# Memorization to Generalization: The Emergence of Diffusion Models from Associative Memory

**Bao Pham**[*]
Department of CS
RPI
phamb@rpi.edu

**Gabriel Raya**[*]
Jheronimus Academy of Data Science
Tilburg University
g.raya@jads.nl

**Matteo Negri**
Department of Physics
University of Rome Sapienza
matteo.negri@uniroma1.it

**Mohammed J. Zaki**
Department of CS
RPI
zaki@cs.rpi.edu

**Luca Ambrogioni**[†]
Donders Institute
Radboud University
l.ambrogioni@donders.ru.nl

**Dmitry Krotov**[†]
MIT-IBM Watson AI Lab
IBM Research
krotov@ibm.com

## Abstract

Hopfield networks are associative memory systems, designed for storing and retrieving specific patterns as local minima of an energy landscape. In the classical Hopfield model, an interesting phenomenon occurs when the model's memorization capacity reaches its critical memory load —*spurious states*, or unintended stable points, emerge at the end of the retrieval dynamics. These particular states often appear as mixtures of the stored patterns, leading to incorrect recall. In this work, we propose that these spurious states are not necessarily a negative feature of retrieval dynamics, but rather that they serve as the onset of generalization. We employ diffusion models, commonly used in generative modeling, to demonstrate that their generalization stems from a phase transition which occurs as the number of training samples is increased. In the low data regime, the model exhibits a strong memorization phase, where the network creates a distinct basin of attraction for each sample in the training set, akin to the Hopfield model below the critical memory load. In the large data regime, a different phase appears where an increase in the training set size fosters the creation of new attractor states that correspond to manifolds of the generated samples. Spurious states appear at the boundary of this transition, and correspond to emergent attractor states, which are absent in the training set, but, at the same time, still have a distinct basin of attraction around them. From the perspective of Hopfield description these spurious states correspond to mixtures of "*fundamental memories*" which facilitate generalization through the superposition of underlying features, resulting in the creation of novel samples. Our findings provide a novel perspective on the memorization-generalization phenomenon in diffusion models via the lens of Hopfield networks, which illuminates the previously underappreciated view of diffusion models as Hopfield networks above the critical memory load.

## 1 Introduction

Originally introduced in [Hop82, Hop84], Hopfield networks have seen a resurgence in interest due to advances in their memorization capacity. Notably, Dense Associative Memories, which are

[*]B. Pham and G. Raya both equally contributed to this work as first authors.
[†]L. Ambrogioni and D. Krotov both equally contributed in advising this work.

Workshop on Scientific Methods for Understanding Deep Learning at NeurIPS 2024.

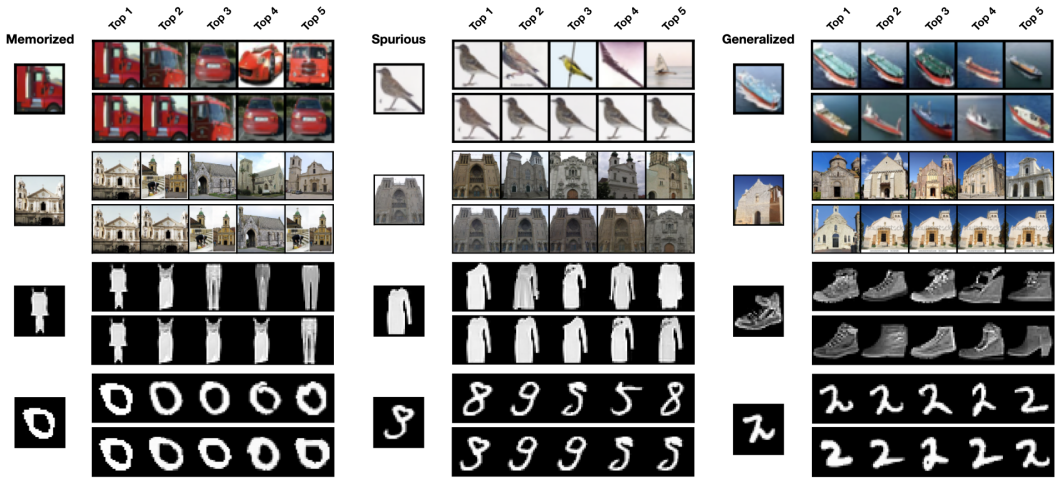


**Figure 1:** Illustration contrasting memorized, spurious, and generalized patterns in their respective column for various datasets. For each target image, its *top-5* nearest neighbors from the training set (***top row***) and the synthetic set (***bottom row***) are shown to highlight the novelty or the commonality of the image with respect to training and synthetic sets. Memorized samples are duplicates of the training set while spurious samples are copies of the synthetic set. In contrast, generalized samples do not belong to any of the two sets, indicating that the model is creative (or has fully generalized).

extensions of Hopfield networks with super-linear memory capacity [KH16, KH18] have paved the way for more sophisticated associative memory systems [DHL$^+$17, AAB$^+$20, ADM20, Kro21, AAAB22, MSS$^+$22, SKZR23, HLP$^+$23, KOT24, dSNMM24, AAAB24], driven in part by their strong connection to the attention mechanism in transformers [RSL$^+$21, KH21]. Simultaneously, generative diffusion models [SDWMG15] have gained considerable popularity, due to their flexibility and accuracy in modeling high dimensional distributions for a variety of domains —ranging from image generation [HJA20, SME20, SSDK$^+$21], audio [CZZ$^+$20, KPH$^+$20, LCY$^+$23], video synthesis [HSG$^+$22, SPH$^+$22, BRL$^+$23, BPH$^+$24], and other scientific applications. However, despite their effectiveness, diffusion models pose challenges related to privacy and security, as concerns grow over their tendency to memorize or replicate training data [SSG$^+$23a, SSG$^+$23b]. Such matters consequently emphasize the need for further understanding of memorization and generalization behaviours in diffusion models.

