When rule learning breaks: Diffusion Fails to Learn Parity of Many Bits

Anonymous Author(s)
Affiliation
Address
email

Abstract

Diffusion models can generate highly realistic samples, but do they learn the latent rules that govern a distribution, and if so, what kind of rule can they learn? We address this question using a controlled *group-parity* benchmark on 6×6 binary images, where each group of G bits must satisfy an even-parity constraint. This setup allows us to precisely tune rule complexity via G and measure both correctness and memorization at the group and sample levels. Using EDM-parameterized Diffusion Transformers of varying depth, we find: (i) learnability depends jointly on G and depth, with deeper models extending—but not eliminating—the range of learnable rules; (ii) successful rule learning exhibits a sharp early transition in accuracy that precedes memorization, creating a temporal window for generalization; (iii) memorization onset follows a steps-per-sample scaling law and is delayed by larger datasets. Theoretically, an energy/score analysis explains the depth dependence through the global multiplicative term in the parity score. Together, these results offer a principled testbed and new insights into the interplay between rule complexity, rule learning, and memorization in diffusion models.

1 Motivation

Recent diffusion models generate strikingly realistic samples across images, audio, and science data. Yet beyond perceptual quality lies a scientific question: *do these models internalize latent rules that govern a data distribution and generate accordingly, if so, what kind of rule can they learn*? Answering this requires tasks where the underlying structure is precise, global, and tunable.

We study this question through the lens of *parity*, a canonical discrete rule that couples many variables multiplicatively and is known to be challenging to learn. Concretely, we construct a controlled benchmark of 6×6 binary images partitioned into D/G disjoint groups of size G, where each group must satisfy even parity. This setting lets us probe whether unconditional diffusion models can (i) learn and *enforce* a global, non-local rule; (ii) recombine valid parts into novel solutions; and (iii) avoid overfitting individual training examples.

Two features make this benchmark especially revealing. First, rule complexity is tunable via the group size G: small G requires only local interactions, whereas large G demands long-range multiplicative dependencies across many bits. Second, we can cleanly separate *correctness* from *memorization*. We evaluate both per-group and per-sample parity accuracy, and we measure memorization at the group and sample levels by exact match against the training set. This enables a direct view of "creativity" as correct but *novel* generations that were not seen during training.

Using EDM-parameterized diffusion transformers (DiT) with controlled depth and capacity, we uncover three consistent phenomena. (1) **Learnability depends jointly on rule complexity and depth.** Small G is learned robustly, while accuracy collapses as G increases; deeper DiTs push the frontier of learnable G but do not eliminate the barrier. (2) **There is a sharp, early *rule-learning transition* that precedes memorization.** When learning succeeds, accuracy rises abruptly well before memorization increases, yielding a clear temporal separation that supports early stopping to preserve generalization. (3) **Memorization follows a steps-per-sample law.** Its onset is largely synchronized across G and aligns when time is measured as gradient updates per example; larger datasets reliably delay memorization without slowing rule acquisition. A simple energy/score analysis clarifies why: the parity score contains a degree-G monomial coupling all bits, which is poorly aligned with the predominantly pairwise structure induced by a single attention block, making depth helpful but not universally sufficient.

2 Backgrounds

Rule learning in Diffusion models There have been several works along this direction. Wang et al. (2024) showed that unconditional diffusion models can learn to generate according to some of the rules in RAVEN's progression matrices encoded as integer arrays, but not all of them. In particular, rules such as the logical operation (AND, OR, XOR) over sets of attribute have been showed to be hard to learn. Similarly, Han et al. (2025) examined rule learning in the pixel space, where the diffusion models can learn the coarse proportional relationship between bars and shadows length, but not precisely, there is usually a non-zero error from the precise rule defined in training set. These previous works prompt this study, where we want to examine what kind of rules can be learned. In this case, we focus on the discrete and abstract rule case.

Memorization and Creativity in Diffusion models The question of when diffusion models are able to generate genuinely novel samples matters both scientifically and for mitigating data leakage. From the rule learning perspective, the model that truly learn the rule should not simply recapitulate the training set, but learn the data manifold underlying it. From the score-matching perspective, if the learned score exactly matches that of the empirical data distribution, then the reverse process reproduces that empirical distribution, and thus does not create new samples beyond the training set (Kamb & Ganguli, 2024; Li et al., 2024; Wang & Vastola, 2024). Yet high-quality diffusion models routinely generate images that are not identical copies of images from the training set. Kamb & Ganguli (2024) take an important step toward reconciling this: when the score network is a simple CNN, its inductive biases (locality and translation equivariance) favor patch wise composition, enabling global samples that are novel while remaining locally consistent "mosaics." Similarly, in Wang & Pehlevan (2025), they noticed score networks with different architectural constraints will learn various approximation of the dataset, and therefore generalize: e.g. linear networks learn the Gaussian approximation, and circular convolutional networks learn the stationary Gaussian process approximation. In Finn et al. (2025), they analyze attention-based diffusion and provide evidence that adding a final self-attention layer promotes global consistency across distant regions, organizing locally plausible features into coherent layouts that move beyond purely patch-level mosaics. Related theoretical work further probes why well-trained diffusion models can generalize despite apparent memorization pressures (Biroli et al., 2024; J. Vastola, 2025; Chen, 2025). These results suggest that departures from exact empirical-score fitting—mediated by inductive biases (both architectural and training dynamics) can explain how diffusion models avoid pure memorization while maintaining visual plausibility (Ambrogioni, 2023). In this work, we study the memorization and generalization dynamics when we have access to the underlying distribution is tractable.

