Analyzing Memorization in Vision: Diffusion Model

Camille Chang
Dept. of Electrical and Computer Engineering, University of California, Davis
tyuchang@ucdavis.edu

Yating Chang
Dept. of Electrical and Computer Engineering, University of California, Davis
cyating@ucdavis.edu

Keywords— Diffusion Probabilistic Models, Generalization in Generative Models, IC Layout Pattern Generation

I. Introduction

Diffusion models are really making waves in the world of generating images, showing off some top-notch results in creating high-quality visuals across different fields. One standout type is the Denoising Diffusion Probabilistic Models, or DDPMs, which have caught a lot of attention. They're not just theoretically solid; they also work really well in handling complex data. Basically, they learn to reverse a noise process, either in a hidden space or right in the pixel space, which helps them create super realistic images, whether they have structured input to guide them.

But it's not all smooth sailing. People are getting worried about how these diffusion models sometimes memorize training data. This is a big deal, especially in areas like semiconductor design, where keeping data private and making sure generated data is diverse really matters. If a model just memorizes the training examples, it can mess with the usefulness and trustworthiness of the generation process. And unlike with other types of models where overfitting is a clear issue, this kind of memorization can be pretty sneaky and requires special tools to spot and measure.

In this study, we're diving into how memorization works compared to generalization in these vision diffusion models, mainly focusing on generating synthetic patterns. We'll check out how similar the generated samples are to the training data, using some handy metrics like Structural Similarity Index (SSIM), entropy, and mutual information to figure out how risky memorization can be. Plus, we're throwing some rare patterns into the training data to see if the model can remember those oddballs or if it can generalize better.

What we found is interesting. Diffusion models are great at generating stuff, but they can also end up repeating specific patterns from the training data, especially if those patterns are simple or pretty rare. This has serious implications for using these models in real-world situations where it's super important to maintain a balance of diversity, privacy, and strong generalization.

II. Datasets and Environment

This study uses grayscale images that show specific patterns found in integrated circuit (IC) design and photolithography. Each image has different features like contacts, metal and vias, which are key for creating semiconductor structures. These patterns are usually sparse and symmetrical, often repeating certain shapes, which makes them tricky for models to learn without overfitting. It's a great way to test how well generative models can memorize things.

Before diving into the study, we made sure all the images were consistent. We resized them all to either 64×64 or 128×128 pixels and adjusted the brightness levels so that the pixel values were between 0 and 1. Even though the original patterns are just black-and-white, this adjustment kept some of the gray variations, allowing models to produce more nuanced outputs during training. We divided the dataset into two parts: one for training the diffusion model and the other for testing the generated samples' similarity and novelty.

A big challenge here is that a lot of these contact layer images have low diversity, so it's tough for generative modeling. The simple structure and repetition can lead to memorization, especially with models that don't have enough regularization. But that makes it perfect for checking how well models balance generalization and memorization. Plus, since these layout patterns are like sensitive design data used in real semiconductor manufacturing, this dataset helps us think about privacy and intellectual property issues when using generative models.

We conduct model training using MetalSet and Contact dataset, which consists of printed and target layout images extracted from real-world semiconductor manufacturing processes. The images are grayscaled and contain a variety of via, line, and corner patterns. All images are resized to 128×128 times pixels and normalized to the [0,1] range. We adopt a latent diffusion pipeline consisting of a pre-trained AutoencoderKL for image encoding and decoding, and a UNet2DModel trained with DDPM noise schedules.

Training is performed on an NVIDIA GPU using a batch size of 32 for 5-10 epochs. To improve generalization, we incorporate both L1 reconstruction loss and entropy regularization.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

III. METRICS

The **Structural Similarity Index (SSIM)** is a perceptual metric used to evaluate the visual similarity between two images, typically a generated image and a reference from the training set. Unlike pixel-wise metrics such as MSE, SSIM incorporates human visual perception by considering luminance, contrast, and structural information. It is especially useful in detecting subtle forms of memorization in generative models, where an output image may look nearly identical to a training sample. The SSIM between two images x and y is computed over small windows (usually 11×11), and is defined as:

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

where

- μ_x, μ_y : means of x and y
- σ_x^2 , σ_y^2 : variances
- σ_{xy} : covariance
- C_1, C_2 : small constants to stabilize the division

The SSIM score is between 0 and 1. Where 0 indicates no similarity and 1 indicates high similarity. We label samples as:

SSIM Range	Interpretation	Risk Category
> 0.98	Extremely similar → likely memorized	High Memorization Risk
0.90 – 0.98	Strong structural similarity	Medium Risk
0.75 – 0.90	Some overlap, but distinguishable	Low Risk
< 0.75	Visually novel / uncommon	Very Low Risk

Table 1 SSIM Range for Memorization Risk

In parallel, we compute the **entropy** of each generated sample to measure pixel-level diversity.

