of $\mathcal{M}$ in the component containing x. In this work, we will consider the LIDs of points $x \in \mathbb{R}^d$ with respect to two specific manifolds: $\mathcal{M}_*$ and $\mathcal{M}_\theta$. We will refer to these quantities as $\text{LID}_*(x)$ and $\text{LID}_\theta(x)$, respectively.

Intuition and the Manifold Hypothesis Before discussing our framework, we review some intuition relating the manifold hypothesis to practical datasets. Manifold structure $\mathcal{M} \subset \mathbb{R}^d$ arises from sets of constraints.

[1] Most authors define a manifold to have a constant dimension over the entire set. Under this common definition, our assumption is referred to as the union of manifolds hypothesis (Brown et al., 2023). We use a more general definition of manifold for brevity.

These can either be very simple, like a set of linear constraints ($\mathcal{M} = \{x \mid Ax = b\}$), or highly complex ($\mathcal{M} = \{x \mid x \text{ is an image of a face}\}$). Locally at a point $x \in \mathcal{M}$, each constraint determines a direction one cannot move without leaving the manifold and violating the structure of the dataset.[2] Hence, a region governed by ℓ independent and active constraints will have dimensionality $\text{LID}(x) = d - \ell$. The value of $\text{LID}(x)$ represents the number of degrees of freedom – valid directions of movement in which the characteristics of the dataset are preserved. Another connection is to complexity. For example, estimates of LID from algorithms like LIDL (Tempczyk et al., 2022) and FLIPD (Kamkari et al., 2024b) have been shown to correspond closely with the complexity of an image; it is reasonable to expect that images with more complex features (i.e., more information) can endure more changes (such as morphing, moving, or changing the colours of different parts of the image) without losing coherence. A limiting example of the LID-complexity connection is a constraint-free dataset of pure random noise, wherein each datapoint contains maximal information and changes in any direction are valid. The notions of constraints, degrees of freedom, and complexity and their relationship to LID will help us understand its connection to memorization in later sections.

A Geometric Framework for Memorization As a motivating example, consider Figure 1, which depicts three possible models $p_\theta(x)$ trained on a dataset $\{x_i\}_{i=1}^n$ that lies on the ground truth manifold $\mathcal{M}_*$. In the first scenario, Figure 1a, the model $p_\theta(x)$ has precisely memorized some of the training data. This is a well-understood mode of memorization; training datapoints are precisely reproduced. To achieve this, the model has learned a 0-dimensional manifold around these datapoints. To our knowledge, Pidstrigach (2022) was the first to point out that a model capable of learning 0-dimensional manifolds can memorize the training data. From this example, we infer that x can be perfectly reproduced when $\text{LID}_\theta(x) = 0$. This indicates suboptimality in the model at the datapoints shown, for which $\text{LID}_*(x) = 2$.

However, memorization can be more complex than simply reproducing a datapoint. For example, Somepalli et al. (2023a) identify instances where layouts, styles, or foreground or background objects in training images are copied without copying the entire image, a phenomenon they refer to as *reconstructive memory*. Webster (2023) surfaces more instances of the same phenomenon and refers to them as *template verbatims*. See Figure 2 for an example. In the region of these points $x \in \mathcal{M}_\theta$, the model is able to generate images with degrees of freedom in some attributes (eg. colour or texture), but is too constrained in other attributes (eg. layout, style, or content). Geometrically, $\mathcal{M}_\theta$ is too constrained compared to the idealized ground truth manifold $\mathcal{M}_*$; i.e., $\text{LID}_\theta(x) < \text{LID}_*(x)$. We depict this situation in Figure 1b, wherein the model has erroneously assigned $\text{LID}_\theta(x) = 1$ for some of the training datapoints.

No memorization is present in Figure 1c, in which the model manifold $\mathcal{M}_\theta$ matches the ground truth manifold $\mathcal{M}_*$.

Two Types of Memorization In light of the above framework, we expect two types of memorization to be of interest. An academic interested in designing DGMs that learn the ground truth distribution correctly will chiefly be interested in avoiding the memorization scenario $\text{LID}_\theta(x) < \text{LID}_*(x)$ (as well as underfitting, wherein $\text{LID}_\theta(x) > \text{LID}_*(x)$). We refer to this first scenario as *overfitting-driven memorization* (OD-Mem). This situation represents a modelling failure in that $p_\theta(x)$ is not generalizing correctly to $p_*(x)$.