Learning parity We focus on parity learning, a versatile testbed that has been widely adopted for understanding both the representational and learning aspects of neural networks (Hahn, 2020; Bhattamishra et al., 2022; Glasgow, 2024; Abbe et al., 2024, 2025). The hardness of parity depends on the number of bits that the parity is computed over, where more bits require a higher boolean sensitivity or larger weight norms in the case of neural networks. For Transformers specifically, learning parity requires growing the MLP norms (Liu et al., 2022; Hahn & Rofin, 2024) and the use of normalization layers (Hahn, 2020; Yao et al., 2021; Chiang & Cholak, 2022). Even when a network is sufficiently expressive, parity is computationally challenging to learn (Kearns, 1998; Barak et al., 2022; Edelman et al., 2023; Wen et al., 2024; Kim & Suzuki, 2025). In this work, we explored parity learning from generative modeling perspective, leveraging this well studied problem to characterize how much modern generative modeling framework, in particular diffusion, could learn these underlying structures.

3 Methods

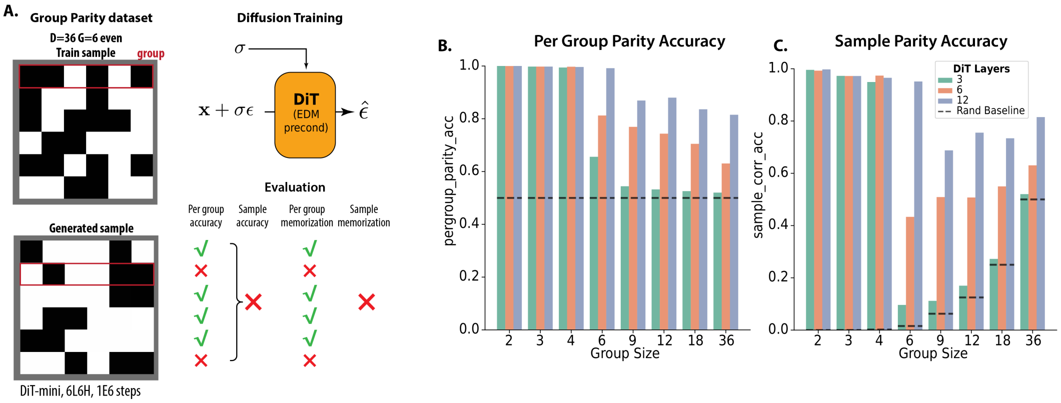

Figure 1: **Schematics of the task and performance evaluation. A.** Structure of the Group Parity dataset and evaluation setup. Each D-dim, 6×6 binary image is divided into G equal-sized groups (here $D = 36$, $G = 6$), with each *group* sufficing even parity (even number of black pixel). DiT models are trained on this dataset, where the generated samples are evaluated by the per group accuracy, sample accuracy and memorization ratio of group and sample. **B.** Per-group parity accuracy as a function of group size for DiT models with 3, 6, or 12 layers (6 heads, 384 d), compared to a random baseline (dashed line), dataset size $N = 4096$. **C.** Sample-level parity accuracy for the same models and group sizes.

Notation Define $\mathcal{S}_d^+ = \{\mathbf{x} \mid \prod_{i=1}^d x_i = 1, x_i \in \{-1, 1\}\}$ to be the set of bit strings with even parity in a d-dimensional boolean cube $\{-1, 1\}^d$; note that $|\mathcal{S}_d^+| = 2^{d-1}$. For example, $\mathcal{S}_3^+ = \{(1, 1, 1), (-1, -1, 1), (-1, 1, -1), (1, -1, -1)\}$. Define $\mathcal{P}_d^+(\mathbf{x}) = |\mathcal{S}_d^+|^{-1} \sum_{\mathbf{y} \in \mathcal{S}_d^+} \delta(\mathbf{x} - \mathbf{y})$, be the mixture of delta measure at all points of the set $\mathcal{S}_d^+$. Further, we define $(\mathcal{S}_d^+)^m = \mathcal{S}_d^+ \times ... \times \mathcal{S}_d^+ \subset \{-1, 1\}^{md}$, where the d bits in each of the m groups satisfy parity relation, and $|(\mathcal{S}_d^+)^m| = (2^{d-1})^m$. We denote the uniform measure corresponding to the set $(\mathcal{S}_d^+)^m$, $(\mathcal{P}_d^+)^m$. Define $\mathcal{U}_d$ as the uniform measure on d-dimensional Boolean cube.

Dataset Design We construct samples $\mathbf{x} \in \mathbb{R}^D$ with length $D = 36$, every G elements are assigned to a group, and $\mathbf{x} \in (\mathcal{S}_G^+)^{D/G}$. Specifically, each group is sampled i.i.d. from the even parity string of length G, $\sim \mathcal{P}_G^+$, and then the D/G groups are concatenated as a sample $\mathbf{x}$ (Fig.1**A**). Then we generate N samples as our training set, where we mandate all training samples to be unique by rejective sampling (though individual groups could repeats). The key design parameter for the dataset will be D, G, N. For diffusion training, we reshape each sample to a 6×6 pixel single channel image.