Recent works [HSK$^+$23, Amb24, RA24] have begun establishing theoretical connections between Dense Associative Memories and generative diffusion models, offering a foundation for bridging the two fields. It has been shown that the logarithm of the probability of the generated samples in diffusion models can be interpreted as the energy function in a commonly used Dense Associative Memory model with the softmax activation function [Amb24]. This makes it possible to apply theoretical tools developed for associative memory to better understand the computational properties of diffusion models. See Appendix C for further details on such connections.

A common feature of Hopfield networks is the phenomenon of spurious patterns. Historically considered as detrimental to pattern recall [HHFP83, AGS85, AMSJ85], spurious patterns can be interpreted as combinations or interpolations of stored patterns, hinting at the network's ability to synthesize new patterns from existing ones. *This blending of fundamental memories resembles the generalization process in generative models, where learned representations are used to generate novel outputs*. In this way, spurious patterns offer a fascinating framework for exploring the balance between memorization —where models store exact patterns from training data —and generalization, where they use underlying structures to create genuinely new samples [KLP$^+$24]. Studying the conditions under which spurious patterns emerge and how they contribute to the network's computation can thus shed light on generalization in both associative memory and contemporary generative models.

Previous works have explored memorization in generative models through various approaches. For instance, [MCD20, dBW21] propose general methods to measure memorization, while [SSG$^+$23a, SSG$^+$23b, YCKR23] study memorization capacity in diffusion models as a function of training data size. Other works focus on understanding generalization, such as [LLZB24], which provides theoretical estimates of the generalization gap, and [KGSM23], which offers spectral analyses of how diffusion models generalize. While these studies shed light on memorization and generalization, they do not fully explore the model's behavior during the transition between these two phases.

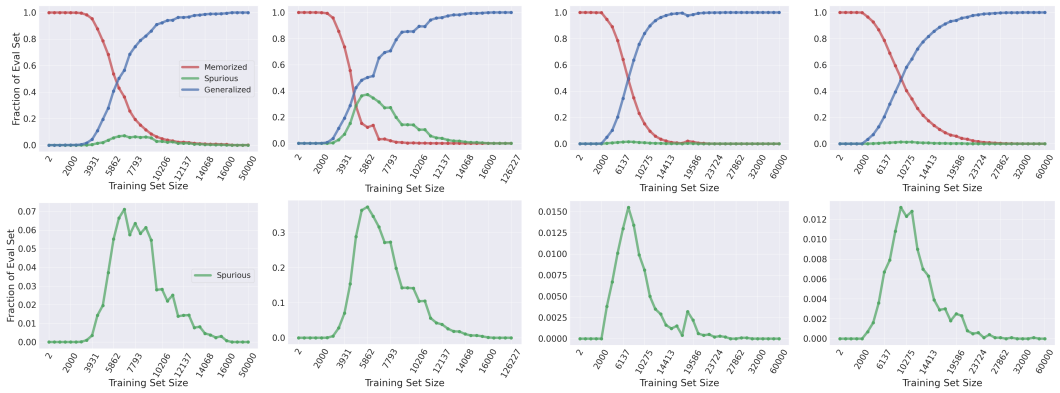


**Figure 2:** Illustrations showing the fractions of memorized, spurious, and generalized samples across different datasets. The fraction of each sample type is with respect to each data split and its corresponding evaluation set of 10k samples. As memorization decreases due to increase in data size, the onset of generalization is initiated with the emergence of spurious patterns, indicating the beginning of attractors' collapse. The fraction of spurious patterns rises and decreases quickly at the boundary between the memorization and generalization phases.

**Contributions** In this work, we center our focus on establishing diffusion models as Hopfield networks that have exceeded their memorization capacity, losing the ability to reliably recall stored data points. To make this connection, we demonstrate that **(1)** diffusion models undergo a phase transition from memorization to generalization as the number of training points increases; **(2)** this onset of generalization is induced through the emergence of spurious patterns which signifies the collapse or reduction of memorized attractors; and lastly, unlike previous works in Hopfield models, **(3)** we provide theoretical descriptions distinguishing spurious states from generalized patterns. By formally establishing the distinction between spurious and generalized patterns, we provide deeper insights on the mechanisms which enable generalization in diffusion (and other generative) models.

## 2 The Onset of Generalization

To characterize the memorization-to-generalization transition, we focus on the model's capacity to generate new samples while still retaining some degree of its ability to replicate memorized data. In simple terms, this transition can be seen as the development from a student — who learns by copying — to an expert, capable of creating original work. During this intermediate phase, the model begins to move beyond replication, but lacks the full creative capacity to generate truly novel patterns with high probability. Following [YCKR23], we extend their memorization capacity $\mathcal{MC}$ to define both spurious and generalization capacities to delineate this transition.

Let $G$ be a diffusion model that maps noise $\mathbf{z}$ drawn from a prior distribution $q$ to a data point in a data space $\mathcal{X}$. Additionally, let $\mathcal{T}$ be a training algorithm that takes in a training set $S = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K\} \subset \mathcal{X}$ and produces a set of parameters $\theta$. Furthermore, let $S' = \{\mathbf{x}'_1, \mathbf{x}'_2, \ldots, \mathbf{x}'_m\}$ be a set of synthetic samples generated by the trained diffusion model $G_\theta$, where the condition $|S'| \gg |S|$ is required to account for duplication in $S'$.