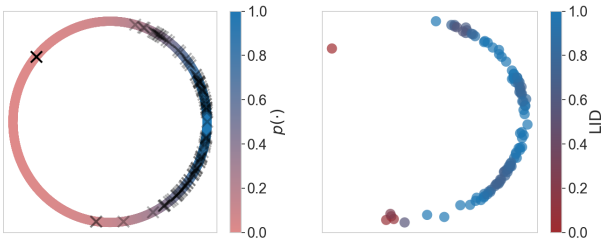
However, an industry practitioner deploying a consumer-facing model might be more interested in hypothetical values of LID_θ *per se*, irrespective of the values of LID_*. For any points $x \in \mathcal{M}_*$ containing trademarked or private information, low values of $\text{LID}_\theta(x)$ will be of concern even if $\text{LID}_\theta(x) = \text{LID}_*(x)$, as this information is likely to be revealed in samples generated from this region. A practitioner would rightly refer to this situation as memorization despite the model generalizing correctly. We refer to this second scenario as *data-driven memorization* (DD-Mem). This certainly happens in practice; for example, conditioning on the title of a specific artwork (e.g. *The Great Wave off Kanagawa* by Katsushika Hokusai (Somepalli et al., 2023a)) is a very strong constraint, leaving few degrees of freedom in the ground truth manifold $\mathcal{M}_*$, but reproducing specific artworks may be undesirable in a production model.

3. Experiments

In this section we verify the geometric framework empirically. In particular, we test that memorized training data has lower LID_θ than unmemorized data on both synthetic toy examples and real world image datasets. Multiple algorithms exist to estimate $\text{LID}_\theta(x)$ for a diffusion model $p_\theta(x)$ (Stanczuk et al., 2022; Horvat & Pfister, 2024). However, to our knowledge, only one is tractable at the scale of Stable diffusion: FLIPD (Kamkari et al., 2024b), which we find can provide sufficiently accurate LID_θ estimates with only 1 model evaluation. (On the other hand, $\text{LID}_*(x)$ is hard to ascertain without the context clues discussed in Section 4.)

Diffusion Model on a von Mises Distribution In an illustrative experiment, we study a von Mises distribution which sits on a 1-dimensional circle in the 2-dimensional plane. Because its support is 1-dimensional, every point x in the support has $\text{LID}_*(x) = 1$. From this distribution we sample 100 training points; both the training points and the ground truth density are depicted in Figure 3. By chance,

[2] This is captured formally by the regular level set theorem of differential geometry (Lee, 2012).



(a) Ground truth manifold. (b) Model samples.

Figure 3: Training a diffusion model on a von Mises distribution.

a single point x_0 sits isolated in a low-density region of the circle, on the upper-left side. Next, we train a DM on this data. Pidstrigach (2022) shows that DMs are capable of learning manifolds of any dimensionality, and we find this to be precisely the case. In Figure 3b we depict 100 samples generated by the DM, colour-coded by their FLIPD estimates (which are scalar-valued). Remarkably, the DM reproduces the isolated sample near-exactly and assigns it $\text{LID}_\theta(x_0) = 0$ despite assigning $\text{LID}_\theta(x) = 1$ to nearly all of its other generated samples. From this experiment we infer that, at least in a simplified setting, memorization does indeed coincide with low LIDs as measured by FLIPD.

Stable Diffusion on LAION, COCO, and Tuxemon For this example, we set $p_\theta(x)$ to Stable Diffusion v1.5 (Rombach et al., 2022). Taking inspiration from the benchmark of Wen et al. (2023), we retrieve memorized LAION (Schuhmann et al., 2022) training images identified by Webster (2023). We focus on the 86 memorized images categorized as "matching verbatim", noting that the other categories of Webster (2023) consist of large numbers of captions that generate samples matching a small set of training images. For non-memorized images, we use a mix of 2000 images sampled from LAION Aesthetics 6.5+, 2000 sampled from COCO (Lin et al., 2014), and all 251 images from the Tuxemon dataset (Tuxemon Project, 2024; Hugging Face, 2024).

We compute FLIPD values for each of the aforementioned images. Note that Stable Diffusion provides two model distributions: the unconditional distribution $p_\theta(x)$ and the conditional distribution $p_\theta(x \mid c)$. Density histograms of FLIPD values with respect to both are depicted in Figure 4. We compare our procedure to a variant of the classifier-free guidance (CFG) norm used to detect memorization by Wen et al. (2023) discussed in more detail in Appendix A.