Model Architecture As a generative modeling problem, we consider the dataset in the continuous space $\mathbb{R}^D$, and solve it with Gaussian diffusion models. Specifically, we used the continuous-time EDM diffusion framework (Karras et al., 2022), and used diffusion transformer (DiT) (Peebles & Xie, 2023) as our function approximator with EDM preconditioning. We started with baseline version `DiT-mini` with 6 layers, 6 heads and hidden dimension 384, and varied the layers (in $\{3, 6, 12\}$) and sizes of the model to examine and effect of model capacity. We keep patch size as 1 to maximize the capacity of attention to model the relation between bits.

Training On every dataset, we use Adam optimizer to train DiT models 10^6 steps , with learning rate 10^{-4} and batch size 256.

Evaluation Throughout training, we generate samples with Heun's 2nd order deterministic sampler (Karras et al., 2022), and evaluate them according to following criterion. First, we evaluate how far the samples are from boolean cube $\{-1, 1\}^D$, and computed $d_{\ell_\infty}(\mathbf{x}) = \max_i ||x_i| - 1|$. We call the samples *invalid* when $d_{\ell_\infty}(\mathbf{x}) > \epsilon$, and calculated the invalid fraction for various $\epsilon = 0.1, 0.01$.

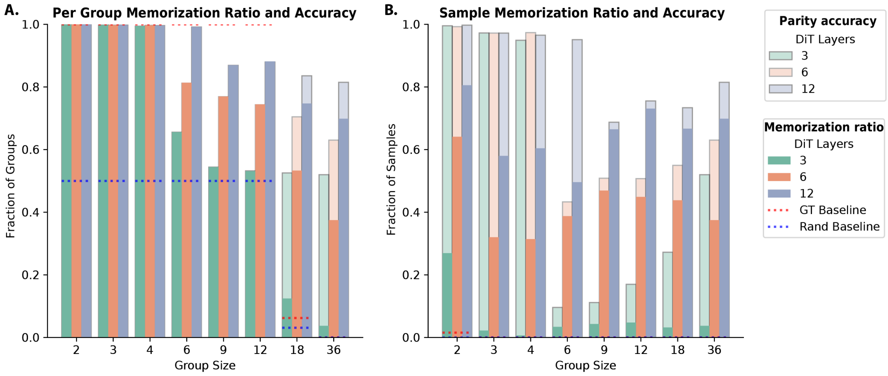

Figure 2: **Memorization and Creativity in Parity Learning.** Memorization ratio overlay on accuracy, for different group size and DiT depth, at group level (**A.**) and sample level (**B.**), with underlying bar plot the same as Fig.1 **B.C.** Red dashed line shows the memorization ratio of the ground truth distribution ($\mathcal{P}_G^+$ for groups, and $(\mathcal{P}_G^+)^{D/G}$ for samples); and blue dashed line shows the memorization ratio of the chance distribution ($\mathcal{U}_G$ for groups and $\mathcal{U}_D$ for samples).

Next, we binarize each element of the sample to $\{-1, 1\}$ and evaluate the binarized sample $\bar{\mathbf{x}}$ according to the parity of groups and samples (Fig. 1**A**). For each group of G elements, we evaluate the correctness of parity of the group, i.e. group parity accuracy, with chance level 2^{-1}. Then for the whole sample, we call it correct if the parity of every D/G group is correct, i.e. sample parity accuracy, for which the chance level is $2^{-D/G}$.

Further, we examine the fraction of the generated samples or groups coincide with those in the training dataset. We term these the *memorization ratio* of group and samples. If we assume the model learned the true distribution i.e. uniform measure on $(\mathcal{S}_G^+)^{D/G}$, then the chance level of sample memorization ratio will be $N/(2^{G-1})^{D/G} = N \cdot 2^{-\frac{G-1}{G}D}$. If the model learned the uniform measure on the boolean cube, then the chance level of sample memorization will be $N \cdot 2^{-D}$.

4 Results

4.1 Parity learning depends on both rule complexity and model depth

Effect of rule complexity. For small group sizes ($G = 2, 3, 4$), all DiT variants achieve near-perfect parity accuracy, indicating that parity rules among few bits are readily captured. However, as G exceeds 6, both per-group and sample-level accuracies (Figure 1**B,C**) decline sharply, with the per-group accuracy deteriorates towards chance level 0.5, and the sample level accuracy degrades towards chance performance at $(1/2)^{D/G}$. This aligns with prior work highlighting the inherent difficulty of learning parity with many interacting bits (Kearns, 1998; Barak et al., 2022; Abbe et al., 2024).

Effect of model depth. Model depth mitigates this difficulty: across G, deeper DiTs consistently outperform shallower ones in both metrics, while we kept head number 6 and latent dimension 384 the same. Notably, a 12-layer DiT consistently achieves near-perfect sample and per-group accuracy for $G = 6$, while 3- and 6-layer models degrade substantially and only learns parity rules up to $G = 4$. Even for large G, deeper networks remain above chance, whereas shallower ones collapse to the chance level of $\mathcal{U}_D$. These results suggest limitation in shallow transformer of learning parity rules of many bits, and that greater depth enhances the ability to integrate information across many bits, increasing capacity for representing the global multiplicative interactions required for large-G parity rules.