**Definition 2.1** (**Memorization Capacity**)**.** The maximum number of training samples $n$ such that the model $G_\theta$ consistently reproduces those samples with high probability. Specifically, this number represents the largest training set size where the probability that the model generates samples belonging to the training set $S$ remains the highest, within a tolerable error margin. This number can thus be described as

$$\mathcal{MC}_S = \max\left\{ K \;\middle|\; \operatorname*{Prob}_{\mathbf{z}\sim q}\left[G_\theta(\mathbf{z}) \in S\right] \geq 1 - \epsilon \right\} \tag{1}$$

where a small value $\epsilon \in (0, 1)$ represents the permissible error margin for the probability and $G_\theta(\mathbf{z})$ is a target sample generated from $G_\theta$, independently from the generation of the synthetic set $S'$, using the noise $\mathbf{z}$ drawn from $q$.

**Hypothesis 2.1.** When the training set size exceeds the memorization capacity, $|S| > \mathcal{MC}$, the model begins to generate spurious patterns, signaling the onset of generalization. These spurious attractors are interpolations between the stored patterns, signaling a transition from strong memorization to initial pattern synthesis.

**Definition 2.2** (**Spurious Capacity**). The number of training samples $n$ which maximizes the probability of samples generated from $G_\theta$ to belong only in the synthetic set $S'$ and not the training set $S$. This number denotes the inflection point at which the model is unable to simply replicate the training data but not yet generalized to the true underlying data distribution. This number can be described as

$$\mathcal{SC}_{S,S'} = \arg\max_K \left\{ \Pr_{\mathbf{z}\sim q} \left[ G_\theta(\mathbf{z}) \notin S \wedge G_\theta(\mathbf{z}) \in S' \right] \right\} \tag{2}$$

**Hypothesis 2.2.** When the training set size exceeds both the memorization capacity, $|S| \gg \mathcal{MC}$, and spurious capacity, $|S| \gg \mathcal{SC}$, the model is near full generalization, closely modeling the underlying data distribution. At this stage, many of the generated samples are genuinely novel, exhibiting no duplication or direct replication of the training data $S$ and the synthetic set $S'$. The fraction of such samples is greater than the fractions of memorized and spurious samples, respectively.

**Definition 2.3** (**Generalization Capacity**). The minimum number of training samples $K$ which maximizes the novelty of samples generated from $G_\theta$ under the condition that they do not belong to either the synthetic set $S'$ or the training set $S$, given as

$$\mathcal{GC}_{S,S'}(G, \mathcal{T}) = \min \left\{ K \;\middle|\; \Pr_{\mathbf{z}\sim q} \left[ G_\theta(\mathbf{z}) \notin S \wedge G_\theta(\mathbf{z}) \notin S' \right] \geq 1 - \epsilon \right\} \tag{3}$$

**Hypothesis 2.3.** When the training set size exceeds the generalization capacity, $|S| > \mathcal{GC}$, the model has reached full generalization, effectively modeling the underlying data distribution. At this stage, generated samples are genuinely novel, exhibiting no duplication or direct replication of the training data $S$ and the synthetic set $S'$.

### 2.1 Detection Metrics

Below, we introduce specific detection metrics to quantify and distinguish between memorized, spurious, and generalized patterns — to tackle the above hypotheses. Using these metrics, we present results which map out the phase transition of different diffusion models trained on various sizes of different datasets (see Figure 2). The metrics defined below rely on three datasets: $S$ — the training set used to train the model $G_\theta$, $S'$ — the synthetic dataset generated from $G_\theta$, and $S^{\text{eval}}$ — the evaluation dataset, which is also generated from $G_\theta$, but independently from $S'$; the size of $S'$ is assumed to be much bigger than the training set size.

**Metric 2.1** (**Memorization Detection**). Following [YCKR23] and using the definition of memorization capacity (1), we define the memorization detection metric $\mathcal{M}$. Given a target pattern $\hat{\mathbf{x}} \in S^{\text{eval}}$, its first and second nearest neighbors $\mathbf{x}_1$ and $\mathbf{x}_2$ are extracted from the training set $S$ according to a distance measure $d(\cdot,\cdot)$ — small distance corresponds to high similarity. Metric $\mathcal{M}$ detects sample $\hat{\mathbf{x}}$ as memorized if it belongs to the training set $S$. Specifically,

$$\mathcal{M}(\hat{\mathbf{x}}, S) = \mathbb{I}\left( \frac{d(\hat{\mathbf{x}}, \mathbf{x}_1)}{d(\hat{\mathbf{x}}, \mathbf{x}_2) + \epsilon} \leq \delta_m \right) \tag{4}$$

where $\mathbb{I}$ represents the indicator function, $\delta_m \in (0, 1)$ is a small threshold value, and $\epsilon$ is a small constant to avoid division by zero, e.g., $\epsilon = 10^{-6}$. In words, $\hat{\mathbf{x}}$ is much closer, and thus considered memorized, to $\mathbf{x}_1$ than any other point in training set $S$.

**Metric 2.2** (**Spurious Detection**). Using the definition of spurious capacity (2), we define the spurious pattern detection metric $\mathcal{S}$. It identifies instances where the model generates outputs that do not belong to the training set but have high similarity with samples from the synthetic set,

$$\mathcal{S}(\hat{\mathbf{x}}, S, S') = \mathbb{I}\left( \frac{d(\hat{\mathbf{x}}, \mathbf{x}'_1)}{d(\hat{\mathbf{x}}, \mathbf{x}_1) + d(\hat{\mathbf{x}}, \mathbf{x}'_1) + \epsilon} \leq \delta_s \right) \wedge \neg \mathcal{M}(\hat{\mathbf{x}}, S) \tag{5}$$

using a small threshold value $\delta_s \in (0, \frac{1}{2})$ as part of the criterion. With respect to the target pattern $\hat{\mathbf{x}}$, $\mathbf{x}_1$ and $\mathbf{x}'_1$ are the first nearest neighbors extracted from the training set $S$ and synthetic set $S'$, respectively. This metric ensures that the ratio is bounded between 0 and 1 while requiring that $d(\hat{\mathbf{x}}, \mathbf{x}'_1) \ll d(\hat{\mathbf{x}}, \mathbf{x}_1)$ for a pattern to be considered as spurious. If the two distances are equal, then the ratio is close to 0.5. Once the spurious patterns are identified via $\mathbb{I}$ given the threshold $\delta_s$, they are also filtered from the memorized sample set to ensure that the two sets are disjointed.