Conditional FLIPD, unconditional FLIPD, and the CFG norm are strong signals of memorization. When used as a score to differentiate memorized from non-memorized samples, their AUROCs are uniformly in the high 90's. Interestingly, the unconditional LID detects memorization well despite the lack of caption information. Detecting memorized training images without the corresponding captions is a novel capability, and notably cannot be done with the CFG norm technique. One might reasonably be concerned that, because LID measures complexity and simple images are more likely to be memorized, the complexity of images may be confounding the results. We address these concerns by juxtaposing our memorization metrics with a quantitative measurement of complexity: the PNG compression size. Figure 4d shows that complexity is a comparatively poor predictor of memorization, attenuating concerns that LID_θ only detects memorization through complexity. This finding is reinforced by results on the Tuxemon dataset, which is less complex (measured both qualitatively and with PNG compression size) but for which we measure higher LIDs.

The LID estimates provided by FLIPD are sometimes negative in value; Kamkari et al. (2024b) justify this as an artefact of estimating the LID using a UNet. Despite underestimating LID in absolute terms, Kamkari et al. (2024b) confirm that FLIPD ranks LID_θ estimates correctly, which is sufficient for the purpose of differentiating memorized from non-memorized examples.

4. Explaining Memorization Phenomena

In this section we explain memorization phenomenona described in related work from the perspective of LID. For additional related work, please see Appendix B. For formal theorem statements and proofs, please see Appendix C.

Duplicated Data and LID It has been broadly observed that memorization occurs when training points are duplicated (Nichol et al., 2022; Carlini et al., 2022; Somepalli et al., 2023a). In Theorem 4.1, we show that datapoint duplication is an example of DD-Mem; duplicated points x_0 indicate $\text{LID}_*(x_0) = 0$, so even a model with good generalization will have $\text{LID}_\theta(x_0) = 0$.

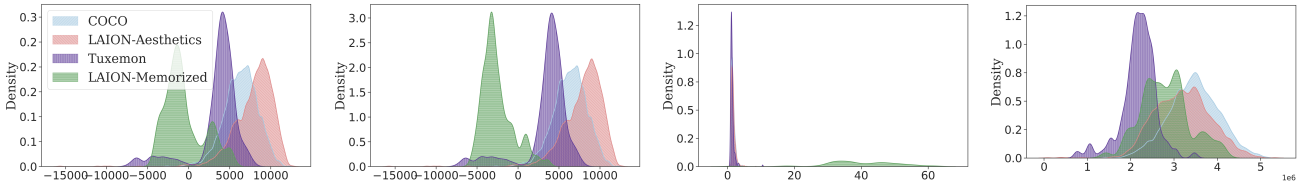
Theorem 4.1 (Informal)**.** *Let $\{x_i\}_{i=1}^n$ be a training dataset independently drawn from $p_*(x)$. Under some regularity conditions, the following hold:*

1. *If duplicates occur in $\{x_i\}_{i=1}^n$ with positive probability, then they occur at a point x_0 such that $\text{LID}_*(x_0) = 0$.*
2. *If $\text{LID}_*(x_0) = 0$ and n is sufficiently large, then duplication will occur in $\{x_i\}_{i=1}^n$ with near-certainty.*

Proof. See Appendix C. □

From this result, we gather that improving model generalization is unlikely to solve this form of memorization unless improvements specifically add inductive biases that prevent $p_\theta(x)$ from learning 0-dimensional points.

We postulate that a similar result applies to "near-duplicated content," in which many similar but non-identical points occur together in the dataset, and in this case LID_* is low but nonzero in the region of the duplicated content.



(a) Unconditional LID (AUROC = 97.6%). (b) Conditional LID (AUROC = 99.2%). (c) CFG norm (AUROC = 99.3%). (d) PNG compression (AUROC = 70.3%).

Figure 4: Density histograms for each memorization metric on the different datasets.

Conditioning and LID Somepalli et al. (2023b) and Yoon et al. (2023) observe that specific conditioning encourages the generation of memorized samples. Here, we point out that conditioning decreases LID, making models more likely to generate memorized samples.