4.2 Memorization and creativity of parity learning

When a model trained on *finite data* succeeds in generating samples consistent with a given parity rule, an immediate follow-up question is: how many of these samples are exact reproductions from the training set, and how many are genuinely *novel*? This relates directly to the notion of *creativity* in generative models (Kamb & Ganguli, 2024). Given the hierarchical nature of our data, novelty can

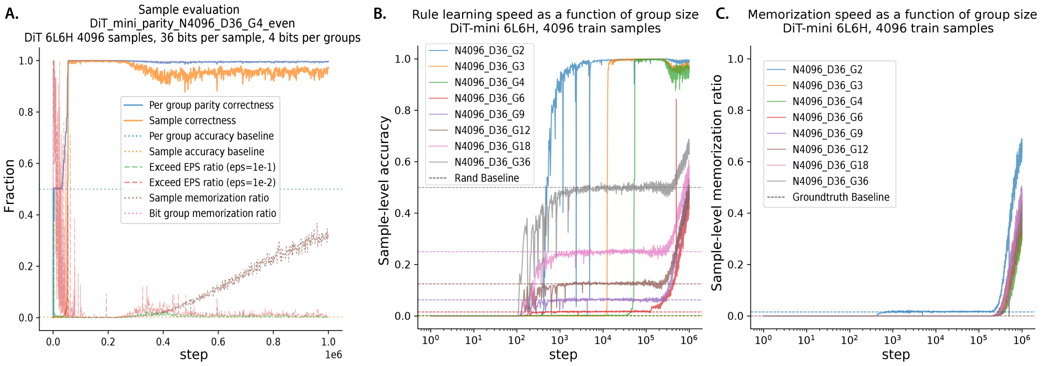

Figure 3: **Learning dynamics of rule acquisition and memorization across parity complexities.** **A.** Training-time evaluation for DiT-mini (6L6H) on the $G = 4$, $N = 4096$ dataset. Invalid-sample ratios decay rapidly, followed by an early, sharp rise in parity accuracy to near-perfect levels. Much later, sample memorization ratio grows steadily until the end of training, following a small bump in invalid ratio (EPS=0.01). **B.** Sample-level accuracy during training for datasets with group sizes $G \in 2, 3, 4, 6, 9, 12, 18, 36$ ($N = 4096$, DiT-mini). Dashed lines indicate random-chance baselines $2^{-D/G}$. For small G, rule learning occurs via a sharp transition, with the transition point shifting later as G increases. For larger G, accuracy rises above chance only at a much later stage, following a gradual process driven primarily by memorization. **C.** Sample-level memorization ratio during training for the same datasets, with dashed lines showing the expected memorization ratio under the ground-truth distribution. Memorization emerges late in training, at similar time across group size G.

be evaluated at two levels: (1) the fraction of *samples* reproduced from the training set, and (2) the fraction of *bit groups* reproduced from the training set.

At our standard dataset size of $N = 4096$, for $G \leq 12$ the training set contains *all* valid even-parity groups. In this regime, novelty at the group level is impossible—any correct group must have appeared in training. The only possible form of creativity is *combinatorial*: assembling previously seen valid groups into novel combinations to form new valid samples.

Combinatorial creativity when rule learning succeeds For small group sizes ($G = 2, 3, 4$), all model variants achieve near-perfect sample accuracy while generating a substantial fraction of *novel* correct samples via recombination (Figure 2**B**). Similarly, when trained on $G = 6$ dataset, over 50% of the 12-layer DiT's generations are novel and correct, indicating strong generalization through recombination rather than pure memorization. On the other hand, the sample memorization ratio is still much higher than the ground truth distribution $(\mathcal{P}_G^+)^{D/G}$ (recall that $D = 36$), showing that the learned distribution still bias towards the combinations encountered in training set.

Deeper models memorize more. Across all datasets and group sizes, deeper models consistently exhibit higher memorization ratios at both the group and sample levels. For $G = 2, 3, 4$, this means that, at matched sample accuracy, deeper models are *less* creative—generating fewer novel combinations—despite achieving the same correctness. This aligns with the broader observation that larger-capacity models tend to memorize more easily (Carlini et al., 2022; Tirumala et al., 2022; Morris et al., 2025).

Memorization under partial rule learning. For $G = 6, 9, 12$, the training set still contains all valid groups, yet models fail to memorize them all—resulting in imperfect per-group accuracy (with the exception of the 12-layer DiT at $G = 6$, which learns the rule fully). This gap is possibly due to the sheer number of valid patterns (2^{G-1}) can exceed the model's memorization capacity for groups.

For $G = 18, 36$, the training set covers only a small fraction of all valid groups, so even an optimal generalizing distribution would have group-level memorization ratios well below 1. Nonetheless, we observe ratios substantially above baseline, indicating a preference for groups seen in training-set over unseen valid ones. The residual gap between parity accuracy and memorization ratio is consistent with chance-level correctness for non-memorized groups (i.e., $\approx 50\%$ parity accuracy in that subset).

4.3 Learning dynamics of generalization and memorization

Next, we examined the learning dynamics of generalization and memorization.

Sharp Rule-Learning Transition Precedes Memorization. Across all datasets where the parity rule is successfully acquired (small to moderate G), we observe a clear temporal separation between the onset of learning parity rule and memorization (Fig. 3A). Initially, the fraction of invalid samples drops rapidly, and sample- and group-level parity accuracies undergo an abrupt, early transition from chance level to near-perfect performance. During this early phase, memorization ratios remain near baseline, indicating that the model has learned to produce valid, rule-conforming samples without simply reproducing training examples. Only much later in training does memorization begin to increase, suggesting that the model first discovers a generative rule-consistent model $(\mathcal{P}_G^+)^{D/G}$ and subsequently drifts toward reproducing specific training samples within it. This clear separation in timescales reinforces prior suggestions that early stopping during diffusion training can preserve the generalizing solution before memorization dominates (Bonnaire et al., 2025).

