Local Coverage Governs Memorization in Diffusion Models

Claudia Merger [1] **Sebastian Goldt** [1]

Abstract

Diffusion models are known to memorize training data, but which samples are most likely to be memorized? While memorization is often treated as a global property, in practice diffusion models simultaneously generate both memorized and novel samples. In this work, we show that memorization is governed by local data coverage. Leveraging the connection between diffusion models and kernel density estimation (KDE), we derive a theoretical criterion that predicts whether a point is memorized or generalized based on the density of training data in its neighborhood and the overall sample complexity. In the high-dimensional limit, this leads to a sharp local transition: regions of low coverage are dominated by isolated training samples (memorization), while dense regions support interpolation and generalization. We validate these predictions empirically, showing that memorization increases with local sparsity and that diffusion models exhibit a coexistence of memorized and novel samples within the same model. Extending this framework to multi-class settings, we further show that classes with higher intra-class diversity (and thus lower local coverage) are more strongly memorized. Our results provide a unified, local view of memorization in diffusion models, explaining when and where memorization occurs in terms of data geometry.

1. Introduction

How does memorization arise *locally* in diffusion models? Diffusion models are powerful generative methods capable of learning complex high-dimensional data distributions from finite datasets. Yet when trained with limited data, they may reproduce training examples rather than generate novel samples, a phenomenon commonly referred to as *memorization* (Somepalli et al., 2022; Carlini et al., 2023; Kadkhodaie et al., 2023). Empirically, memorization decreases as the number of training samples increases, suggesting a fundamental relationship between finite-sample effects and generative generalization.

Theoretical work has begun to analyze this phenomenon through the close connection between diffusion models and kernel density estimation (KDE) (Pham et al., 2024; Ambrogioni, 2023; Biroli & Mézard, 2024; Achilli et al., 2025a; Lucibello & Mézard, 2024). In this picture, each training sample contributes a local kernel, and generalisation arises from the superposition of many such kernels, see Figure 1a for a sketch. When kernels overlap strongly, the model assigns finite probability to the space between examples; when overlap is weak, generation can collapse onto individual training points. These analyses predict a global phase transition at a critical sample complexity, where the model changes from memorizing training examples to generating novel samples. Recent work has also gone beyond this global perspective. First, Achilli et al. (2026) showed that memorization may emerge progressively through the loss of manifold dimensions, leading to a form of *geometric memorization* in which some directions of variability are lost before exact copying occurs. Second, Garnier-Brun et al. (2026) showed a coexistence of memorization and generalization in early-stopped diffusion models. This suggests that memorization is richer than a single abrupt transition.

Additionally, empirical observations indicate an additional level of heterogeneity that remains theoretically unexplained. Memorizing diffusion models often reproduce only a subset of their training data rather than the entire dataset (Fang et al., 2025; Carlini et al., 2023). Memorization is also known to depend on guidance strength, conditioning, architecture, and dataset structure (Somepalli et al., 2023; Kim et al., 2025; Gu et al., 2025; Yoon et al., 2023). Thus, at finite sample size, a diffusion model often generates a mixture of copied and novel samples see Figure 1 c) f) and i). This raises a natural question:

Which regions of data space, classes, or individual samples are most likely to be memorized?

Because memorized samples may expose private, copyrighted, or otherwise sensitive training data, predicting

[1]SISSA, Trieste, Italy. Correspondence to: Claudia Merger <cmerger@sissa.it>, Sebastian Goldt <sgoldt@sissa.it>.

Published at ICML 2026 Workshop on the Impact of Memorization on Trustworthy Foundation Models, Seoul, South Korea. PMLR 306, 2026.

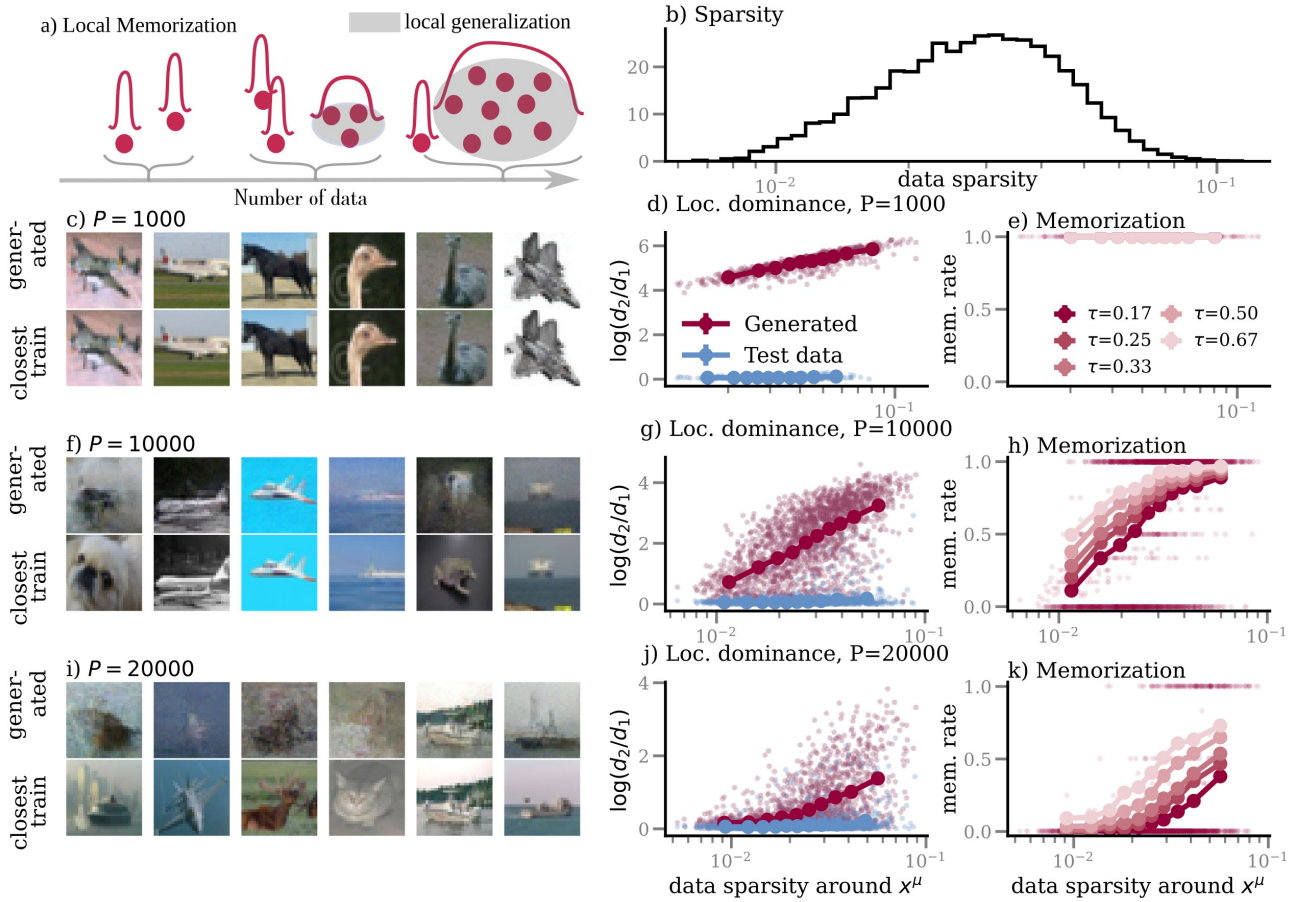

Figure 1. **Local memorization.** a) Sketch of kernel density approximation and local memorization phenomenon. b) Distribution of data sparsity for 20,000 samples from CIFAR-10. c) Generated samples + closest training image (measured by cosine similarity) for diffusion model trained on $P = 1000$ training examples. d) Local dominance as a function of sparsity. For each generated sample $\hat{x}$, we compute distances d_1 and d_2 to its nearest and second-nearest training samples, and define dominance via their ratio. Each point corresponds to a training sample x^μ, aggregating generated samples assigned to it. Scatter shows raw values, markers indicate binned averages. Blue points show the same quantity computed using test data, providing a baseline without memorization. e) Memorization rate per training sample x^μ, defined as the fraction of generated samples satisfying $d_1/d_2 < \tau$. Scatter shows raw values for $\tau = 0.1$, markers indicate binned averages across thresholds. f)–k) Same measurements for models trained on larger datasets. As the number of training samples increases, both dominance and memorization decrease, and their dependence on sparsity weakens. Overall, sparse regions exhibit strong single-sample dominance and higher memorization, while dense regions promote interpolation across multiple training points.

where memorization occurs provides a tool for auditing and mitigating risks in generative models. In this work, we give a concrete criterion for memorization by introducing a *local* theory of memorization in diffusion models. Our central hypothesis is simple: memorization is governed not only by the total number of training samples, but by their *local coverage*. Regions of data space containing many nearby examples support interpolation and generalization, whereas isolated regions are prone to sample retrieval. In Figure 1 we show an example of this behavior for diffusion models with U-net architecture (Ronneberger et al., 2015) trained on subsets of the CIFAR-10 image dataset (Krizhevsky, 2009). We find that more isolated training data are preferentially memorized, whereas data points in regions of higher local density (lower local sparsity in Figure 1) are memorized less.