**Metric 2.3** (**Generalization Detection**). Lastly, following the definition of generalization capacity (3), we define the generalized pattern detection metric $\mathcal{G}$ as

$$\mathcal{G}(\hat{\mathbf{x}}, S, S') = \neg \mathcal{M}(\hat{\mathbf{x}}, S) \wedge \neg \mathcal{S}(\hat{\mathbf{x}}, S, S') \tag{6}$$

which identifies a given target sample $\hat{\mathbf{x}}$ as generalized if it is not classified either as memorized nor spurious using the above detection metrics $\mathcal{M}$ and $\mathcal{S}$.

## 3 Experiments

Using our detection metrics, we computed the fractions of memorized, spurious, and generalized samples for various sizes of training data for different datasets (see Figure 2). These datasets include MNIST [Den12], FashionMNIST [XRV17], CIFAR10 [KNH14], and LSUN-Church [YZS$^+$15] scaled down to $64 \times 64$ resolution using center-crop and down-scale. For each dataset, we trained a DDPM-based diffusion model [HJA20] for $M = 38$ different training set sizes obtained by using a fixed random seed to split. The same training setting as DDPM was used with the exception of using no random flip, modifying batch size, and the channel multipliers in the Unet backbone to accommodate for the dataset's dimensionality. For each model $\alpha = 1, ..., M$, trained on the training set $S_\alpha$, an evaluation set $S_\alpha^{\text{eval}}$ of 10k samples was generated. Additionally, a synthetic set $S'_\alpha$ was generated for each model. To account for duplication and the diversity of the dataset, we ensured that each synthetic set is 16 times the size of the corresponding training set $|S'_\alpha| = 16|S_\alpha|$ for CIFAR10 and LSUN-Church. Meanwhile, for FashionMNIST and MNIST, we kept the synthetic size as four times the size of the corresponding training set $|S'_\alpha| = 4|S_\alpha|$ due to their limited diversity. Each sample $\hat{\mathbf{x}} \in S_\alpha^{\text{eval}}$ was classified as memorized, spurious, or generalized using the above metrics (Eqs. 4, 5, 6). The ratio of the sizes of these three sets with respect to $|S_\alpha^{\text{eval}}|$ =10k was used to plot the curves in Figure 2. For each dataset our smallest model was trained on the training set of $|S_1| = 2$ data points, and the largest model was trained on the entire original training set $S_M = S$. Please refer to Appendix B for more details on training, selection of the $M$ data sizes, and detection metrics' hyperparameters for each dataset.

## 4 Discussion

The results in Figure 2 clearly demonstrate the transition from memorization to generalization as the dataset size increases, validating the above hypotheses (2.1, 2.2, 2.3). Meanwhile, the collected samples show distinct characteristics in each of the considered phases (see Figure 1). For example, in the small data regime, the diffusion model predominantly replicates the training data (see Figure 1 memorized panel, or additional samples in Figure 3 of the Appendix). As the data size surpasses the memorization capacity $\mathcal{MC}$, we observe a critical transition where the memorization fraction declines and spurious patterns begin to emerge. Such patterns are primarily composites of the data points and exhibit strong duplication in the synthetic set (see Figure 1 spurious panel, or additional samples in Figure 4 of the Appendix). The emergence of such patterns aligns with the onset of generalization, as the model moves away from strictly reproducing the training set and starts generating novel combinations of learned features. Once the training size surpasses the spurious capacity $\mathcal{SC}$, the creation of memorized and spurious samples starts to decrease, signaling the transition to full generalization regime. When $\mathcal{GC}$ is surpassed, the model completely loses its replication ability and no duplicate samples are detected in either training or synthetic sets (see Figure 1 generalized panel, or additional samples in Figure 5 of the Appendix).

## 5 Conclusion

In this work, we establish a novel connection between diffusion models and Hopfield models, demonstrating how diffusion models experience a phase transition from memorization to generalization as the training data size increases. Moreover, unlike the general knowledge in associative memory, we show that spurious patterns are not merely negative artifacts but are, in fact, indicators of the onset of generalization. These patterns emerge as the model begins to learn and leverage common features across the data. By highlighting this transition, we offer a deeper understanding of how diffusion models evolve toward generalization, with spurious patterns serving as a bridge between the memorization and generalization phases. Future work could further investigate the size and dynamics of the basins of attraction in each phase to provide a more comprehensive view of how diffusion models generalize.

## References


[AAAB22] Linda Albanese, Francesco Alemanno, Andrea Alessandrelli, and Adriano Barra. Replica symmetry breaking in dense hebbian neural networks. *Journal of Statistical Physics*, 189(2):24, 2022.

[AAAB24] Linda Albanese, Andrea Alessandrelli, Alessia Annibale, and Adriano Barra. About the de almeida–thouless line in neural networks. *Physica A: Statistical Mechanics and its Applications*, 633:129372, 2024.

[AAB+20] Elena Agliari, Francesco Alemanno, Adriano Barra, Martino Centonze, and Alberto Fachechi. Neural networks with a redundant representation: Detecting the undetectable. *Physical review letters*, 124(2):028301, 2020.

[ADM20] Elena Agliari and Giordano De Marzo. Tolerance versus synaptic noise in dense associative memories. *The European Physical Journal Plus*, 135(11):1–22, 2020.

[AGS85] Daniel J. Amit, Hanoch Gutfreund, and H. Sompolinsky. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys. Rev. Lett.*, 55:1530–1533, Sep 1985.

[Amb24] Luca Ambrogioni. In search of dispersed memories: Generative diffusion models are associative memory networks. *Entropy*, 26(5):381, 2024.