Higher-bit parity delays rule-learning transition When the group size G increases, the sharp accuracy transition is systematically delayed (Fig. 3B). For small G (≤ 4), this jump in accuracy occurs within the first few thousand steps. In contrast, for large G, accuracy remains at chance for an extended period before eventually rising. In extreme cases (e.g., $G \geq 18$), this increase is not due to genuine rule learning but rather to a slow, memorization-driven improvement at late training stages. This pattern reflects the greater difficulty of learning high-order parity relations, which require integrating information multiplicatively across many bits.

Sample memorization emerges at similar times across G. Interestingly, the onset of sample-level memorization is largely independent of G (Fig. 3C). Across all datasets, memorization begins only after a prolonged period of stable accuracy—whether that accuracy was achieved through genuine rule learning (small G) or remains near chance (large G). The synchronized late rise in memorization suggests that it is governed more by total optimization time and model capacity than by rule complexity, consistent with a gradual overfitting process that unfolds after the model has stabilized its score estimates for the training distribution.

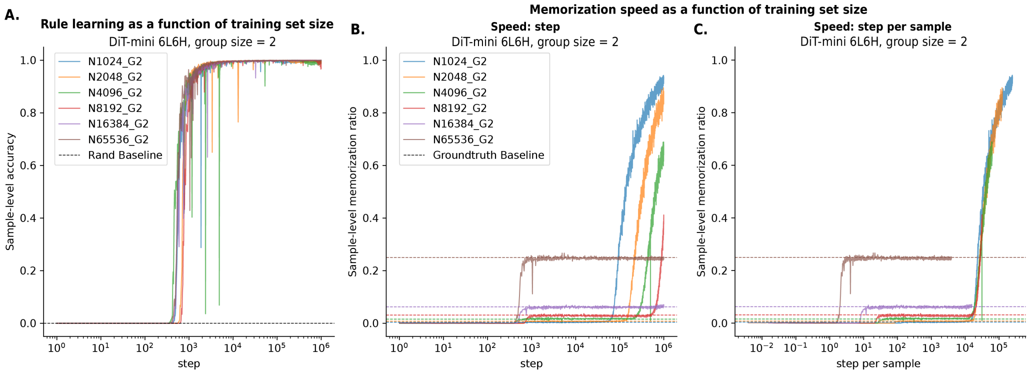

Figure 4: **Learning dynamics of rule acquisition and memorization across dataset size. A.** Dynamics of sample parity accuracy across dataset scale, at $G = 2$, DiT-mini. **B. C.** Dynamics of sample memorization ratio across dataset scales, the dynamics are plotted as a function of step (**B.**) and step per sample (step x batch size/ dataset size) (**C.**). Colored dashed lines denotes the memorization ratio expected from the ground truth distribution.

4.4 Scaling law of generalization and memorization

Is the difficulty of rule learning due to limited dataset size? We next investigate how dataset size N affects the dynamics of rule learning and memorization (Fig. 4, Fig. 5).

Rule learning dynamics are relatively invariant to dataset size. Across dataset scales ranging from $N = 1{,}024$ to $N = 65{,}536$, sample-level accuracy follows a similar trajectory as a function of training step (esp. nearly identical for $G = 2, 3$, Fig. 4A). All curves exhibit the same early, sharp transition from chance to near-perfect accuracy, indicating that the onset and speed of rule acquisition are essentially independent of the number of training samples—at least for small G where the rule is consistently learnable (Fig. 6 7 for all group sizes G). When rule complexity is on the edge of learnability ($G = 6$), increasing dataset scale can help or hinder rule learning (Fig. 6).

Memorization is delayed by larger datasets. In contrast, sample-level memorization shows a strong dependence on dataset size (Fig. 4B). Across rule complexity G, larger datasets consistently postpone the onset of memorization to later training steps, Specifically, at our largest dataset scales ($N = 16384, 65536$), excessive memorization do not happen, and the memorization ratio stays at the expected level from ground truth. When we rescaled each memorization curve by N, (Fig.4C), the memorization curves align well as a function of "steps per sample" (i.e. step × batch size $/N$). This alignment suggests that the key parameter governing sample memorization is the number of gradient steps per training example rather than the raw step count. On our dataset, this memorization happens around $\sim 10^4$ steps per example.

Implications for training strategy. These results suggest that, for learnable rules, increasing dataset size does not hinder the model's ability to acquire the underlying structure but can substantially extend the generalization phase before overfitting begins. This reinforces the view—also supported by our temporal scale separation results—that early stopping can preserve a generalizing solution, and that larger datasets naturally widen the safe window before memorization dominates.

5 Why Diffusion Transformers Struggle with Parity

We explored this question through the corresponding energy of parity.

Continuous-space Energy model of Parity Consider the following energy,

$$E_d(\mathbf{x}) = \sum_{i=1}^{d}(x_i^2 - 1)^2 \; + \; \lambda_p \big(\prod_{i=1}^{d} x_i - 1 \big)^2 \;\; , \;\; \lambda_p > 0 \tag{1}$$

The first term encourages $\mathbf{x}$ to be on the boolean cube, and the second term encourages the parity to be 1.The set of all energy minima are the bit sequence with correct parity $\arg\min_{\mathbf{x}} E_d(\mathbf{x}) = \mathcal{S}_d^+$. Further, one can show that $p_d(\mathbf{x}; \beta) \propto \exp(-\beta E_d(\mathbf{x}))$ when $\beta \to \infty, p_d(\mathbf{x}) \to \mathcal{P}_d^+(\mathbf{x})$, as the distribution converged to a uniform distribution over the minima.