To formalize this intuition, we extend the KDE framework (Lucibello & Mézard, 2024; Biroli & Mézard, 2024; Achilli et al., 2025b) to a local setting. We quantify the local coverage around a point $x \in \mathbb{R}^N$ by the probability mass contained in a ball of radius h around x,

$$p_{\text{in}}(x) = \int_{B_h(x)} d\rho(x'), \tag{1}$$

where $B_h(x)$ denotes a N-dimensional hypersphere centered at x and ρ the density of the data from which the training points are drawn. This quantity measures how densely the data distribution ρ populates the neighborhood of x. In the high-dimensional regime in which the number

of samples scales as $P = e^{\alpha N}$, we define

$$\nu_{\text{in}}(x) = \lim_{N\to\infty} \tfrac{1}{N} \ln p_{\text{in}}(x)\,. \tag{2}$$

Our main result shows that memorization at a point x is controlled by the pair $(\nu_{\text{in}}(x), \alpha)$ with a sharp transition:

$$\begin{cases} x \in A_{\text{mem}} & \text{if } \nu_{\text{in}}(x) < -\alpha \\ x \in A_{\text{gen}} & \text{otherwise} \end{cases} \tag{3}$$

where we refer to A_{gen} as the region where the model correctly interpolates, and A_{mem} is a region where the learned density is characterized by isolated peaks centered on training examples. This result shows that depending on local coverage, the same diffusion model may exhibit retrieval-like behavior in some regions of space and generative interpolation in others. Points in low-coverage regions ($\nu_{\text{in}}(x)$ small) are memorized, while points in high-coverage regions are generalized. This local theory explains several empirical phenomena. Most crucially, it predicts a **coexistence of memorization and generalization:** copied and novel samples can arise simultaneously from the same model. Moreover, for models trained on multimodal data, it explains **Class-dependent memorization:** classes with larger intra-class diversity (and therefore lower local coverage) are memorized more strongly. We validate these predictions on diffusion models trained on standard image datasets such as CIFAR-10 and CelebA (Liu et al., 2015). Consistent with theory, training examples originating from regions of lower density, as well as classes with larger intra-class diversity are memorized more strongly.

2. Background

We briefly recall a key observation: diffusion models trained on finite data behave like KDE. In brief, diffusion models learn the score $\nabla \ln \rho_t(x)$ of a random variable $x_t \sim \rho_t$ that has been corrupted by a noising process.

$$x_t(\epsilon_t) = \sqrt{\bar\alpha_t} x_0 + \sqrt{1-\bar\alpha_t}\epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, \mathrm{Id})\ , \tag{4}$$

where $\bar\alpha_t \in (0,1)$ is a decreasing function in t. To make the connection to KDE explicit, consider an idealized setting in which the model has infinite capacity and is trained to optimality. In this case, for each t, the learned function $\epsilon_\theta(\cdot, t)$ can be treated as an arbitrary function $\epsilon(\cdot, t)$ and the loss can be minimized functionally. Setting this functional derivative to zero yields:

$$\epsilon(x,t) \propto \nabla_x \ln \textstyle\sum_\mu \mathcal{N}(x; \bar\alpha_t x^\mu, (1-\bar\alpha_t)\mathrm{Id})\ . \tag{5}$$

This identity reveals that, when trained to convergence on a finite number of samples, diffusion models learn a Gaussian mixture centered at the training data points. In other words, diffusion models trained on a finite dataset recover the score of the empirical distribution convolved with Gaussian noise, rather than the true underlying density ρ. In the low-noise limit ($\bar\alpha_t \to 1$), this mixture becomes sharply peaked around individual training samples, and the model effectively reproduces them. At finite noise, there is a direct correspondence between Equation (5) and KDE. KDE attempts to approximate a density ρ using a mixture of kernel functions K centered at the data points:

$$Z(x) = \tfrac{1}{P} \textstyle\sum_\mu K\left(\frac{|x-x^\mu|}{h}\right)\ . \tag{6}$$

The correspondence with KDE becomes explicit by identifying the kernel as Gaussian with bandwidth $h = \sqrt{1-\bar\alpha_t}$ up to a rescaling of the variables x, x^μ by $\sqrt{\bar\alpha_t}$. In the ideal case, one has sufficient number of samples P such that Z becomes a smooth approximation of ρ. This behavior with P is one hypothesis to explain memorization and generalization in diffusion models.

This perspective suggests that generalization is not uniform across the data space. Even when the total number of samples is large, the quality of the KDE approximation depends on local sample density. Regions with high data density are well approximated, while sparse regions remain dominated by individual kernels. In the following, we formalize this intuition by deriving conditions under which local regions exhibit memorization or generalization.

3. Results

3.1. Kernel density estimation with hard spheres

To study kernel density estimation (KDE), we will now characterize the behavior of the log-density

$$z(x) = \tfrac{1}{N} \ln Z(x), \tag{7}$$

where $Z(x)$ is the mixture of kernel functions K centered at the data points defined in Equation (6). The difficulty in establishing the behavior of z is that it is a random quantity that depends on the draw of the P training points from ρ. Our goal is to compute its distribution over such draws to derive general properties of high dimensional KDE. To this end, we restrict ourselves to one particular kernel, namely

$$K(x) = \Theta(R^2 - x^2)/V_R \tag{8}$$

where V_R is the volume of the N-dimensional sphere with radius R. This choice of kernel allows us to express our results in a particularly simple form. In high dimensions, we expect that this choice of kernel is qualitatively equivalent to a Gaussian one with standard deviation R, as the mass of both distributions concentrates near a thin shell at radius R when $N \to \infty$.

In Section B, we give an overview of our method and additional results on general kernels K. Detailed derivations are

found in Section C. Assuming that $\nu_{\text{in}}(x)$ remains $\mathcal{O}(1)$ as $N \to \infty$, we obtain the following sharp characterization of $z(x)$:

$$z(x) \to \begin{cases} \nu_{\text{in}}(x) & \text{if } \nu_{\text{in}}(x) \geq -\alpha \\ -\infty & \text{if } \nu_{\text{in}}(x) < -\alpha \end{cases}. \tag{9}$$

The divergence $z(x) \to -\infty$ for $\nu_{\text{in}}(x) < -\alpha$ reflects the fact that, in these regions, the kernel density estimate vanishes with high probability. This result shows that the behavior of $z(x)$ is controlled by the comparison between the local coverage $\nu_{\text{in}}(x)$ and the sample complexity α, establishing a sharp dichotomy between two regimes. In regions where $\nu_{\text{in}}(x) \geq -\alpha$, the kernel density estimate provides a smooth approximation of the underlying distribution, corresponding to a generalization regime. In contrast, in regions where $\nu_{\text{in}}(x) < -\alpha$, the estimate vanishes with high probability, indicating that no nearby samples are present. In the context of diffusion models, such regions are dominated by isolated training points. If probability mass is assigned to these locations, it must concentrate on individual samples, leading to memorization. This establishes the coexistence of memorized and generalized regions, as described in Eq. (3).

3.2. Measuring local sparsity and dominance

We now empirically investigate how local data density influences memorization behavior in diffusion models trained on image data, details on the training procedure are found in Section E. While our theory predicts a sharp transition as a function of local coverage, here we test its qualitative consequences using empirical proxies for density. We quantify local sparsity by the average squared distances $d_1, \dots d_k$ to the k nearest training samples, as well as local dominance by $\Delta(\hat{x}) = \log(d_2/d_1)$, which to which degree generated samples are dominated by close training data. Memorization is defined by the threshold criterion $d_1/d_2 < \tau$, see Section 3.2 for details.

Results. The middle column of Figure 1 shows the average dominance as a function of local sparsity for different training set sizes. We observe a clear **monotonic increase of dominance with sparsity** for generated samples: sparse regions exhibit strong single-sample dominance, while dense regions show low dominance and interpolation. In contrast, the test-data baseline remains nearly flat, indicating that this effect is not explained by nearest-neighbor geometry alone. As the number of training samples increases, the overall level of dominance decreases and the dependence on sparsity weakens, consistent with the expectation that higher sample density reduces isolated regions. The corresponding results for the memorization rate are shown in the right column of Figure 1. Morover, we see a clear **coexistence of memorization and generalization** that depends on data sparsity. Moreover, while the precise behavior depends on τ, the same qualitative trend is observed: **memorization increases with sparsity**. In the appendix Section D we show the outcome of the same experiment on downsampled CelebA, as well as extend the picture to the multi-class setting. The results of these experiments are also consistent with the hypothesis. These results support a local version of the KDE picture: diffusion models exhibit a continuous transition from interpolation in dense regions to single-sample dominance in sparse regions.

4. Discussion

Summary. In this work, we test the predictive power of kernel density estimation to explain local memorization in diffusion models. We show that memorization in diffusion models is not a global failure mode, but a *localized phenomenon* driven by the geometry of the data distribution. In particular, isolated regions of the data space can remain memorized even when the majority of generated samples are novel.