Entropy is a fundamental measure of uncertainty or information content in a distribution, widely used in image analysis to quantify complexity. In the context of generative models, entropy can be used to evaluate the diversity or repetitiveness of generated images. A high-entropy image indicates greater structural variation, while low entropy often implies simplicity or potential memorization of training patterns. Given a grayscale image, we first compute its normalized histogram $\{pi\}_{i=1}^{N}$, where pi is the probability of pixel intensity level i, and N is the number of bins (e.g., 256 for 8-bit images). The Shannon entropy is then defined as:

$$H(X) = -\sum_{i=1}^{N} p_i \log_2(p_i)$$

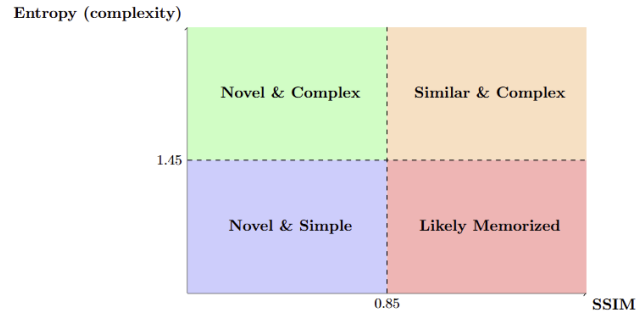

Fig.1 Entropy vs SSIM Memorization Risk Quadrant

Also, to evaluate the similarity between generated and training images, we adopt Mutual Information (MI).

Mutual Information (MI) is a statistical measure of dependency between two random variables or images, used to assess how much one image reveals about another. In the context of memorization detection, MI complements SSIM by detecting shared structural content even when perceptual similarity is low. For two grayscale images X and Y, MI is computed from their joint and marginal distributions:

$$\mathrm{MI}(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

Here, $p(x,y)$ is the joint probability of pixel intensities x and y, while $p(x)$ and $p(y)$ are their marginals. MI is zero when the two images are completely independent and increases as shared structure or redundancy grows. In practice, MI is estimated using normalized 2D histograms over pixel values. High MI with low SSIM may indicate subtle or non-obvious memorization behavior in a generative model. Unlike SSIM, MI can detect hidden structural correlations even when visual similarity is low, making it useful for identifying subtler forms of memorization. Together, SSIM and MI provide a robust framework for assessing both perceptual and statistical similarity, helping to reveal whether a generative model is producing novel outputs or replicating training data.

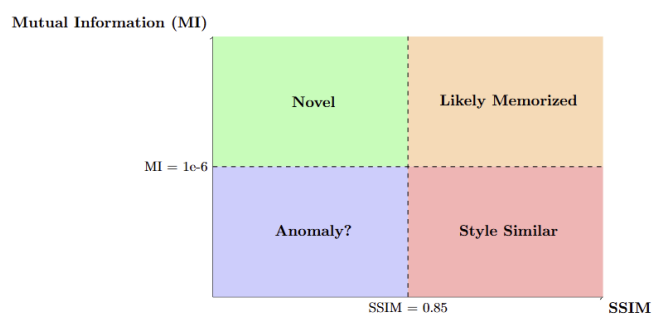

Fig.1 MI vs SSIM Memorization Risk Quadrant

IV. RESULTS

To assess memorization behavior in the diffusion model, we evaluated each generated image using three complementary metrics: Structural Similarity Index (SSIM), Mutual Information (MI), and Shannon Entropy. Each

generated sample was compared to its most similar training image to obtain SSIM and MI scores, while entropy was computed independently to capture intrinsic complexity.

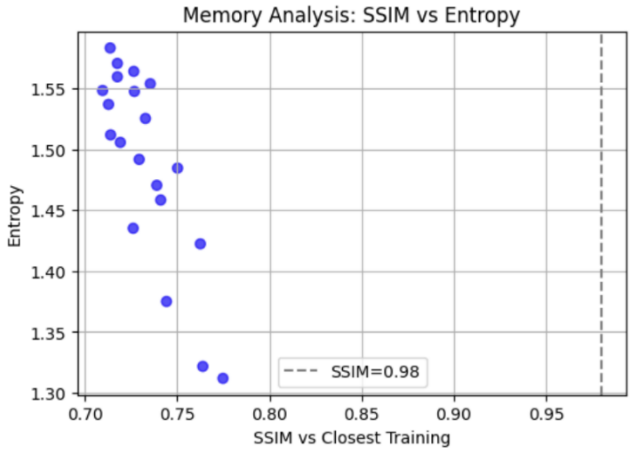

Fig.3 Memory Analysis of SSIM vs Entropy

Figure 3 shows a strong inverse relationship between SSIM and entropy: samples with lower SSIM values generally display higher entropy, suggesting that perceptually novel images tend to be more complex and less structured. Conversely, low-entropy samples — which are simpler and potentially repetitive — tend to have higher SSIM scores, indicating greater visual overlap with training data.