Thus, at low noise regime ($\beta \to \infty$), the score network in the diffusion model should be approximating the gradient of the energy, i.e. the score reads

$$\nabla \log p_d(\mathbf{x}; \beta) = -\beta \nabla_{\mathbf{x}} E_d(\mathbf{x}) \tag{2}$$

$$\nabla_{x_i} E_d(\mathbf{x}) = 4x_i(x_i^2 - 1) + 2\lambda_p \prod_{j=1, j\neq i}^{d} x_j \big(\prod_{j=1}^{d} x_j - 1\big) \tag{3}$$

Notably, the 1st term is a function local to the bit, the 2nd term in the score depends on the product of all d bits, which requires aggregation and broadcasting of global information.

Implication for DiT The local term can be easily learned by the MLP modules which operates on individual bits. Learning the local term effectively push samples onto the boolean hypercube, thus minimize the invalid samples. As we showed empirically, the EPS deviation decays rapidly, so learning this term is efficient and easy.

However, the 2nd term requires multiplication of all G bits. If the attention is becoming one-hot, then each layer only aggregates product information of two bits. Thus deeper transformer models can aggregate more bits through sequential application of attention, explaining their strength in learning rules of higher G. In Appendix B.1, we construct energy functions that correspond to the multi-head self-attention layers and MLP layers which make this point more concrete.

6 Discussion

We introduced a controlled group-parity testbed to probe whether diffusion models can learn and generalize precise rules of different global levels. Across variations in model depth, dataset size, and rule complexity (G), we found a clear learnability threshold that shifts with depth; a consistent temporal separation between an early *rule-learning transition* and later memorization; before memorization start, models learning low rule complexity exhibit combinatorial creativity and discover the ground truth distribution; further the memorization onset time scale linearly with dataset size. Our energy/score analysis tied the observed depth dependence to the degree-G multiplicative interaction term in the parity score, which is misaligned with the predominantly pairwise interactions that a single attention block can directly encode.

Score complexity and spectral bias. The parity score naturally decomposes into a local term and a global multiplicative term $\prod_{j=1}^{G} x_j$, whose polynomial degree grows with G. In the Fourier/Walsh–Hadamard basis, higher-degree interactions correspond to higher-frequency components, and it is well established that neural networks exhibit a *spectral bias*, fitting low-frequency components before high-frequency ones (Canatar et al., 2021; Wang & Pehlevan, 2025). This framework offers a natural explanation for our learning dynamics: small-G components emerge early in training, while large-G components appear only much later—if they appear at all. When the latter are not learned from data, accuracy improvements in late training tend to come from memorization rather than genuine rule acquisition. A more formal score-complexity analysis could help predict the point at which models shift from generalizing to overfitting, and explain how architectural constraints shape this transition.

Implications for natural data. Our findings suggest that relations involving many-way interactions are inherently difficult for current diffusion architectures. In naturalistic settings, this may underlie the difficulty of learning certain abstract reasoning rules. For example, prior studies on the RAVEN progression matrices found that XOR-type relations over multiple attributes are especially hard for diffusion models (Wang et al., 2024); our results indicate that the same complexity–spectral-bias bottleneck may be responsible. The broader implication is that scientific or physical constraints depending on large-scale multiplicative structure—such as conservation laws involving many coupled quantities—may not be faithfully learned without targeted architectural or training interventions.

Pathways to improved rule learning. The gap between theoretical capacity and observed performance invites several possible remedies. One is to modify the architecture to enable *global broadcasting* of information—through dedicated register tokens, global memory units, or structured multiplicative interactions—so that the model can aggregate and disseminate the features required for large-G parity in a single step. Another is to enrich training with auxiliary objectives that explicitly require detecting and representing parity-like dependencies, such as masked group-parity prediction, to encourage the formation of suitable internal representations. Finally, a curriculum that gradually increases G during training could scaffold the acquisition of higher-order rules, allowing the network to build on simpler cases before tackling more complex ones.

Broader outlook. Although parity is synthetic, it isolates a fundamental limitation: global rules with high interaction order are not well aligned with the inductive biases of current diffusion transformers. Addressing this limitation is critical for applications where rule adherence is as important as perceptual fidelity, including symbolic reasoning, structured design, and scientific modeling. Our group-parity testbed provides a controlled setting in which to explore both the failure modes and potential solutions, and offers a stepping stone toward architectures that can internalize and apply abstract, combinatorial rules from data.

References

Emmanuel Abbe, Samy Bengio, Aryo Lotfi, and Kevin Rizk. Generalization on the unseen, logic reasoning and degree curriculum. *Journal of Machine Learning Research*, 25(331):1–58, 2024.

Emmanuel Abbe, Elisabetta Cornacchia, Jan Hazla, and Donald Kougang-Yombi. Learning high-degree parities: The crucial role of the initialization. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL `https://openreview.net/forum?id=OuNIWgGGif`.

Luca Ambrogioni. In search of dispersed memories: Generative diffusion models are associative memory networks, 2023. URL `https://arxiv.org/abs/2309.17290`.

B. Barak, Benjamin L. Edelman, Surbhi Goel, S. Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: Sgd learns parities near the computational limit. *Neural Information Processing Systems*, 2022. doi: 10.48550/arXiv.2207.08799.