[AMSJ85] Y. Abu-Mostafa and J. St. Jacques. Information capacity of the hopfield model. *IEEE Transactions on Information Theory*, 31(4):461–464, 1985.

[BPH+24] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024.

[BRL+23] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[CHN+23] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23, USA, 2023. USENIX Association.

[CZZ+20] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*, 2020.

[dBW21] Gerrit J.J. Van den Burg and Chris Williams. On memorization in probabilistic deep generative models. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[Den12] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[DHL+17] Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168(2):288–299, May 2017.

[dSNMM24] Saul José Rodrigues dos Santos, Vlad Niculae, Daniel C McNamee, and Andre Martins. Sparse and structured hopfield networks. In *Forty-first International Conference on Machine Learning*, 2024.

[Efr83] Bradley Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American statistical association*, 78(382):316–331, 1983.

[ET94] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

[HHFP83] John J. Hopfield, John J. Hopfield, David I. Feinstein, and Richard G. Palmer. 'unlearning' has a stabilizing effect in collective memories. *Nature*, 304:158–159, 1983.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020.

[HLP+23] Benjamin Hoover, Yuchen Liang, Bao Pham, Rameswar Panda, Hendrik Strobelt, Duen Horng Chau, Mohammed Zaki, and Dmitry Krotov. Energy transformer. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 27532–27559. Curran Associates, Inc., 2023.

[Hop82] J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

[Hop84] John J Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the national academy of sciences*, 81(10):3088–3092, 1984.

[HSG+22] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

[HSK+23] Benjamin Hoover, Hendrik Strobelt, Dmitry Krotov, Judy Hoffman, Zsolt Kira, and Duen Horng Chau. Memory in plain sight: A survey of the uncanny resemblances between diffusion models and associative memories. *arXiv preprint arXiv:2309.16750*, 2023.

[Hyv05] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[KGSM23] Zahra Kadkhodaie, Florentin Guth, Eero P Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representation. *arXiv preprint arXiv:2310.02557*, 2023.

[KH16] Dmitry Krotov and John J. Hopfield. Dense associative memory for pattern recognition. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[KH18] Dmitry Krotov and John Hopfield. Dense associative memory is robust to adversarial inputs. *Neural computation*, 30(12):3151–3167, 2018.

[KH21] Dmitry Krotov and John J Hopfield. Large associative memory problem in neurobiology and machine learning. In *International Conference on Learning Representations*, 2021.

[KLP+24] Silvio Kalaj, Clarissa Lauditi, Gabriele Perugini, Carlo Lucibello, Enrico M Malatesta, and Matteo Negri. Random features hopfield networks generalize retrieval to previously unseen examples. *arXiv preprint arXiv:2407.05658*, 2024.

[KNH14] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. The cifar-10 dataset. *online: http://www. cs. toronto. edu/kriz/cifar. html*, 55(5):2, 2014.

[KOT24] Ryo Karakida, Toshihiro Ota, and Masato Taki. Hierarchical associative memory, parallelized mlp-mixer, and symmetry breaking. *arXiv preprint arXiv:2406.12220*, 2024.

[KPH+20] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

[Kro21] Dmitry Krotov. Hierarchical associative memory. *arXiv preprint 2107.06446*, 2021.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

[LCY+23] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 21450–21474, 2023.

[LLZB24] Puheng Li, Zhong Li, Huishuai Zhang, and Jiang Bian. On the generalization properties of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

[MCD20] Casey Meehan, Kamalika Chaudhuri, and Sanjoy Dasgupta. A non-parametric test to detect data-copying in generative models. In *International Conference on Artificial Intelligence and Statistics*, 2020.

[MSS+22] Beren Millidge, Tommaso Salvatori, Yuhang Song, Thomas Lukasiewicz, and Rafal Bogacz. Universal hopfield networks: A general framework for single-shot associative memory models. In *International Conference on Machine Learning*, pages 15561–15583. PMLR, 2022.

[RA24] Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

[RSL+21] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David Kreil, Michael K Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021.

[SDWMG15] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015.

[SE19] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[SKCK17] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017.

[SKZR23] Bishwajit Saha, Dmitry Krotov, Mohammed J Zaki, and Parikshit Ram. End-to-end differentiable clustering with associative memories. In *International Conference on Machine Learning*, pages 29649–29670. PMLR, 2023.

[SME20] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[SPH+22] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[SSDK+21] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[SSG+23a] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6048–6058, 2023.

[SSG+23b] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803, 2023.

[VdOKE+16] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, and Alex Graves. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016.

[Vin11] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

[XRV17] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint 1708.07747*, 2017.

[YCKR23] TaeHo Yoon, Joo Young Choi, Sehyun Kwon, and Ernest K Ryu. Diffusion probabilistic models generalize when they fail to memorize. In *ICML 2023 Workshop on Structured Probabilistic Inference Generative Modeling*, 2023.

[YZS+15] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[YZS+23] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Comput. Surv.*, 56(4), nov 2023.