Limitations. Our analysis relies on the assumption that $\frac{1}{N} \ln p_{\text{in}}(x)$ satisfies a large deviation principle, analogous to concentration assumptions used in prior work (Lucibello & Mézard, 2024; Achilli et al., 2025a), which leads to an abrupt transition between memorization and generalization. We hypothesize that fluctuations are at the core of the more gradual increase observed empirically.Incorporating such effects is an important direction for future work. A second limitation stems from the choice of the metric space underlying the KDE approximation: we assumed that diffusion models operate in the ambient space. In practice, however, models may implicitly operate in a lower-dimensional representation, for example through projection onto a data manifold (Achilli et al., 2025a; 2026) or by exploiting symmetries in the data (Kamb & Ganguli, 2025). In such cases, the effective kernel should be understood as acting in this learned metric space, which depends on model architecture and may help explain the observed dependence of memorization on model capacity (Yoon et al., 2023), which has already been shown explicitly for autoencoder architectures by Zhang et al. (2025).

Outlook. Our results suggest that controlling memorization requires shaping the *local geometry* of the data representation. In particular, learning representations that increase local coverage, e.g. by mapping data to lower-dimensional spaces may reduce memorization but reduces model expressivity, presenting a tradeoff between increased local coverage and preserving generative performance. Another important direction concerns training dynamics. Memorization is known to depend on training time and optimization hyperparameters (Bonnaire et al., 2025; Favero et al., 2025; Wu et al., 2025; Garnier-Brun et al., 2026). Understanding how optimization dynamics affect local memorization is a promising direction for future research.

References

Achilli, B., Ambrogioni, L., Lucibello, C., Mézard, M., and Ventura, E. The Capacity of Modern Hopfield Networks under the Data Manifold Hypothesis, March 2025a. URL http://arxiv.org/abs/2503.09518. arXiv:2503.09518 [cond-mat].

Achilli, B., Ambrogioni, L., Lucibello, C., Mézard, M., and Ventura, E. Memorization and generalization in generative diffusion under the manifold hypothesis. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(7):073401, July 2025b. ISSN 1742-5468. doi: 10.1088/1742-5468/ade136. URL https://doi.org/10.1088/1742-5468/ade136.

Achilli, B., Ventura, E., Silvestri, G., Pham, B., Raya, G., Krotov, D., Lucibello, C., and Ambrogioni, L. Losing dimensions: Geometric memorization in generative diffusion, March 2026. URL http://arxiv.org/abs/2410.08727. arXiv:2410.08727 [stat].

Ambrogioni, L. In search of dispersed memories: Generative diffusion models are associative memory networks, November 2023. URL http://arxiv.org/abs/2309.17290. arXiv:2309.17290 [cs, stat].

Ambrogioni, L. In Search of Dispersed Memories: Generative Diffusion Models Are Associative Memory Networks. *Entropy*, 26(5):381, May 2024. ISSN 1099-4300. doi: 10.3390/e26050381. URL https://www.mdpi.com/1099-4300/26/5/381.

Biroli, G. and Mézard, M. Kernel Density Estimators in Large Dimensions, October 2024. URL http://arxiv.org/abs/2408.05807. arXiv:2408.05807 [cs].

Biroli, G., Bonnaire, T., de Bortoli, V., and Mézard, M. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, November 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-54281-3. URL https://www.nature.com/articles/s41467-024-54281-3.

Bonnaire, T., Urfin, R., Biroli, G., and Mézard, M. Why Diffusion Models Don't Memorize: The Role of Implicit Dynamical Regularization in Training, May 2025. URL http://arxiv.org/abs/2505.17638. arXiv:2505.17638 [cs].

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wallace, E. Extracting Training Data from Diffusion Models, January 2023. URL http://arxiv.org/abs/2301.13188. arXiv:2301.13188 [cs].

Deng, L. The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. *IEEE Signal Processing Magazine*, 29(6):141–142, November 2012. ISSN 1558-0792. doi: 10.1109/MSP.2012.2211477. URL https://ieeexplore.ieee.org/document/6296535.

Dotsenko, V. Replica solution of the random energy model. *EPL (Europhysics Letters)*, 95(5):50006, September 2011. ISSN 0295-5075, 1286-4854. doi: 10.1209/0295-5075/95/50006. URL https://iopscience.iop.org/article/10.1209/0295-5075/95/50006.

Fang, Z., Jiang, Z., Chen, H., Zhang, X., Tang, K., Li, X., and Li, J. A Closer Look on Memorization in Tabular Diffusion Model: A Data-Centric Perspective, August 2025. URL http://arxiv.org/abs/2505.22322. arXiv:2505.22322 [cs].

Favero, A., Sclocchi, A., and Wyart, M. Bigger Isn't Always Memorizing: Early Stopping Overparameterized Diffusion Models, September 2025. URL http://arxiv.org/abs/2505.16959. arXiv:2505.16959 [cs].

Garnier-Brun, J., Biggio, L., Beltrame, D., Mézard, M., and Saglietti, L. Biased Generalization in Diffusion Models, March 2026. URL http://arxiv.org/abs/2603.03469. arXiv:2603.03469 [cs].

George, A. J., Veiga, R., and Macris, N. Denoising Score Matching with Random Features: Insights on Diffusion Models from Precise Learning Curves, February 2025. URL http://arxiv.org/abs/2502.00336. arXiv:2502.00336 [cs].

Gu, X., Du, C., Pang, T., Li, C., Lin, M., and Wang, Y. On Memorization in Diffusion Models, February 2025. URL http://arxiv.org/abs/2310.02664. arXiv:2310.02664 [cs].

Jeon, D., Kim, D., and No, A. Understanding and Mitigating Memorization in Generative Models via Sharpness of Probability Landscapes, August 2025. URL http://arxiv.org/abs/2412.04140. arXiv:2412.04140 [cs].

Kadkhodaie, Z., Guth, F., Simoncelli, E., and Mallat, S. Generalization in diffusion models arises from geometry-adaptive harmonic representation. *ArXiv*, abs/2310.02557:null, 2023. doi: 10.48550/arXiv.2310.02557. URL https://www.semanticscholar.org/paper/a8724abaf519ab9113cf9dcc4c6d17f984de52cf.

Kadkhodaie, Z., Guth, F., Simoncelli, E. P., and Mallat, S. Generalization in diffusion models arises from geometry-adaptive harmonic representations, April

2024. URL http://arxiv.org/abs/2310.02557. arXiv:2310.02557 [cs].

Kamb, M. and Ganguli, S. An analytic theory of creativity in convolutional diffusion models. June 2025. URL https://openreview.net/forum?id=ilpL2qACla.

Kim, J., Kim, S., and Lee, J.-S. How Diffusion Models Memorize, September 2025. URL http://arxiv.org/abs/2509.25705. arXiv:2509.25705 [cs].

Krizhevsky, A. Learning Multiple Layers of Features from Tiny Images. 2009. URL https://www.semanticscholar.org/paper/Learning-Multiple-Layers-of-Features-from-Tiny-Krizhevsky/5d90f06bb70a0a3dced62413346235c02b1aa086.

Li, S., Chen, S., and Li, Q. A Good Score Does not Lead to A Good Generative Model, January 2024. URL http://arxiv.org/abs/2401.04856. arXiv:2401.04856 [cs].

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep Learning Face Attributes in the Wild. pp. 3730–3738. IEEE Computer Society, December 2015. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.425. URL https://www.computer.org/csdl/proceedings-article/iccv/2015/8391d730/12OmNzGlRCR.

Lucibello, C. and Mézard, M. Exponential Capacity of Dense Associative Memories. *Physical Review Letters*, 132(7):077301, February 2024. doi: 10.1103/PhysRevLett.132.077301. URL https://link.aps.org/doi/10.1103/PhysRevLett.132.077301.

Pham, B., Raya, G., Negri, M., Zaki, M. J., Ambrogioni, L., and Krotov, D. Memorization to Generalization: The Emergence of Diffusion Models from Associative Memory. November 2024. URL https://openreview.net/forum?id=zVMMaVy2BY.

Pidstrigach, J. Score-Based Generative Models Detect Manifolds. October 2022. URL https://openreview.net/forum?id=AiNrnIrDfD9.

Ronneberger, O., Fischer, P., and Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation, May 2015. URL http://arxiv.org/abs/1505.04597. arXiv:1505.04597 [cs].

Ross, B. L., Kamkari, H., Wu, T., Hosseinzadeh, R., Liu, Z., Stein, G., Cresswell, J. C., and Loaiza-Ganem, G. A Geometric Framework for Understanding Memorization in Generative Models, March 2025. URL http://arxiv.org/abs/2411.00113. arXiv:2411.00113 [stat].

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models, December 2022. URL http://arxiv.org/abs/2212.03860. arXiv:2212.03860 [cs].

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Understanding and Mitigating Copying in Diffusion Models. November 2023. URL https://openreview.net/forum?id=HtMXRGbUMt.

Wen, Y., Liu, Y., Chen, C., and Lyu, L. DETECTING, EXPLAINING, AND MITIGATING MEMO- RIZATION IN DIFFUSION MODELS. 2024.

Wu, Y.-H., Marion, P., Biau, G., and Boyer, C. Taking a Big Step: Large Learning Rates in Denoising Score Matching Prevent Memorization. In *Proceedings of Thirty Eighth Conference on Learning Theory*, pp. 5718–5756. PMLR, July 2025. URL https://proceedings.mlr.press/v291/wu25a.html.

Yoon, T., Choi, J. Y., Kwon, S., and Ryu, E. K. Diffusion Probabilistic Models Generalize when They Fail to Memorize. July 2023. URL https://openreview.net/forum?id=shciCbSk9h#all.