Proposition 4.2 (Informal)**.** *Let $x_0 \in \mathcal{M}_\theta$, and let us denote by $\mathit{LID}_\theta(x_0 \mid c)$ the LID of the conditional distribution $p_\theta(x \mid c)$ at the point x_0. We then have*

$$\mathit{LID}_\theta(x_0 \mid c) \leq \mathit{LID}_\theta(x_0). \tag{1}$$

Proof. See Appendix C for the formal proof. Intuitively, conditioning can be interpreted as adding an additional constraint to the manifold $\mathcal{M}_\theta$ of generated data. For example, one can constrain Stable Diffusion samples stylistically by conditioning: $p_\theta(x \mid$ "photorealistic"$)$. Any meaningful constraint will reduce the model manifold ($\mathcal{M}_{\theta|c} \subsetneq \mathcal{M}_\theta$), in which case LID can only decrease or stay the same. □

Using LID as a heuristic notion for the complexity of a datapoint, Proposition 4.2 can be understood as a "local" analogue of the fact that entropy bounds conditional entropy: $H(x \mid c) \leq H(x)$ (a distribution-level notion).

The Classifier-Free Guidance Norm and LID Classifier-free guidance (CFG) is a way to improve the quality of conditional generation. Whereas standard conditional generation would employ the score function $s_\theta(x; t, c)$, which refers to a neural estimate at time t of the conditional score, CFG increases the strength of conditioning by using the following modified score:

$$s_\theta^{\text{CFG}}(x; t) = s_\theta(x; t, \emptyset) + \alpha(\underbrace{s_\theta(x; t, c) - s_\theta(x; t, \emptyset)}_{\text{CFG vector}}), \tag{2}$$

where α is a hyperparameter for "guidance strength" and $s_\theta(x; t, \emptyset)$ refers to conditioning on the empty string (here we formulate diffusion models using stochastic differential equations (Song et al., 2021)).

Following on from Proposition 4.2, Wen et al. (2023) identify that specific conditioning inputs c generate memorized samples when the CFG vector has a large magnitude. For our context, it is sufficient to observe that a large CFG magnitude will generally result in a large magnitude of the CFG score $s_\theta^{\text{CFG}}(x; t)$.

It is understood in the literature that large $\|s_\theta^{\text{CFG}}(x; t)\|$, and its explosion as $t \to 0$, is common for high-dimensional data (Vahdat et al., 2021) and necessary to generate samples from low-dimensional manifolds (Lu et al., 2023). It has been empirically observed that this explosion occurs faster as the dimensionality gap increases between $\mathcal{M}_\theta$ and the ambient data space, which is one reason that diffusion modelling on lower-dimensional latent space improves performance (Loaiza-Ganem et al., 2022; 2024). The largest $\|s_\theta^{\text{CFG}}(x; t)\|$ values should thus generate points with the largest dimensionality difference from $\mathbb{R}^d$; i.e., points with the smallest LID. Hence we infer that controlling the score norm should increase $\text{LID}_\theta(x)$ and reduce memorization, a fact confirmed empirically by Wen et al. (2023).

Complexity and LID Somepalli et al. (2023b) also highlight image complexity as a potential factor in determining memorization. Using the heuristic understanding that LID corresponds to complexity as proposed in Section 2, we infer that low-complexity images $x \in \mathcal{M}_*$ have low $\text{LID}_*(x)$. This fact suggests that, like with duplication, memorization of low-complexity images is an example of DD-Mem.

5. Conclusions

Throughout this work, we have drawn connections between the geometry of a DGM and its propensity to memorize. First, we showed that the notion of LID provides a systematic way of understanding different types of memorization. Second, in experiments, LID proved to be a promising way to detect memorization. Third, we explained how memorization phenomena described by prior work can be understood from the perspective of LID. We offered several connections, including the insight that some instances of memorization in DMs are due *not* to the DM's inability to generalize (OD-Mem), but rather to low-LID ground truth (DD-Mem).

Having demonstrated the utility of our geometric framework, we expect LID to be a useful lens to direct future research in DGMs. For example, controlling LID might be a more effective means of preventing memorization in DGMs than studying their generalization directly. Although the manifold hypothesis does not apply directly to discrete data such as language, some intuition described in this work carries over, and generalizations or parallels to the concepts here may offer insights for the language-modelling space.

References

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

Bradski, G. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

Brown, B. C., Caterini, A. L., Ross, B. L., Cresswell, J. C., and Loaiza-Ganem, G. Verifying the union of manifolds hypothesis for image data. In *International Conference on Learning Representations*, 2023.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2022.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Chen, C., Liu, D., and Xu, C. Towards memorization-free diffusion models. *arXiv preprint arXiv:2404.00922*, 2024.