[ZIE+18] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

# A Appendix

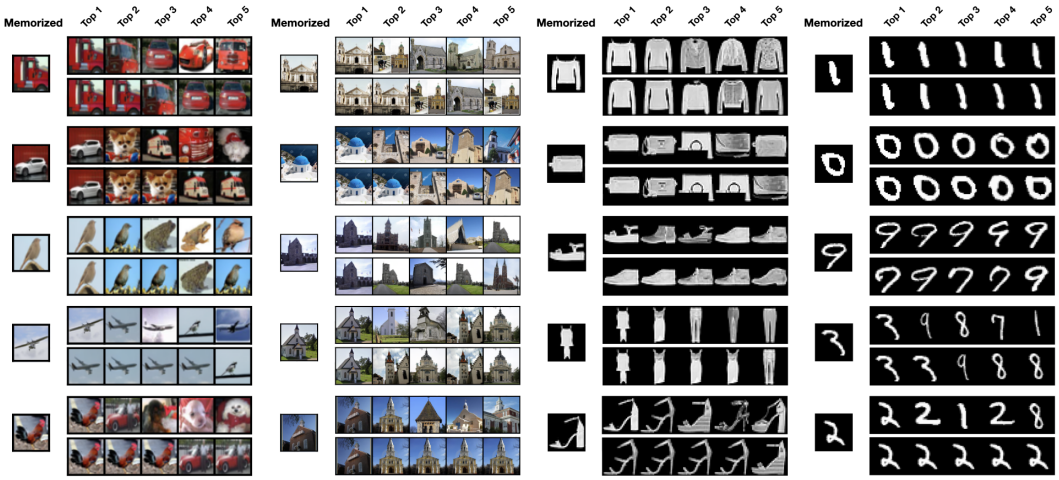


**Figure 3:** Visualization of memorized patterns and their nearest neighbors for different datasets. The ***top row*** illustrates nearest neighbors from the training set while the ***bottom row*** depicts those from the synthetic set. Memorized samples are duplicates of the training set. During the strong memorization phase, duplicates are also found within the synthetic set. Note, even though our memorization metric does not utilize the synthetic set, we are showing the nearest neighbors obtained from it, for consistency.

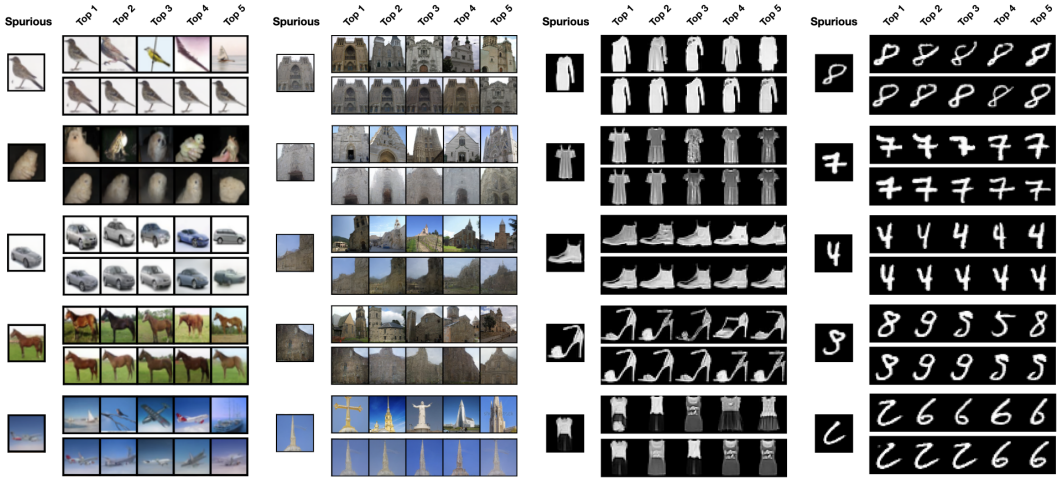


**Figure 4:** Visualization of spurious patterns and their *top-5* nearest neighbors for different datasets. The ***top row*** illustrates nearest neighbors from the training set while the ***bottom row*** shows those from the synthetic set. Spurious patterns are demonstrated to arise from the onset of generalization where the mixing of training points begins. Since the basin of attractions is yet to collapse completely, spurious patterns can be duplicated with high probability, as seen in the synthetic set, much like memorized patterns. Hence, spurious samples lack the uniqueness to be considered as generalized samples.

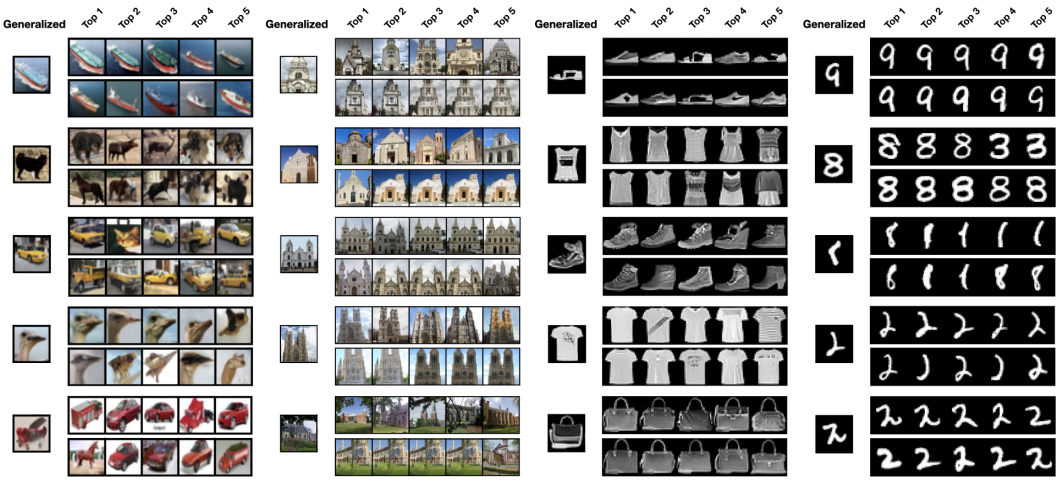


**Figure 5:** Visualization of generalized patterns and their nearest neighbors for different datasets. The ***top row*** illustrates nearest neighbors from the training set while the ***bottom row*** depicts those from the synthetic set. Generalized samples are novel samples, which have little to no resemblance to their nearest neighbors in training and synthetic sets, that fully trained models can generate.