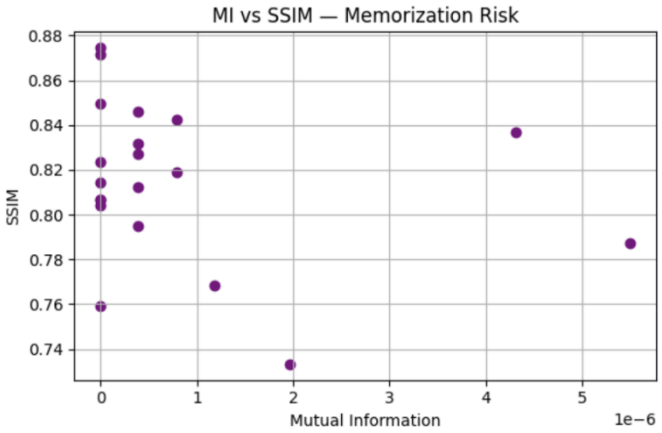

Fig.4 Memory Analysis of SSIM vs. MI

Figure 4 shows that most generated samples have relatively low MI values, clustered below $1x10^{-6}$, with corresponding SSIM scores ranging between 0.73 and 0.88. This suggests that, although some samples are perceptually similar to training images (moderate SSIM), they do not share substantial pixel-wise structure. Notably, only a few outliers show both higher MI and SSIM, indicating hidden memorization in the diffusion model.

Together, these results indicate that the model mostly generates novel outputs but occasionally reproduces patterns that are visually or structurally close to training data. The combined use of SSIM, entropy, and MI enables a understanding of memorization risk, where outliers in 2D plot can be flagged for further review. This framework offers a scalable and interpretable approach for further analysis of generative model behavior.

V. Conclusion

In this study, we proposed a metric-based framework to evaluate memorization risk in diffusion-based generative models. By analyzing generated samples using Structural Similarity Index (SSIM), Shannon Entropy, and Mutual Information (MI), we were able to characterize the extent to which the model reproduced patterns from its training data. Our results demonstrate that while the diffusion model is capable of producing diverse and novel outputs, certain samples exhibit strong visual or statistical similarity to training examples, suggesting potential memorization.

The combination of SSIM with entropy and MI provided complementary insights: SSIM captured perceptual similarity, entropy quantified internal image complexity, and MI revealed structural dependencies. This multi-metric analysis enabled quadrant-based interpretation of sample behavior and allowed us to flag outliers with high memorization potential.

Overall, the framework offers a practical and scalable method for post-hoc memorization evaluation, supporting safer and more responsible use of generative models, particularly in applications where data privacy or content originality is critical.

References

[1] G. R. Reddy, K. Madkour and Y. Makris, "Machine Learning-Based Hotspot Detection: Fallacies, Pitfalls and Marching Orders," 2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), Westminster, CO, USA, 2019, pp. 1-8, doi: 10.1109/ICCAD45719.2019.8942128

[2] Sumedha and R. Rohilla, "Exploring Vision Transformer model for detecting Lithography Hotspots," 2022 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON), Bengaluru, India, 2022, pp. 200-205, doi: 10.1109/CENTCON56610.2022.10051370

[3] Zheng, Su, et al. "Lithobench: Benchmarking ai computational lithography for semiconductor manufacturing." *Advances in Neural Information Processing Systems* 36 (2023): 30243-30254.

[4] Engin Afacan, Nuno Lourenço, Ricardo Martins, Günhan Dündar,Review: Machine learning techniques in analog/RF integrated circuit design, synthesis, layout, and test,Integration, Volume 77,2021, Pages 113-130, ISSN 0167-9260

[5] Z. Chai, Y. Zhao, W. Liu, Y. Lin, R. Wang and R. Huang, "CircuitNet: An Open-Source Dataset for Machine Learning in VLSI CAD Applications With Improved Domain-Specific Evaluation Metric and Learning Strategies," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 42, no. 12, pp. 5034-5047, Dec. 2023, doi: 10.1109/TCAD.2023.3287970.