Dai, B. and Wipf, D. Diagnosing and enhancing VAE models. In *International Conference on Learning Representations*, 2019.

Daras, G., Dimakis, A. G., and Daskalakis, C. Consistent diffusion meets tweedie: Training exact ambient diffusion models with noisy data. *arXiv preprint arXiv:2404.10177*, 2024.

Dinh, L., Krueger, D., and Bengio, Y. NICE: Non-linear independent components estimation. *arXiv:1410.8516*, 2014.

Folland, G. *Real Analysis: Modern Techniques and Their Applications*. Wiley, 2013.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Gu, X., Du, C., Pang, T., Li, C., Lin, M., and Wang, Y. On memorization in diffusion models. *arXiv preprint arXiv:2310.02664*, 2023.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Horvat, C. and Pfister, J.-P. Intrinsic dimensionality estimation using normalizing flows. In *Advances in Neural Information Processing Systems*, 2022.

Horvat, C. and Pfister, J.-P. On gauge freedom, conservativity and intrinsic dimensionality estimation in diffusion models. *arXiv preprint arXiv:2402.03845*, 2024.

Hugging Face. Tuxemon. https://huggingface.co/datasets/diffusers/tuxemon, 2024.

Kadkhodaie, Z., Guth, F., Simoncelli, E. P., and Mallat, S. Generalization in diffusion models arises from geometry-adaptive harmonic representation. In *The Twelfth International Conference on Learning Representations*, 2023.

Kamkari, H., Ross, B. L., Cresswell, J. C., Caterini, A. L., Krishnan, R. G., and Loaiza-Ganem, G. A geometric explanation of the likelihood OOD detection paradox. In *International Conference on Machine Learning*, 2024a.

Kamkari, H., Ross, B. L., Hosseinzadeh, R., Cresswell, J. C., and Loaiza-Ganem, G. A geometric view of data complexity: Efficient local intrinsic dimension estimation with diffusion models. *arXiv:2406.03537*, 2024b.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Lee, J. M. *Introduction to Smooth Manifolds*. Springer New York, 2012.

Levina, E. and Bickel, P. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems*, 2004.

Li, P., Li, Z., Zhang, H., and Bian, J. On the generalization properties of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2023.

Li, S., Chen, S., and Li, Q. A good score does not lead to a good generative model. *arXiv preprint arXiv:2401.04856*, 2024.

Li, X., Thickstun, J., Gulrajani, I., Liang, P. S., and Hashimoto, T. B. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Loaiza-Ganem, G., Ross, B. L., Cresswell, J. C., and Caterini, A. L. Diagnosing and fixing manifold overfitting in deep generative models. *Transactions on Machine Learning Research*, 2022.

Loaiza-Ganem, G., Ross, B. L., Hosseinzadeh, R., Caterini, A. L., and Cresswell, J. C. Deep generative models through the lens of the manifold hypothesis: A survey and new connections. *arXiv:2404.02954*, 2024.

Lu, Y., Wang, Z., and Bal, G. Mathematical analysis of singularities in the diffusion model under the submanifold assumption. *arXiv:2301.07882*, 2023.

Meehan, C., Chaudhuri, K., and Dasgupta, S. A non-parametric test to detect data-copying in generative models. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Nichol, A., Ramesh, A., Mishkin, P., Dariwal, P., Jang, J., and Chen, M. DALL·E 2 Pre-Training Mitigations, June 2022. URL https://openai.com/blog/dall-e-2-pre-training-mitigations/.

Orrick, W. H. Andersen v. stability ai ltd., 2023. URL https://casetext.com/case/andersen-v-stability-ai-ltd.

Pidstrigach, J. Score-based generative models detect manifolds. In *Advances in Neural Information Processing Systems*, 2022.

Pizzi, E., Roy, S. D., Ravindra, S. N., Goyal, P., and Douze, M. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14532–14542, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294, 2022.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265, 2015.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6048–6058, 2023a.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Understanding and mitigating copying in diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 47783–47803, 2023b.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Stanczuk, J., Batzolis, G., Deveney, T., and Schönlieb, C.-B. Your diffusion model secretly knows the dimension of the data manifold. *arXiv:2212.12611*, 2022.