Zhang, Z., Li, X., Li, X., Shi, L., Wu, M., Tao, M., and Qu, Q. Generalization of Diffusion Models Arises with a Balanced Representation Space. October 2025. URL https://openreview.net/forum?id=57THeGgNAN.

A. Related Work

A growing body of empirical and theoretical work has established that diffusion models can memorize training data, particularly in low-data regimes. Memorization is not a uniform phenomenon: it depends strongly on properties of the training setup. In particular, it is exacerbated by data duplication (Carlini et al., 2023), prompt conditioning and classifier free guidance (Somepalli et al., 2023; Wen et al., 2024; Kim et al., 2025; Gu et al., 2025), model capacity (Yoon et al., 2023; George et al., 2025; Zhang et al., 2025) and training duration (Bonnaire et al., 2025; Favero et al., 2025), (Garnier-Brun et al., 2026) whereas it decreases with learning rate (Wu et al., 2025) and dataset size (Kadkhodaie et al., 2024; Somepalli et al., 2023). These findings suggest that memorization is a structured and predictable effect, rather than a rare failure mode.

Strategies to measure and mitigate memorization include determining the local intrinsic dimension of generated points (Ross et al., 2025), and estimating how "peaked" the estimated density is around a given point (Jeon et al., 2025) as well as identifying (Wen et al., 2024; Kim et al., 2025) memorized prompts. Importantly, these strategies share a common intuition: isolated data points produce a highly peaked local density, and produce samples of low intrinsic dimension. This perspective is naturally consistent with a kernel density estimation (KDE) view of diffusion models. Isolated training points induce highly concentrated local density estimates and generate low-dimensional samples, while dense regions support smoother interpolation. Similarly, conditioning can be interpreted as restricting the effective sample set contributing to the density estimate, thereby increasing local sparsity and promoting memorization. In this sense, existing empirical methods already implicitly rely on a local density perspective on memorization.

On the theoretical side, several works have established a connection between diffusion models and kernel density estimation (Pidstrigach, 2022; Ambrogioni, 2024; Li et al., 2024). Within this framework, one can determine a global "collapse phase" in diffusion models as a function of effective noise in the backward/sampling process (Biroli et al., 2024), where the model begins to reproduce training data. More generally, the sample complexity required for smooth interpolation is known to scale exponentially with the ambient or manifold dimension(Lucibello & Mézard, 2024; Achilli et al., 2025b) of the data. Recent work has further shown that memorization can emerge gradually through a loss of manifold dimensions Achilli et al. (2026) and that this process may be spatially inhomogeneous. However, existing theoretical analyses remain largely global, (Lucibello & Mézard, 2024; Biroli & Mézard, 2024; Achilli et al., 2025b) focused on establishing *global* phase transitions, focusing on memorization or generalization of *typical samples from the data distribution*. In contrast, we develop a local theory of memorization that explicitly links memorization to data coverage at the level of individual points. This provides a principled explanation for the heterogeneous memorization patterns observed in practice.

B. Theory overview: Kernel density estimation with hard spheres

In this appendix, we give an overview of the steps undertaken to derive our central result. Detailed calculations are given in the next section Section C. Recall that, to study kernel density estimation (KDE), we will now characterize the behavior of the log-density $z(x) = \frac{1}{N} \ln Z(x)$, where $Z(x)$ is the mixture of kernel functions K centered at the data points defined in Equation (6). The difficulty in establishing the behavior of z is that it is a random quantity that depends on the draw of the P training points from ρ. Our goal is to compute its distribution over such draws to derive general properties of high dimensional KDE.

To this end, we use a method introduced by Dotsenko (2011), which allows us to compute the cumulative distribution function (CDF) of $z(x)$ via an exponential transform. We define

$$\begin{aligned} W_N(y,x) =& \overline{\exp\left(-\exp\left(-N\left(y-z(x)\right)\right)\right)} \\ \overset{N\to\infty}{\longrightarrow}& \ \overline{\Theta(y-z(x))} \end{aligned} \tag{10}$$

where the $\overline{f(z(x))}$ denotes the average over draws of the P data points from ρ and Θ is the Heaviside function with convention $\Theta(0) = e^{-1}$. In the limit $N \to \infty$, this quantity converges to the cumulative distribution function of $z(x)$. If K is bounded from above (e.g. Gaussian K), we can expand the outer exponential function in powers of $\overline{Z^n(x)}$ and exchange the order of averaging and summation. This reduces the problem to computing moments of the form $\overline{Z^n(x)}$, which can be evaluated exactly using the Fourier transforms of ρ and K (see Section C). We then obtain the final expression

$$W_N(y,x) = \exp\left\{e^{\alpha N} \ln\left[1 - I_\circ\left(y,x\right)\right]\right\} \tag{11}$$

where

$$I_\circ\left(y,x\right) = \int d\tilde{x}\rho\left(\tilde{x}\right)\left(1 - e^{-K\left[\frac{x-\tilde{x}}{h}\right]\exp[-N(\alpha+y)]}\right) \leq 1\,. \tag{12}$$

This expression eliminates the explicit dependence on the random dataset, reducing the problem to an integral over the data distribution ρ. Although this integral remains high-dimensional, its structure depends only on the distribution of kernel values $K\left(\frac{x-\tilde{x}}{h}\right)$. In particular, if the logarithm of the kernel concentrates, this integral simplifies significantly. Even without such assumptions, the distribution of kernel values can be estimated empirically from data. This formulation makes explicit that the behavior of $z(x)$ is controlled by the statistics of distances between x and samples drawn from ρ, providing a direct link between KDE performance and local data geometry.

We now analyze the limit $N \to \infty$. Depending on the scaling of $\lim_{N\to\infty} e^{\alpha N} I_\circ(y,x) =: i_\circ(y,x)$, we obtain three regimes:

$$\lim_{N\to\infty} W_N(y,x) = \begin{cases} 0 & \text{if } i_\circ(y,x) = \infty \\ e^{-i_\circ(y,x)} & \text{if } i_\circ(y,x) = \mathcal{O}(1) \\ 1 & \text{if } i_\circ(y,x) = 0 \end{cases} \tag{13}$$

As a function of y, we find that $I_\circ$ decreases and thus W_N increases. A typical scenario is that the distribution of z concentrates with $N \to \infty$ thus $\Pr(\{y > z(x)\})$ becomes a step function, i.e. the measure of points y where the second condition in Equation (13) is fulfilled must shrink to zero. However, our results could in principle be used to compute fluctuations in $z(x)$, corresponding to a non-vanishing interval in y where the second condition is fulfilled.

Hard spheres. We now restrict ourselves to one particular kernel, namely

$$K(x) = \Theta(R^2 - x^2)/V_R \tag{14}$$

where V_R is the volume of the N-dimensional sphere with radius R. As we will see, this choice of kernel allows us to express Equation (13) in a particularly simple form.

In high dimensions, we expect that this choice of kernel is qualitatively equivalent to a Gaussian one with standard deviation R, as the mass of both distributions concentrates near a thin shell at radius R when $N \to \infty$. We fix $R = 2\pi e$, for which the volume V_R remains $\mathcal{O}(1)$ as $N \to \infty$. The effective scale of the kernel relative to the data is controlled by the bandwidth h, which can be interpreted as a rescaling of the data points by a factor $1/h$. Unless otherwise specified, we set $h = 1$.

Under these assumptions, the expression for $I_\circ$ simplifies significantly. Assuming that $\nu_{\text{in}}(x)$ remains $\mathcal{O}(1)$ as $N \to \infty$, we obtain the following sharp characterization of $z(x)$:

$$z(x) \to \begin{cases} \nu_{\text{in}}(x) & \text{if } \nu_{\text{in}}(x) \geq -\alpha \\ -\infty & \text{if } \nu_{\text{in}}(x) < -\alpha \end{cases}, \tag{15}$$

see Section C for a derivation.

This result shows that the behavior of $z(x)$ is controlled by the comparison between the local coverage $\nu_{\text{in}}(x)$ and the sample complexity α. Regions with sufficiently high local density behave as smooth KDE estimates, while regions with low density are dominated by the absence of nearby samples.

The divergence $z(x) \to -\infty$ for $\nu_{\text{in}}(x) < -\alpha$ reflects the fact that, in these regions, the kernel density estimate vanishes with high probability. Indeed,

$$\begin{aligned} \Pr(\{Z(x) = 0\}) = (1 - p_{\text{in}}(x))^P &= e^{P \ln(1 - e^{N\nu_{\text{in}}(x)})} \\ &\to \Theta(-[\alpha + \nu_{\text{in}}(x)]) \end{aligned}$$

On the other hand, when α is large enough, then $Z(x)$ approaches $p_{\text{in}}(x)$, hence the kernel density estimate converges to a local average of ρ over a sphere.

This establishes a sharp dichotomy between two regimes. In regions where $\nu_{\text{in}}(x) \geq -\alpha$, the kernel density estimate provides a smooth approximation of the underlying distribution, corresponding to a generalization regime. In contrast, in regions where $\nu_{\text{in}}(x) < -\alpha$, the estimate vanishes with high probability, indicating that no nearby samples are present.

In the context of diffusion models, such regions are dominated by isolated training points. If probability mass is assigned to these locations, it must concentrate on individual samples, leading to memorization. This establishes the coexistence of memorized and generalized regions, as described in Eq. (3).