S. Bhattamishra, Arkil Patel, Varun Kanade, and Phil Blunsom. Simplicity bias in transformers and their ability to learn sparse boolean functions. *Annual Meeting of the Association for Computational Linguistics*, 2022. doi: 10.48550/arXiv.2211.12316.

Giulio Biroli, Tony Bonnaire, Valentin De Bortoli, and Marc Mézard. Dynamical regimes of diffusion models, 2024. URL `https://arxiv.org/abs/2402.18491`.

Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mézard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training, 2025. URL `https://arxiv.org/abs/2505.17638`.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12(1):2914, May 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-23103-1. URL `https://doi.org/10.1038/s41467-021-23103-1`.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. *arXiv preprint arXiv: 2202.07646*, 2022.

Zhengdao Chen. On the Interpolation Effect of Score Smoothing, February 2025. URL `http://arxiv.org/abs/2502.19499v1`.

David Chiang and Peter Cholak. Overcoming a theoretical limitation of self-attention. *arXiv preprint arXiv: 2202.12172*, 2022.

Benjamin L. Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Pareto frontiers in neural feature learning: Data, compute, width, and luck. *arXiv preprint arXiv: 2309.03800*, 2023.

Emma Finn, T. Anderson Keller, Manos Theodosis, and Demba E. Ba. Origins of creativity in attention-based diffusion models, 2025. URL `https://arxiv.org/abs/2506.17324`.

Margalit Glasgow. SGD finds then tunes features in two-layer neural networks with near-optimal sample complexity: A case study in the XOR problem. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL `https://openreview.net/forum?id=HgOJlxzB16`.

Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020.

Michael Hahn and Mark Rofin. Why are sensitive functions hard for transformers? *arXiv preprint arXiv: 2402.09963*, 2024.

Yujin Han, Andi Han, Wei Huang, Chaochao Lu, and Difan Zou. Can Diffusion Models Learn Hidden Inter-Feature Rules Behind Images?, February 2025.

John J. Vastola. Generalization through variance: How noise shapes inductive biases in diffusion models, April 2025.

Mason Kamb and Surya Ganguli. An analytic theory of creativity in convolutional diffusion models, 2024. URL `https://arxiv.org/abs/2412.20292`.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022.

Michael Kearns. Efficient noise-tolerant learning from statistical queries. *J. ACM*, 45(6):983–1006, November 1998. ISSN 0004-5411. doi: 10.1145/293347.293351. URL `https://doi.org/10.1145/293347.293351`.

Juno Kim and Taiji Suzuki. Transformers provably solve parity efficiently with chain of thought. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL `https://openreview.net/forum?id=n2NidsYDop`.

Sixu Li, Shi Chen, and Qin Li. A good score does not lead to a good generative model, 2024. URL `https://arxiv.org/abs/2401.04856`.

Bingbin Liu, Jordan T Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. *arXiv preprint arXiv:2210.10749*, 2022.

John X. Morris, Chawin Sitawarin, Chuan Guo, Narine Kokhlikyan, G. Edward Suh, Alexander M. Rush, Kamalika Chaudhuri, and Saeed Mahloujifar. How much do language models memorize? *arXiv preprint arXiv: 2505.24832*, 2025.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers, March 2023. URL `http://arxiv.org/abs/2212.09748`. arXiv:2212.09748 [cs].

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290, 2022.

Binxu Wang and Cengiz Pehlevan. An Analytical Theory of Spectral Bias in the Learning Dynamics of Diffusion Models, March 2025.

Binxu Wang and John Vastola. The unreasonable effectiveness of gaussian score approximation for diffusion models and its applications. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=I0uknSHM2j`.

Binxu Wang, Jiaqi Shang, and Haim Sompolinsky. Diverse capability and scaling of diffusion and auto-regressive models when learning abstract rules, November 2024. URL `http://arxiv.org/abs/2411.07873`. arXiv:2411.07873.

Kaiyue Wen, Huaqing Zhang, Hongzhou Lin, and Jingzhao Zhang. From sparse dependence to sparse attention: Unveiling how chain-of-thought enhances transformer sample efficiency. *International Conference on Learning Representations*, 2024. doi: 10.48550/arXiv.2410.05459.

Shunyu Yao, Binghui Peng, Christos Papadimitriou, and Karthik Narasimhan. Self-attention networks can process bounded hierarchical languages. *arXiv preprint arXiv:2105.11115*, 2021.