Stein, G., Cresswell, J., Hosseinzadeh, R., Sui, Y., Ross, B., Villecroze, V., Liu, Z., Caterini, A. L., Taylor, E., and Loaiza-Ganem, G. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Tempczyk, P., Michaluk, R., Garncarek, L., Spurek, P., Tabor, J., and Golinski, A. LIDL: Local intrinsic dimension estimation using approximate likelihood. In *International Conference on Machine Learning*, pp. 21205–21231, 2022.

Tuxemon Project. Tuxemon, 2024. URL https://tuxemon.org/.

Vahdat, A., Kreis, K., and Kautz, J. Score-based generative modeling in latent space. In *Advances in Neural Information Processing Systems*, 2021.

Vyas, N., Kakade, S. M., and Barak, B. On provable copyright protection for generative models. In *International Conference on Machine Learning*, pp. 35277–35299. PMLR, 2023.

Webster, R. A reproducible extraction of training images from diffusion models. *arXiv preprint arXiv:2305.08694*, 2023.

Webster, R., Rabin, J., Simon, L., and Jurie, F. This person (probably) exists. identity membership attacks against gan generated faces. *arXiv preprint arXiv:2107.06018*, 2021.

Wen, Y., Liu, Y., Chen, C., and Lyu, L. Detecting, explaining, and mitigating memorization in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023.

Yi, M., Sun, J., and Li, Z. On the generalization of diffusion model. *arXiv preprint arXiv:2305.14712*, 2023.

Yoon, T., Choi, J. Y., Kwon, S., and Ryu, E. K. Diffusion probabilistic models generalize when they fail to memorize. In *ICML 2023 Workshop on Structured Probabilistic Inference {\&} Generative Modeling*, 2023.

Zhang, H., Zhang, J., Srinivasan, B., Shen, Z., Qin, X., Faloutsos, C., Rangwala, H., and Karypis, G. Mixed-type tabular data synthesis with score-based diffusion in latent space. *arXiv preprint arXiv:2310.09656*, 2023.

Zheng, Y., He, T., Qiu, Y., and Wipf, D. P. Learning manifold dimensions with conditional variational autoencoders. In *Advances in Neural Information Processing Systems*, 2022.

A. Experimental Details

A.1. Experimental Details

Here we provide additional experimental details for the Stable Diffusion experiment in Section 3. For all methods, we "tune" the associated hyperparameters by hand to get the best AUROC performance, but anecdotally their performance is stable (and good) across all sufficiently small timesteps ($t \leq 0.1$ under the SDE framework of Song et al. (2021) wherein $t \in [0, 1]$).

FLIPD For both conditional and unconditional FLIPD, we use a timestep $t = 0.1$ and a single Hutchinson sample to estimate the trace component.

CFG Norm Note that while Wen et al. (2023) use the generation process to measure whether a synthesized image has been memorized, we focus here on detecting whether real, training-set images have been memorized, which requires some methodological changes. To compute a memorization score, we take k Euler steps forward using the conditional score $s_\theta(x; t, c)$ with the probability flow ODE (Song et al., 2021) until time t_0 to get a point at $x_0 \in \mathbb{R}^d$. We then compute the CFG norm $\|s_\theta(x_0; t_0, c) - s_\theta(x; t_0, \emptyset)\|$. We use timestep $t_0 = 0.01$ and 3 Euler steps.

PNG Compression Size We use the maximum compression level of 9 with the cv2 package (Bradski, 2000).

B. Related Work

Here we briefly summarize some additional literature on LID and memorization in the context of diffusion models.

Detecting and Preventing Memorization for Image Models A number of authors have studied the task of surfacing memorized training examples from generated ones, finding that (i) L2 distance in pixel space works poorly (Carlini et al., 2023), (ii) recalibrating L2 in various ways works better for detecting images in small datasets such as CIFAR-10 (Carlini et al., 2023; Yoon et al., 2023; Stein et al., 2023), and (iii) using retrieval techniques such as distance in SSCD feature space (Pizzi et al., 2022) works better still (Somepalli et al., 2023a). However, retrieval techniques are generally too expensive to be deployed in conjunction with a live model. To this end, mechanistic detection techniques as well as guidance and training techniques have been proposed (Wen et al., 2023; Chen et al., 2024; Daras et al., 2024).