C. Derivation

In this appendix, we give a detailed derivation of the results presented in Section B.

C.1. Average over draws of samples

Our starting point is the expansion

$$W_N(y,x) = \overline{\exp\left(-\exp\left(-N\left(y - z(x)\right)\right)\right)}$$

Recall that the notation $\overline{f(z)}$ denotes the average over the data set, meaning that when we insert the definition of $z(x)$ and explicitly state the average, we find

$$\begin{aligned} W_N(y,x) &= \int \prod_\mu d\rho(x^\mu) \exp\left(-e^{-N(y+\alpha)} \sum_\mu K(x - x^\mu)\right) \\ &= \left\{\int d\rho(\tilde{x}) \exp\left[-e^{-N(y+\alpha)} K\left(x - \tilde{x}\right)\right]\right\}^P \end{aligned}$$

which directly yields Equation (11).

C.2. Hard spherical kernels

We now outline how to derive the result for kernels that are hard spheres. Using the definition of p_{in}, we find that the integral $I_\circ$ defined in Equation (12) decomposes into two areas: those where $K = 0$ and those where $K = 1$. Using this distinction, we then find

$$I_\circ\left(y,x)\right) = p_{\text{in}}(x)\left(1 - \exp\left(-\exp\left\{-N\left(y+\alpha\right)\right\}\right)\right) \tag{16}$$

now additionally assuming that

$$\nu(x) = \frac{1}{N} \ln p_{\text{in}}(x) \tag{17}$$

scales as $\mathcal{O}_N(1)$, we find that

$$W_N(x,y) \to \begin{cases} \Theta\left(y - \nu_{\text{in}}(x)\right) & y + \alpha \geq 0 \\ \Theta\left(-\left(\nu_{\text{in}}(x) + \alpha\right)\right) & \text{else} \end{cases}$$

which yields the probability that $z \leq y$ in the limit. The first line shows that when z is larger than α, we recover $z \to \nu_{\text{in}}(x)$.

The second line must be treated with care. If $\nu(x) + \alpha > 0$, then this implies $y < \nu_{\text{in}}(x)$, meaning y is strictly smaller that z, i.e. y is just not large enough to have had the jump in the cumulative probability. On then other hand, when $\nu_{\text{in}}(x) + \alpha < 0$, then we have that $P_N\left(\{z_N(x) \leq y\}\right) = 1$ everywhere, meaning that z is smaller than any finite value of y. In this case, the jump happens at $z \to -\infty$, or zero local density.

Summarizing these findings we find that the cumulative distribution has the shape of a step function. Hence the value of $z(x)$ concentrates on the "jump location", given by

$$z(x) \to \begin{cases} \nu_{\text{in}}(x) & \text{if } \nu_{\text{in}}(x) \geq -\alpha \\ -\infty & \text{if } \nu_{\text{in}}(x) < -\alpha \end{cases} \tag{18}$$

which exactly the expression we report in Section B.

C.3. Convergence of the exponential transform to the CDF.

We justify that the averaged exponential transform probes the cumulative distribution function of $z_N(x)$. For notational simplicity we suppress the dependence on x and write

$$z_N := z_N(x), \qquad W_N(y) := \exp\{-\exp[-N(y - z_N)]\}.$$

Assume that $z_N \Rightarrow z$ in distribution and let $F(y) = \mathbb{P}(z \leq y)$ be the limiting cumulative distribution function. Let us suppose that F is continuous in y (i.e. there is no atomic mass at $z_N = y$). We will now show that

$$\lim_{N\to\infty} \overline{W_N(y)} = F(y).$$

The proof proceeds by constructing an upper and lower bound on W and showing that the bound becomes super-exponentially tight in N. For the lower bound, let us fix $\delta > 0$. If $z_N \leq y - \delta$, then

$$W_N(y) = \exp\{-\exp[-N(y - z_N)]\} \geq \exp\{-e^{-N\delta}\}.$$

In other words, using the indicator function $\mathbf{1}$, we know that

$$W_N(y) \geq e^{-e^{-N\delta}} \mathbf{1}\{z_N \leq y - \delta\}.$$

Taking expectations gives

$$\mathbb{E}[W_N(y)] \geq e^{-e^{-N\delta}} \mathbb{P}(z_N \leq y - \delta).$$

We now take care of the upper bound. Since $0 \leq W_N(y) \leq 1$, we know that

$$\begin{aligned} W_N(y) &\leq \mathbf{1}\{z_N < y + \delta\} + W_N(y)\mathbf{1}\{z_N \geq y + \delta\} \\ &\leq \mathbf{1}\{z_N < y + \delta\} + e^{-e^{N\delta}} \mathbf{1}\{z_N \geq y + \delta\}. \end{aligned}$$

Taking the average yields

$$\overline{W_N(y)} \leq \mathbb{P}(z_N < y + \delta) + e^{-e^{N\delta}} \mathbb{P}(z_N \geq y + \delta).$$

Combining the two bounds yields

$$e^{-e^{-N\delta}} \mathbb{P}(z_N \leq y - \delta) \leq \overline{W_N(y)} \leq \mathbb{P}(z_N < y + \delta) + e^{-e^{N\delta}} \mathbb{P}(z_N \geq y + \delta).. \tag{19}$$

Taking $N \to \infty$, using $z_N \to z$, and then sending $\delta \downarrow 0$, we find

$$\lim_{N\to\infty} \overline{W_N(y)} = F(y)$$

for every continuity point y of F. Thus the averaged transform $\overline{W_N(y)}$ converges to the CDF of the limiting random variable z.

C.4. Relation to Gumbel-type extreme value statistics

In this appendix we clarify the relation between the exponential transform used in Eq. (10) and Gumbel-type extreme value statistics. Let us firs clarify what we mean by Gumbel-type extreme value statistics: for independent random variables $u_1, \ldots, u_P$, one has

$$\mathbb{P}\left(\max_{\mu} u_\mu < y\right) = \left[1 - \mathbb{P}(u_\mu \geq y)\right]^P. \tag{20}$$

which is the standard structure associated with Gumbel-type extreme value statistics and, identifying $I_\circ$ with $\mathbb{P}(u_\mu \geq y)$, has strong similarity with Equation (11).

We will first show that letting $N \to \infty$ at $P < \infty$ fixed, W_N has a natural extreme-value interpretation. However, this does not mean that the calculation replaces the KDE by its largest term, because in the main result, we let $P \to \infty$ with $\frac{1}{N} \ln P = \alpha$ fixed. In this case, W_N is a soft probe of the cumulative distribution function of the full log-KDE.

For the sake of comparison, however, let us now keep $P < \infty$ and all samples fixed. We then find that W_N can be rewritten as

$$W_N(y, x) = \prod_{\mu=1}^{P} \exp\left\{-K(x - x^\mu)e^{-N(y+\alpha)}\right\}. \tag{21}$$

Now define the single-sample contribution $u_\mu(x) := \frac{1}{N} \ln K(x - x^\mu) - \alpha$. Then each factor in Eq. (21) can be written as

$$\exp\left\{-K_\mu(x)e^{-N(y+\alpha)}\right\} = \exp\left\{-\exp[N(u_\mu(x) - y)]\right\}. \tag{22}$$

If we were to fix P and all $u_\mu(x)$, and let $N \to \infty$ this converges to a hard threshold:

$$\exp\left\{-\exp[N(u_\mu(x) - y)]\right\} \xrightarrow[N\to\infty]{} \Theta(y - u_\mu(x)). \tag{23}$$

Consequently, if this hard-threshold limit is taken termwise before accounting for the exponentially many samples, Eq. (21) becomes

$$W_N(y, x) \approx \prod_{\mu=1}^{P} \Theta(y - u_\mu(x)) = \Theta\left(y - \max_\mu u_\mu(x)\right). \tag{24}$$

This is the origin of the formal similarity with Gumbel's law: the product form resembles the cumulative distribution function of a maximum. In regimes where the KDE is dominated by a single large kernel contribution, $z_N(x)$ is well approximated by

$$\widetilde{z}_N(x) := \max_\mu \frac{1}{N} \log \frac{K_\mu(x)}{P} = \max_\mu u_\mu(x). \tag{25}$$

In such max-dominated regimes, the above Gumbel-type interpretation directly describes the statistics of $z_N(x)$.

However, to obtain a generalization regime, we must send P to infinity too. Consequently, in the generalization regime, exponentially many samples may contribute to the KDE at any point. Then the sum cannot be replaced by its largest term. This distinction can be seen explicitly for the hard-sphere kernel. Let $m_N(x)$ be the number of training samples inside the corresponding ball around x with volume $V_R = 1$. Then

$$Z_N(x) = \frac{m_N(x)}{P}, \qquad \Rightarrow \qquad z_N(x) = \frac{1}{N} \log m_N(x) - \alpha. \tag{26}$$

If the local mass of the ball satisfies $\ln p_{\text{in}}(x) = N\nu_{\text{in}}(x) + o(N)$, then the typical number of samples in the ball is

$$m_N(x) \sim P p_{\text{in}}(x) = e^{N(\alpha+\nu_{\text{in}}(x))+o(N)} \tag{27}$$

whenever $\alpha + \nu_{\text{in}}(x) > 0$. In this regime,

$$z_N(x) \to \nu_{\text{in}}(x), \tag{28}$$

up to subexponential prefactors. By contrast, the largest single contribution is only

$$\widetilde{z}_N(x) = -\alpha \tag{29}$$

provided that at least one sample lies inside the ball. Thus, whenever $\nu_{\text{in}}(x) > -\alpha$, the full KDE exponent $z_N(x)$ differs from the maximum contribution $\widetilde{z}_N(x)$.