# B Additional Details on Transition Mapping

## B.1 Points Selection

For the computation of Figure 2, we follow the experimentation of [YCKR23] and conduct a sparse search starting at $0.5$k data size and doubling it all the way to the total data size $|S|$, i.e., $K = 0.5\text{k}, 1\text{k}, 2\text{k}, \ldots, N$. We then identified two transitional critical points, $A$ and $B$ (see Table 1), to further highlight the memorization-generalization transition. Point $A$ indicates the initial drop in memorization while $B$ signals the plateauing of memorization. To capture the details of the memorization-generalization transition, we perform a fine-grained search of 30 linearly spaced points, from $A$ to $B$. For regions outside of the transition, using linearly spacing, we sample 5 points from $|S| = 2$ to point $A$, and another 5 points from $B$ to the total dataset size $|S|$. In other words, we train a separate model for each selected data size and generate the corresponding evaluation and synthetic sets. Then, we compute the memorization, spurious, and generalization fractions, see Figure 2. Due to limited sample size, we employ the bootstrapping method [Efr83, ET94] to obtain more robust estimates of these fractions. Specifically, for each data size points $\alpha = 1 \ldots M$, we perform 10k bootstrapping iterations, where in each iteration, we randomly select 10k samples from each $S_\alpha^{\text{eval}}$ with replacement and compute each fraction. We report the mean and the standard deviation as error bar, computed from bootstrapping, for each data size point in Figure 2. However, we must note that the error bar is too small to be seen.

| **Dataset** | **Start Point** (Data Size) | **Point A** (Data Size) | **Point B** (Data Size) | **End Point** (Data Size) |
|---|---|---|---|---|
| CIFAR10 | 2 | 2,000 | 16,000 | 50,000 |
| LSUN-Church | 2 | 2,000 | 16,000 | 126,227 |
| FashionMNIST | 2 | 2000 | 32,000 | 60,000 |
| MNIST | 2 | 2000 | 32,000 | 60,000 |

**Table 1:** Table showing the critical points $A$ and $B$ of each datasets, as well as the starting point and ending point or the total dataset size. We use 30 linearly spaced points between $A$ and $B$, and 5 points for the other two ends.

### B.2 Detection Details

For high dimensional datasets, CIFAR10 and LSUN-Church, we utilize LPIPS [ZIE+18] as the function $d(\cdot, \cdot)$, with AlexNet [KSH12] as the backbone, for both memorization (4) and spurious (5) detection metrics. LPIPS is a commonly used perceptual metric that compares the similarity between two images based on their deep feature representations, offering a more nuanced evaluation of image similarity than pixel-wise comparisons. It has been shown to better align with human judgment of visual similarity, making it ideal for assessing the quality and diversity of generated samples in these high-dimensional image datasets. For simpler datasets like MNIST and FashionMNIST, where images are single-channel and less complex, we found that $l2$-distance suffices for both memorization and spurious detection tasks.

For the consideration of the detection thresholds, we first focused on selecting an optimal value for $\delta_m$ since [YCKR23, CHN+23] have found the memorization metric (4) to be very robust and accurate in identifying memorized samples. Consequently, similarly to [YCKR23], we have found that $\delta_m = \frac{1}{3}$ works well with the four datasets. To select the optimal spurious threshold, we first detect the set of memorized samples across $M$ data sizes and set our initial $\delta_s = 0.3$. We tune this threshold by manually inspecting at the least 10 spurious samples, identified via the spurious metric (5), across all $M$ sizes. If, on average, there are a minimum of 9 well-looking spurious samples, we select that $\delta_s$ as the optimal value; otherwise, we decrease it and repeat the process again. Overall, we selected $\delta_s = 0.28$ for MNIST and $\delta_s = 0.3$ for FashionMNIST; $\delta_s = 0.175$ for CIFAR10 and $\delta_s = 0.25$ for LSUN-Church.

### B.3 Model Training Details

For each point in our transition plots in Figure 2, we train a DDPM-based diffusion model, where the score model is a PixelCNN++ based Unet [VdOKE+16, SKCK17]. We keep the variances, $\beta_{\text{min}} = 10^{-4}$ and $\beta_{\text{max}} = 2 \times 10^{-2}$, timesteps $T = 1000$, and learning rate $lr = 2 \times 10^{-4}$ for all models and datasets. Each model has 2 residual blocks [HZRS16] for each down- and up- sampling layer, while an attention block is placed at 16x resolution. We only modified the channel multipliers for each model based on the complexity of the dataset, see Table 2. For generation or inference, we use the exponential moving average (EMA) of each trained model, as delineated in [HJA20], which was obtained with the decay value set as 0.9999 during training. We did not use random flipping in training our models, but we did use dropout (of value 0.1) for the training of CIFAR10 models. Lastly, for of the each training set $S_i$, they were split from the original dataset given a specified size $n$, and using the same random seed value for all data splits.

| Dataset | Initial Latent | Channel Multipliers | Number of Parameters | Batch Size | Training Iterations |
|---|---|---|---|---|---|
| CIFAR10 | 128 | (1, 2, 2, 2) | 35.7M | 128 | 500,000 |
| LSUN-Church | 96 | (1, 1, 2, 2, 4, 4) | 61.7M | 64 | 800,000 |
| FashionMNIST | 128 | (1, 2, 2) | 24.5M | 128 | 400,000 |
| MNIST | 128 | (1, 2, 2) | 24.5M | 128 | 400,000 |