Explaining Memorization There is an active field of work attempting to explain why and how memorization occurs in DMs. Li et al. (2024) show that memorization can occur even in the presence of generalization. This claim is superficially similar to our definition of data-driven memorization, which we assert can occur in the presence of perfect generalization, but our work rests on a different, geometric definition of generalization. Both DD-Mem and the assertions of Li et al. (2024) would appear to conflict with the thesis and findings of Yoon et al. (2023) that diffusion models only generalize when they fail to memorize. However, the latter actually focuses on what we refer to as *overfitting driven memorization*, in which case their work agrees with our framework. Further work has suggested both that diffusion models do (Kadkhodaie et al., 2023; Li et al., 2023) or do not (Yi et al., 2023) generalize well under different assumptions and contexts.

DGM-Based LID Estimation As opposed to classical LID estimators (e.g., Levina & Bickel (2004)), which are constructed to estimate the dimension of $\mathcal{M}_*$, DGM-based LID estimation estimates the dimensionality of $\mathcal{M}_\theta$, the manifold learned by a DGM. These types of estimators are available for many types of DGMs, and in addition to being useful for memorization, have found utility in out-of-distribution detection (Kamkari et al., 2024a). In the literature, LID estimators for normalizing flows (Dinh et al., 2014) have been proposed using the singular values of their Jacobians (Horvat & Pfister, 2022; Kamkari et al., 2024a) or their density estimates (Tempczyk et al., 2022). Dai & Wipf (2019) and Zheng et al. (2022) proposed estimators for VAEs using the structure of their posterior distribution. Several authors have proposed estimators for DMs as well (Stanczuk et al., 2022; Horvat & Pfister, 2024); we focus on that of Kamkari et al. (2024b) because it is the most computationally tractable.

C. Proofs

We restate each theorem in full formality below along with their proofs.

Throughout this section, we let P_* and P_θ be the probability measures of the ground truth data and model, respectively. We assume that the respective supports of P_* and P_θ are $\mathcal{M}_*, \mathcal{M}_\theta \subset \mathbb{R}^d$, Riemannian submanifolds of the Euclidean space $\mathbb{R}^d$ with metrics g_* and g_θ respectively. As mentioned in Section 2, we take a lax definition of manifold which allows them to vary in dimensionality in different components. A single manifold under our definition is equivalent to a disjoint union of manifolds under the more standard definition.

C.1. Theorem 4.1

For the purpose of this theorem, we assume that P_* admits a continuous, real-valued probability density $p_*(x)$ with respect to the Riemannian measure on $\mathcal{M}_*$, which we denote by $\mu_{\mathcal{M}_*}$. This assumption, though standard (Loaiza-Ganem et al., 2022; Tempczyk et al., 2022), has a strong impact on this theorem. Dropping this assumption would invalidate Theorem 4.1 by allowing, roughly speaking, for point masses belonging to higher-dimensional components $\mathcal{M}_*$ on which duplication can occur. However, even if the assumption does not hold, the theorem arguably still holds on an intuitive level because the point mass is a "0-dimensional object" even if it belongs to a higher-dimensional subset of $\mathcal{M}_*$.

Lemma C.1. *Let $x_0 \in \mathcal{M}_*$. The following are equivalent:*

(a) $P_*(\{x_0\}) > 0$, *and*

(b) $\mathit{LID}(x_0) = 0$.

Proof.

$(a) \implies (b)$ Assume $P_*(\{x_0\}) > 0$.

$$0 < P_*(\{x_0\}) = \int_{\{x_0\}} p_*(x) \mathrm{d}\mu_{\mathcal{M}_*}(x), \tag{3}$$

which necessitates $\mu_{\mathcal{M}_*}(\{x_0\}) > 0$. If we had $\mathrm{LID}_*(x_0) > 0$ this would incur a contradiction: letting (U, ϕ) be a chart around x_0, then by the definition of $\mu_{\mathcal{M}_*}$,

$$0 < \mu_{\mathcal{M}_*}(\{x_0\}) = \int_{\phi(\{x_0\})} \sqrt{\det(g_*)} d\lambda, \tag{4}$$

where λ is the Lebesgue measure on $\mathbb{R}^{\mathrm{LID}_*(x_0)}$. Due to the singleton domain of integration, this would be impossible *unless* $\mathrm{LID}_*(x_0) = 0$, in which case $\mu_{\mathcal{M}_*}$ would instead be the counting measure by convention.