The Gumbel-like structure of Eq. (11) should therefore be understood as follows: the exponential transform is a soft thresholding device which, after factorization over samples, resembles an extreme-value observable. If one hardens this threshold before accounting for the exponentially many samples, one obtains a maximum-probe interpretation. Keeping the transform soft through the large-N calculation, however, retains the collective contribution of exponentially many moderate terms and therefore probes the full log-KDE $z_N(x)$.

D. Additional Experiments

D.1. Local sparsity and dominance

In Figure 2, we report the analogous experiment to Figure 1 for CelebA data, which we downsample to 32×32 greyscale pixels. We find that both local dominance and memorization increase with local sparsity. In comparison to CIFAR-10, we observe that the CelebA dataset appears to have fewer samples in very low density regions (compare logarithmic scale of Figure 2 d) to Figure 1d).

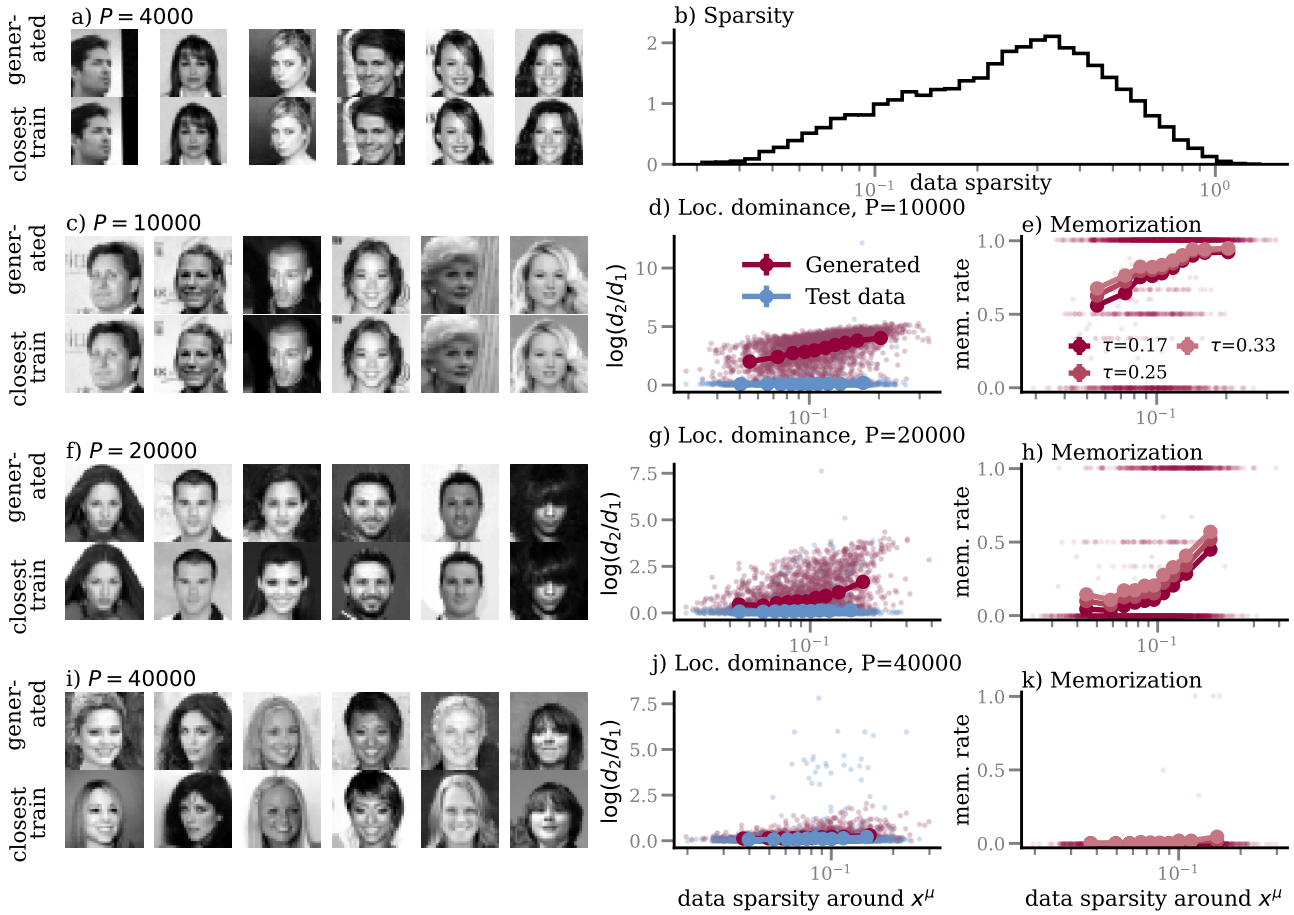

Figure 2. **Local memorization in CelebA data.** a) Sketch of kernel density approximation and local memorization phenomenon. b) Distribution of data sparsity for 40,000 samples from CelebA. c) Generated samples + closest training image (measured by cosine similarity) for diffusion model trained on $P = 4000$ training examples. d) Local dominance as a function of sparsity. For each generated sample $\hat{x}$, we compute distances d_1 and d_2 to its nearest and second-nearest training samples, and define dominance via their ratio. Each point corresponds to a training sample x^μ, aggregating generated samples assigned to it. Scatter shows raw values, markers indicate binned averages. Blue points show the same quantity computed using test data, providing a baseline without memorization. e) Memorization rate per training sample x^μ, defined as the fraction of generated samples satisfying $d_1/d_2 < \tau$. Scatter shows raw values for $\tau = 0.1$, markers indicate binned averages across thresholds. f)–k) Same measurements for models trained on larger datasets. As the number of training samples increases, both dominance and memorization decrease, and their dependence on sparsity weakens. Overall, sparse regions exhibit strong single-sample dominance and higher memorization, while dense regions promote interpolation across multiple training points.

D.2. Class-dependent Memorization

We now study memorization at a coarser level, focusing on **class-dependent effects**. In a multi-class setting, the KDE picture predicts that classes with more concentrated support (i.e. lower local sparsity) should be memorized less than classes with higher intra-class variability. A schematic illustration is shown in Figure 3a. To formalize this intuition, consider a mixture of class-conditional densities ρ_c, such that

$$\rho(x) = \textstyle\sum_c w_c \rho_c(x),$$

where the weights w_c sum to one, and we assume that all weights are order one in N. We define the local in-distribution log-density for class c as

$$\nu_{\mathrm{in},c}(x) := \frac{1}{N} \ln p_{\mathrm{in},c}(x) = \frac{1}{N} \ln \int_{B_h(x)} d\rho_c(x'), \tag{30}$$