**Table 2:** Table displaying both model and training configurations for each dataset.

## C Diffusion Models and Hopfield Models

### C.1 Diffusion Models

Given a dataset of i.i.d. samples $\mathbf{x}_0$ drawn from an unknown data distribution $p(\mathbf{x}_0)$, diffusion models are a class of generative models, which aims to approximate $p(\mathbf{x}_0)$ by placing a reversible process that maps data to noise and back. The mapping to noise (or forward process) is described by the following stochastic differential equation (SDE) [SSDK+21],

$$\mathrm{d}\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)dt + g(t)\mathrm{d}\mathbf{w}_t, \tag{7}$$

which transforms the given data distribution into a simpler distribution, e.g., isotropic Gaussian distribution. Here, $\mathbf{f}(\mathbf{x}_t, t)$ represents the drift term which guides the diffusion process while $g(t)$ represents the diffusion coefficient that controls the noise at each time step $t$. In contrast, the reverse process, removes the injected noise at each $\tau = T - t$ step and it is described as,

$$\mathrm{d}\mathbf{x}_\tau = [g(\tau)^2 \nabla_{\mathbf{x}_\tau} \log p(\mathbf{x}_\tau, \tau) - \mathbf{f}(\mathbf{x}_\tau, \tau)]\mathrm{d}\tau + g(\tau)\mathrm{d}\mathbf{w}_\tau, \tag{8}$$

where $\nabla\tau$ is an infinitesimal positive step. To effectively solve this equation, it is crucial to reliably estimate the score $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$. This is done via the parameterization of $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ as a neural network $s_\theta(\mathbf{x}, t)$, where $\theta^*$ is obtained using methods for denoising score matching across multiple times steps [Hyv05, Vin11, SE19]. The general description of this optimization problem, given by [YZS$^+$23], is formulated as

$$\theta^* = \arg\min_\theta \mathbb{E}_{t\sim\mathcal{U}(1,T),\, \mathbf{x}_0\sim p(\mathbf{x}_0),\, \mathbf{x}_t\sim p(\mathbf{x}_t|\mathbf{x}_0)} \left[\lambda(t)\|s_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log p_{0,t}(\mathbf{x}_t|\mathbf{x}_0)\|^2\right] + C, \quad (9)$$

where $p(\mathbf{x}_t|\mathbf{x}_0)$ is the forward process, $\lambda(t)$ is a positive weighting function, $\mathcal{U}(1, T)$ is a uniform distribution over the set $\{1, 2, \ldots, T\}$, and $C$ is a constant which does not depend $\theta$.

### C.2 Diffusion Models as Hopfield Models

Hopfield networks are the original energy-based systems, which seek to store data patterns as memories and perform accurate retrievals. These systems utilize an explicit formulation of an energy function, with properties that define them as a dynamical system, to perform pattern recall. The energy function of the classical system, the Hopfield network [Hop82], is described as $E(\sigma) = -\frac{1}{2}\sigma^\top \mathbf{W}\sigma$, where $\mathbf{W} = \sum_j^N \mathbf{y}_j\mathbf{y}_j^\top$ is the memory matrix which stores patterns $\mathbf{y}_j \sim \mathcal{D}$ through Hebbian learning and $\sigma$ is a binary query to the system. Although expressive, the Hopfield network can only store a small number of patterns ($\approx 0.14N$) and consequently, often recall false or spurious patterns [AGS85, AMSJ85]. However, with recent revisions, such networks have evolved into Dense Associative Memory or Modern Hopfield networks, capable of storing exponential number of patterns [KH16, DHL$^+$17, RSL$^+$21, KH21]. Specifically, for the model with the softmax activation function the energy function is given by

$$E(\mathbf{x}) = -\beta^{-1} \log\left(\sum_{j=1}^N e^{\beta \mathbf{x}^\top \mathbf{y}_j}\right) + \frac{1}{2}\|\mathbf{x}\|_2^2, \quad (10)$$

where $\beta$ is the inverse temperature which controls the overlap among the memories $\mathbf{y}_j$ —is theoretically linked to the attention mechanism found in transformers [RSL$^+$21, HLP$^+$23].

Meanwhile, generative diffusion models can be interpreted as modern associative memory networks, due to their strong link to energy-based models [SE19]. As demonstrated in [Amb24, RA24], the energy of diffusion models $E_{\text{DM}}(\mathbf{x}, t)$ can be described as

$$E_{\text{DM}}(\mathbf{x}, t) = -\sigma^2 \log p_t(\mathbf{x}) = -\sigma^2 \log \mathbb{E}_{\mathbf{y}\sim p_{\text{data}}}\left(e^{-\frac{\|\mathbf{x}-\mathbf{y}_j\|_2^2}{2\tau\sigma^2}}\right) + C, \quad (11)$$

where $\tau = T - t$ and the constant $C$ can be omitted as it does not depend on $x$. By assuming the stored patterns are normalized such that $\|\mathbf{y}_j\|_2^2 = 1$ and divide $E_{\text{DM}}(\mathbf{x}, t)$ by $\sigma^2$, it leads to

$$\frac{E_{\text{DM}}(\mathbf{x}, t)}{\sigma^2} = -\log\left(\sum_{j=1}^N e^{\frac{\mathbf{x}^\top \mathbf{y}_j}{\tau\sigma^2}}\right) + \frac{\|\mathbf{x}\|_2^2}{2\tau\sigma^2}, \quad (12)$$

where setting $\beta^{-1}(t) = (T - t)\sigma^2 = \tau\sigma^2$ and multiply it to both sides yields

$$\begin{aligned} E_{\text{DAM}}(\mathbf{x}, t) &= \beta^{-1}(t)\frac{E_{\text{DM}}(\mathbf{x}, t)}{\sigma^2} \\ &= -\beta^{-1}(t) \log\left(\sum_{j=1}^N e^{\frac{\mathbf{x}^\top \mathbf{y}_j}{\tau\sigma^2}}\right) + \frac{1}{2}\|\mathbf{x}\|_2^2 \\ &= -\beta^{-1}(t) \log\left(\sum_{j=1}^N e^{\beta(t)\,(\mathbf{x}^\top \mathbf{y}_j)}\right) + \frac{1}{2}\|\mathbf{x}\|_2^2, \end{aligned} \quad (13)$$

as the energy of the Dense Associative Memory network with the softmax activation function. As a note, in general $\beta(t)$ is fixed in Hopfield networks and thus the energy minimization is deterministic unlike the stochasticity exhibited in the dynamics of diffusion models [Amb24].