$(b) \implies (a)$ Suppose $\mathrm{LID}(x_0) = 0$. This implies that $\{x_0\}$ is an open set in the subspace topology of $\mathcal{M}_*$, meaning there exists an open set $V \subset \mathbb{R}^d$ such that $V \cap \mathcal{M}_* = \{x_0\}$. But since $x_0 \in \mathrm{supp} P_*$, any open set containing x_0 must have positive probability. Moreover, since $V \setminus \{x_0\} = V \setminus \mathcal{M}_*$ is an open set with no intersection with the support, $P_*(V \setminus \mathcal{M}_*) = 0$. So $P_*(\{x_0\}) = P_*(V \cap \mathcal{M}_*) = P_*(V \cap \mathcal{M}_*) + P_*(V \setminus \mathcal{M}_*) = P_*(V) > 0$.

□

Theorem C.2 (Formal Restatement of Theorem 4.1)**.** *Let $\{x_i\}_{i=1}^n$ be a training dataset drawn independently from $p_*(x)$.*

1. *If duplicates occur in $\{x_i\}_{i=1}^n$ with positive probability, then they will occur at a point x_0 such that $\mathit{LID}_*(x_0) = 0$.*
2. *If $\mathit{LID}_*(x_0) = 0$ then the probability of duplication in $\{x_i\}_{i=1}^n$ will converge to 1 as $n \to \infty$.*

Proof. 1. Due to Lemma C.1, it suffices to show that any duplicates in $\{x_i\}_{i=1}^n$ must occur at a point x_0 such that $P_*(\{x_0\}) > 0$. Equivalently, we can show that if $P_*(\{x_0\}) = 0$ for every x_0, then $P_*(x_1 = x_2) = 0$. Assume that $P_*(\{x_0\}) = 0$ for every $x_0 \in \mathcal{M}_*$. Since x_1 and x_2 are independent, $P_*(x_1 = x_2) = P_* \times P_*(D)$, where $D = \{(x, x) \in \mathcal{M}_* \times \mathcal{M}_* \mid x \in \mathcal{M}_*\}$. We then have:

$$P_* \times P_*(D) = \int_D \mathrm{d}P_* \times P_*(x_1, x_2) = \int_{\mathcal{M}_*} \int_{\{x_2\}} \mathrm{d}P_*(x_1) \mathrm{d}P_*(x_2) = \int_{\mathcal{M}_*} P_*(\{x_2\}) \mathrm{d}P_*(x_2) = 0, \tag{5}$$

where the second equality follows from a standard result in measure theory (see e.g. Theorem 7.26 in Folland (2013)), and the last equality follows by assumption. This finishes this part of the proof.

2. Suppose $\text{LID}_*(x_0) = 0$. By Lemma C.1, we have $P_*(\{x_0\}) > 0$. In this case, $P_*(\{x_i = x_0\}) > 0$ for all $i \in \{1, \dots, n\}$, meaning that

$$\begin{aligned} P_*(\{x_i = x_j \text{ for some } i, j \geq 1, i \neq j\}) &\geq P_*(\{x_i = x_j = x_0 \text{ for some } i, j \geq 1, i \neq j\}) && (6)\\ &\geq 1 - P_*(\{x_i \neq x_0 \text{ for } i \geq 2) && (7)\\ &= 1 - P_*(\{x_2 \neq x_0\}) \cdots P_*(\{x_n \neq x_0\}) && (8)\\ &= 1 - (1 - P_*(\{x_0\}))^{n-1} && (9)\\ &\longrightarrow 1\,, && (10) \end{aligned}$$

where the last line depicts the limiting behaviour as $n \to \infty$.

□

C.2. Proposition 4.2

Here we presume the joint distribution of model samples and k-dimensional conditioning inputs $(x, y) \in \mathbb{R}^{d+k}$ has support $S \subset \mathbb{R}^{d+k}$ such that $\{x : (x, c) \in S \text{ for some } c \in \mathbb{R}^k\} = \mathcal{M}_\theta$. We define the *conditional support* of x given c to be $S(c) = \{x : (x, c) \in S\}$.

Proposition C.3 (Formal)**.** *Let $x_0 \in \mathcal{M}_\theta$ and $c \in \mathbb{R}^k$. Suppose that $S(c)$ is also a submanifold of $\mathbb{R}^d$ and denote its LID by $\mathit{LID}_\theta(x_0 \mid c)$. We then have*

$$\mathit{LID}_\theta(x_0 \mid c) \leq \mathit{LID}_\theta(x_0). \tag{11}$$

Proof. If $S(c)$ is a submanifold of $\mathbb{R}^d$, then it is also a submanifold of $\mathcal{M}_\theta$. The inequality follows directly. □