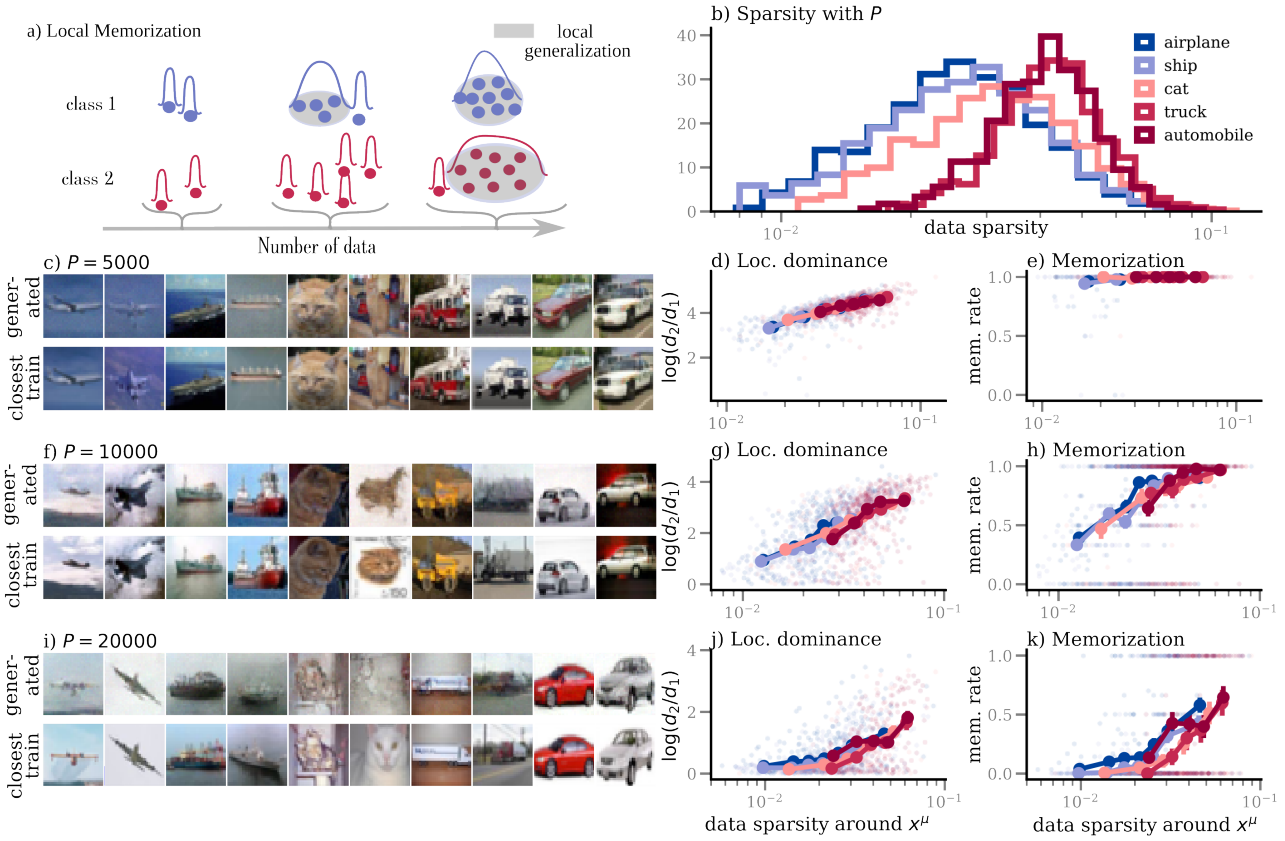

Figure 3. **Class-wise memorization.** a) Schematic illustration of KDE for two classes with different local sparsities. b) Distribution of per-class sparsity around training points for different subclasses of CIFAR-10, measured from 20,000 training samples. c) Generated samples together with their closest training example (measured by cosine similarity) for a diffusion model trained on $P = 1000$ samples, sorted to contain 2 nearest neighbors for each of the four classes shown in b). d) Local dominance as a function of sparsity. For each generated sample $\hat{x}$, we compute distances d_1 and d_2 to its nearest and second-nearest training samples, and define dominance via their ratio. Each point corresponds to a training sample x^μ, aggregating generated samples assigned to it. Scatter shows raw values, markers indicate binned averages. Different colors correspond to different classes. e) Memorization rate of each sample x^μ, defined as the fraction of generated samples satisfying $d_1/d_2 < \tau = 1/3$. f) - k) report the same measures, but for diffusion models trained on larger datasets. "Airplane" and "ship" samples lie in denser regions and exhibit lower dominance and memorization, while "truck" and "automobile" samples are sparser and more frequently memorized, consistent with the prediction that local coverage controls class-dependent memorization.

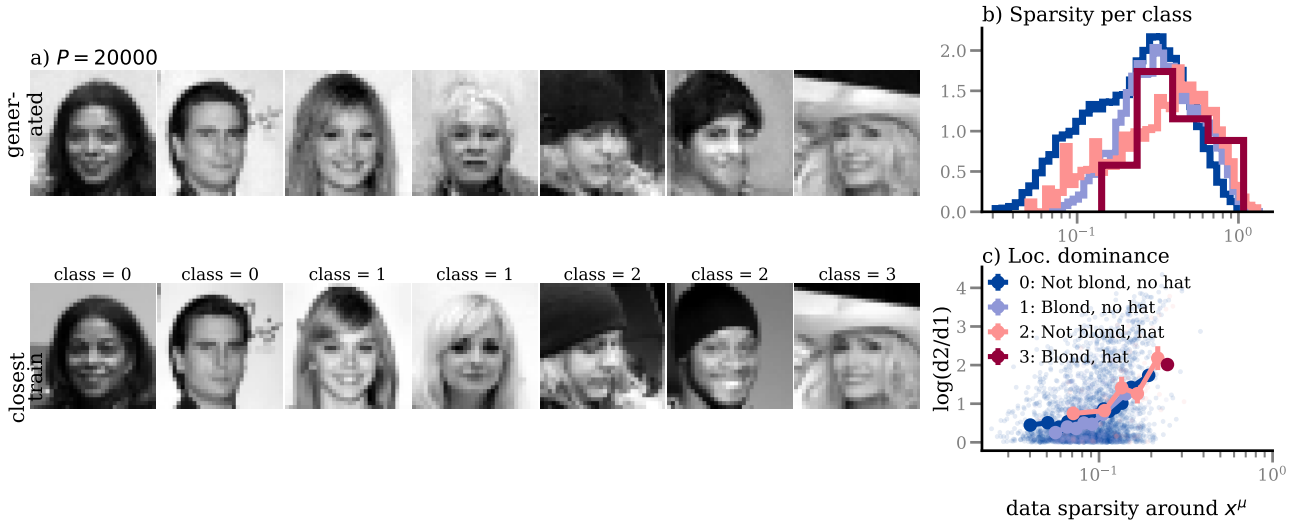

Figure 4. **Attribute-conditioned memorization in CelebA.** a) Generated samples + closest training image (measured by cosine similarity) for diffusion model trained on $P = 20000$ training examples, sorted by classes. b) Distribution of data sparsity for 40000 samples from CelebA. c) Local dominance of training sample x^μ against local data sparsity around x^μ for a diffusion model trained on $P = 20000$ training examples from CelebA.

and assume that $\nu_{\text{in},c}(x)$ remains $\mathcal{O}(1)$ in the large-N limit. Then the total in-distribution density satisfies

$$\begin{aligned}\nu_{\text{in}}(x) &= \frac{1}{N} \ln \sum_c e^{N\left(\nu_{\text{in},c}(x) + N^{-1} \ln w_c\right)} \\ &\to \max_c \nu_{\text{in},c}(x),\end{aligned}$$

i.e. it is dominated by the locally densest class. Consequently, we obtain the same scaling behavior as in Equation (15):

$$z(x) \to \begin{cases} \max_c \nu_{\text{in},c}(x) & \text{if } \max_c \nu_{\text{in},c}(x) \geq -\alpha, \\ -\infty & \text{otherwise,} \end{cases} \tag{31}$$

showing that memorization is governed by the class with the highest local coverage around x. This leads to a clear prediction: classes with higher intra-class variability (and thus lower local coverage) should be more prone to memorization.

We now demonstrate empirically that such class-dependent density differences naturally arise. As a concrete example, we isolate the memorization behavior of different sub-classes of CIFAR-10. In Figure 3b we show that the classes "airplane" and "ship" in CIFAR-10 have, on average, considerably lower intra-class diversity than the classes "truck" and "automobile", likely due to more homogeneous blue background colors in the former two classes. Correspondingly, we observe both higher local dominance, and higher memorization in more diverse classes.

The same trends emerge for two additional experiments. First, when one sorts the CelebA datasets into classes according to whether images have the attribute "wearing hat" or "blond", images where "wearing hat" is true are more diverse than those where it is false. Correspondingly, portraits featuring hats are memorized more, see Figure 4. Second, we train diffusion models on a mixture of MNIST (Deng, 2012) and CIFAR-10 images. MNIST, consisting of handwritten digits, is structurally simpler and occupies a much more concentrated region of image space, whereas CIFAR-10 exhibits substantially higher variability. Treating MNIST and CIFAR-10 as two distinct classes, we observe that MNIST samples consistently exhibit lower local sparsity than CIFAR-10 samples. As predicted by the theory, this translates into reduced memorization, see Figure 5. These results support the hypothesis that classes with higher intra-class diversity are memorized more strongly than classes with lower intra-class diversity.

D.2.1. Attribute-dependent Memorization, CelebA

The CelebA dataset consists of celebrity portraits that are annotated with "attributes" such as hair color or accessories. We construct four classes from these attributes, conditioning on "blond" and "not blond" as well as "wearing hat" and the

opposite. Again, we find that local dominance (and thus memorization) increases with data sparsity, see Figure 4. Samples from classes where "wearing hat" is true are typically more diverse. Therefore this class has a higher average sparsity per class,and is therefore more likely to be memorized. These classes are not balanced: approximately one sixth of images has attribute "blond" and approximately 5 % of images have attribute "wearing hat". Consequently the diffusion model generates fewer samples that are memorized data points with these attributes, and even fewer samples are closest to training data where both attributes are true. This is reflected in the lower bin resolution in Figure 4b) and fewer such points appearing in Figure 4 c).

D.2.2. CLASS-DEPENDENT MEMORIZATION, CIFAR-10+MNIST

In a separate experiment, we train diffusion models on a mixture of MNIST (Deng, 2012) and CIFAR-10 images. MNIST, consisting of handwritten digits, is structurally simpler and occupies a much more concentrated region of image space, whereas CIFAR-10 exhibits substantially higher variability. Treating MNIST and CIFAR-10 as two distinct classes, we observe that MNIST samples consistently exhibit lower local sparsity than CIFAR-10 samples. As predicted by the theory, this translates into reduced memorization: in Figure 5, both local dominance and memorization rates increase with sparsity. For the comparison of CIFAR and MNIST, all distances are computed on ℓ_2-normalized samples, so that sparsity reflects relative geometric structure (angular similarity) rather than differences in overall scale. We verified that using unnormalized samples yields qualitatively similar results. On average, MNIST points lie in denser regions and are therefore memorized less than CIFAR-10 points.

E. Experimental Details

E.1. Measuring local sparsity, dominance, and memorization

We give a detailed definition of the experimental measures reported.

Local sparsity. For each training sample x^μ, we quantify local sparsity using nearest-neighbor distances, which serve as a proxy for inverse local coverage. Concretely, we define

$$s(x^\mu) = \frac{1}{Nk} \sum_{i=1}^{k} \|x^\mu - x^{\mu_i}\|^2, \tag{32}$$

where $\{x^{\mu_i}\}$ are the k nearest neighbors of x^μ. Larger values of $s(x^\mu)$ correspond to sparser regions of the data distribution. We also validate that the results remain consistent for different choices of k in $\{5, 10, 20, 50\}$. The results reported in the figures correspond to $k = 10$.

Local dominance. Given a generated sample $\hat{x}$, we compute its distances $d_1(\hat{x})$ and $d_2(\hat{x})$ to the nearest and second-nearest training samples, and define the *dominance*

$$\Delta(\hat{x}) = \log \frac{d_2(\hat{x})}{d_1(\hat{x})}. \tag{33}$$

Large values of Δ indicate that a single training point dominates the local score estimation around the generated sample, while $\Delta \approx 0$ corresponds to interpolation between multiple neighbors. We assign each generated sample to its nearest training point and compute, for each x^μ, the average dominance over all generated samples assigned to it. To disentangle model-specific effects from dataset geometry, we construct a baseline by replacing generated samples with held-out test data, processed in the same way.

Memorization rate. We further report a binary memorization metric used in several previous studies (Wu et al., 2025; Yoon et al., 2023; Bonnaire et al., 2025) based on the condition $d_1/d_2 < \tau$.

E.2. Training Diffusion models

Model. We train denoising diffusion models using a standard discrete-time formulation with $T = 1000$ timesteps. The score network is parameterized by a U-Net architecture (Ronneberger et al., 2015) with convolutional filters. All models operate directly in pixel space.

Training procedure. Models are trained using the standard diffusion objective with mean squared error loss. Optimization is performed using Adam with learning rate 10^{-4} and batch size 100. Training proceeds for a fixed number of 10^6 gradient steps for all models.

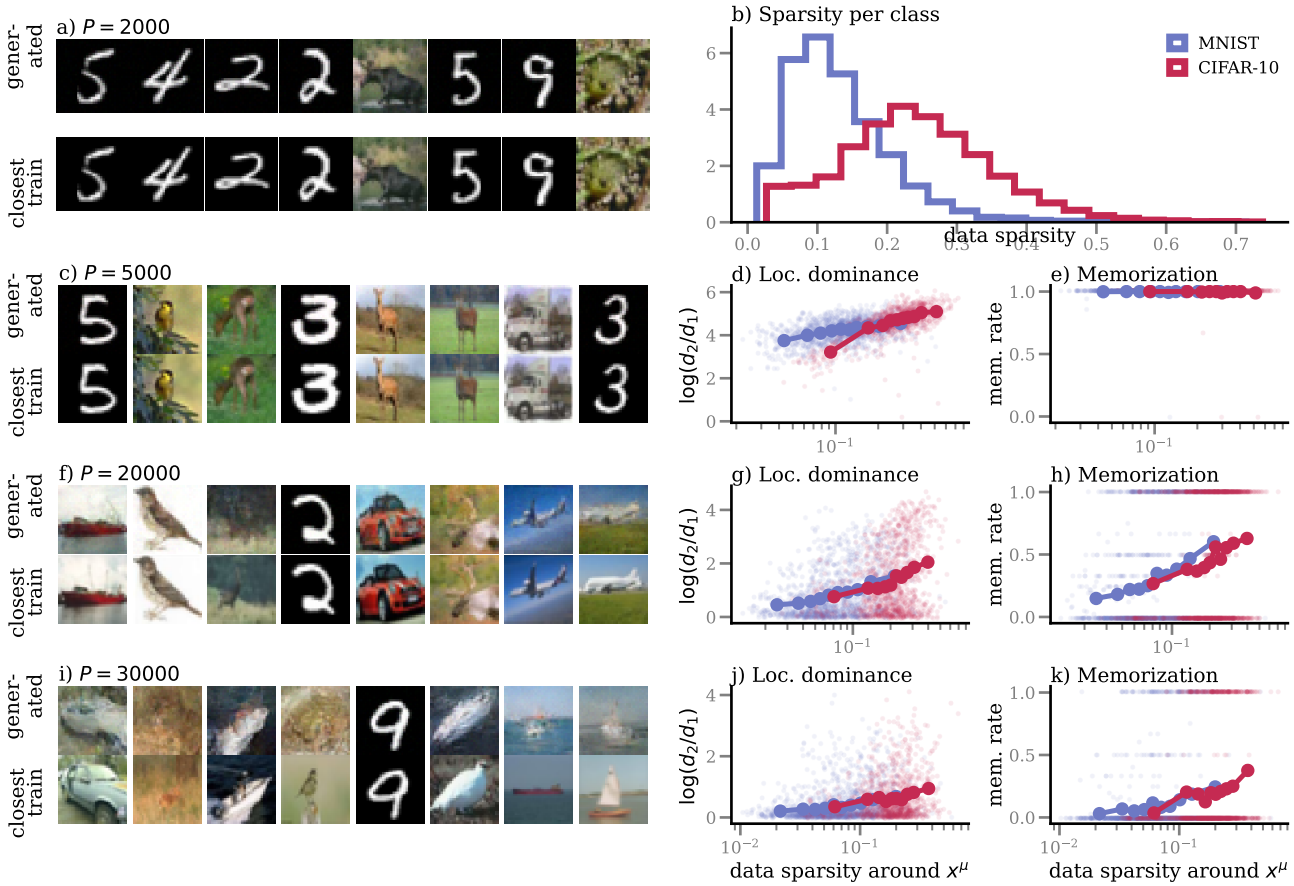

Figure 5. **Class-wise memorization on combined MNIST and CIFAR-10.** a) Generated samples together with their closest training example (measured by cosine similarity) for a diffusion model trained on $P = 2000$ samples. b) Distribution of per-class sparsity around training points for both classes, measured from 20,000 training samples. c) Generated samples together with their closest training example (measured by cosine similarity) for a diffusion model trained on $P = 5000$ samples. d) Class-wise local dominance of a training point x^μ as a function of local data sparsity. Local dominance is defined via the ratio between the distance d_1 to the closest generated sample and the distance d_2 between that generated sample and the second-closest training point. MNIST points are shown in blue, CIFAR-10 points in red. e) Memorization rate of each sample x^μ, defined as the fraction of generated samples satisfying $d_1/d_2 < \tau = 1/3$. Different classes are shown as different colors. f)–k) Same measurements for models trained on larger datasets. As the number of training samples increases, both dominance and memorization decrease, and their dependence on sparsity weakens.

Evaluation For all experiments, the dataset is deterministically split into a training set of size P and a held-out test set. The same split is reused for all evaluations to ensure comparability across checkpoints and metrics. At evaluation time, we generate $N_{\text{gen}} = 4000$ samples from the trained diffusion model. We compare generated samples to training and test data using either cosine similarity or normalized ℓ_2 distance. In the cosine case, all inputs are flattened and ℓ_2-normalized before computing similarities. Memorization is quantified at the level of individual generated samples. For each generated sample $\tilde{x}$, we compute its two nearest training neighbors and define distances d_1 and d_2. A sample is considered memorized if $\frac{d_1}{d_2} < \tau$ for a range of thresholds $\tau \in \{1/6, 1/4, 1/3, 1/2, 2/3\}$. This yields both per-sample and per-training-point memorization statistics. Local data density is estimated using nearest-neighbor statistics computed solely on the training set. For each training sample, we compute the distances to the k nearest neighbors for $k \in \{2, 5, 10, 20, 50\}$ and average over these distances as defined in Equation (32) These statistics serve as proxies for local sparsity and are later correlated with memorization behavior.

E.3. CIFAR-10 Experiments

We construct the CIFAR-10 dataset by combining the original training and test splits (60,000 images total), and then randomly partitioning them into three equal subsets (train/validation/test), each containing approximately 20,000 images. All images are represented as 32×32 RGB tensors normalized to $[0, 1]$. No additional preprocessing or augmentation is applied.

E.4. MNIST + CIFAR-10 Experiments

To study memorization under controlled differences in local data density, we construct a combined dataset from CIFAR-10 and MNIST, which exhibit markedly different intrinsic complexities. CIFAR-10 images are highly variable, while MNIST digits occupy a much more concentrated region of image space. This setup allows us to induce systematic differences in local sparsity across classes.

We merge the original training and test splits of both datasets. To ensure compatibility, MNIST images are resized to 32×32 and converted to three channels by duplicating the grayscale channel, resulting in RGB tensors consistent with CIFAR-10. CIFAR-10 images are used without modification. All images are represented as 32×32 RGB tensors normalized to $[0, 1]$, and no additional preprocessing or augmentation is applied.

To isolate dataset-level effects, we ignore the original class labels and instead assign a binary label indicating the dataset of origin (CIFAR-10 or MNIST). The two datasets are then balanced by subsampling the larger dataset such that both contribute an equal number of samples. The resulting dataset therefore consists of an equal mixture of CIFAR-10 and MNIST images.

E.5. CelebA Experiments

This dataset consists of high-resolution face images together with 40 binary attributes per image. All images are resized to 32×32 pixels, and converted to grayscale. The resulting images are represented as single-channel tensors normalized to $[0, 1]$. No additional data augmentation is applied.