A Extended Results

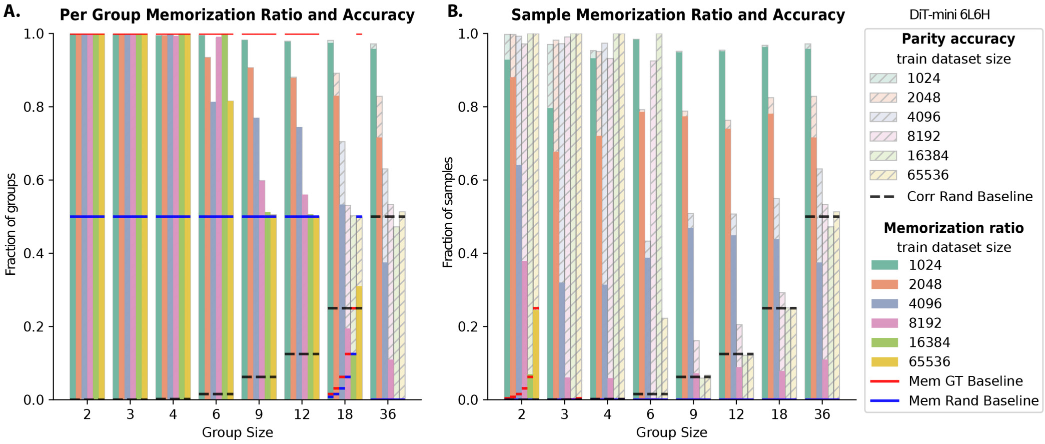

Figure 5: **Memorization and Creativity in Parity Learning across dataset scales.** Memorization ratio overlay on accuracy, for different group size and training dataset scale, at group level (**A.**) and sample level (**B.**), with similar format as Fig. 2. Red solid line shows the memorization ratio of the ground truth distribution ($\mathcal{P}_G^+$ for groups, and $(\mathcal{P}_G^+)^{D/G}$ for samples); and blue solid line shows the memorization ratio of the chance distribution ($\mathcal{U}_G$ for groups and $\mathcal{U}_D$ for samples). Black dashed line shows the chance level accuracy.

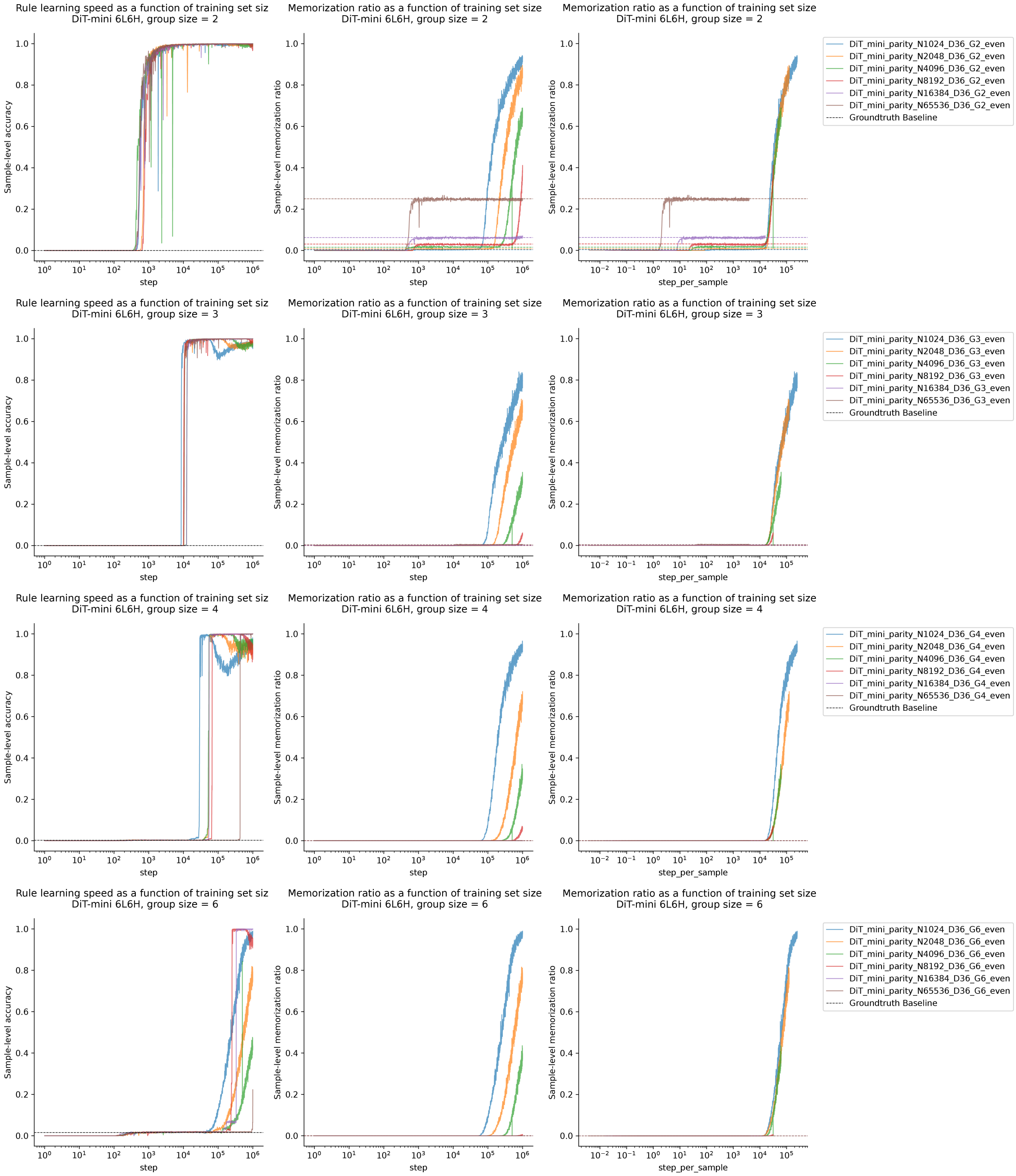

Figure 6: **Learning dynamics of rule acquisition and memorization across dataset size,** $G = 2, 3, 4, 6$**. Left.** Dynamics of sample parity accuracy across dataset scale, DiT-mini. **Mid. Right.** Dynamics of sample memorization ratio across dataset scales, the dynamics are plotted as a function of step (**Mid.**) and step per sample (step x batch size/ dataset size) (**Right.**). Colored dashed lines denotes the memorization ratio expected from the ground truth distribution. Similar format as Fig.4.

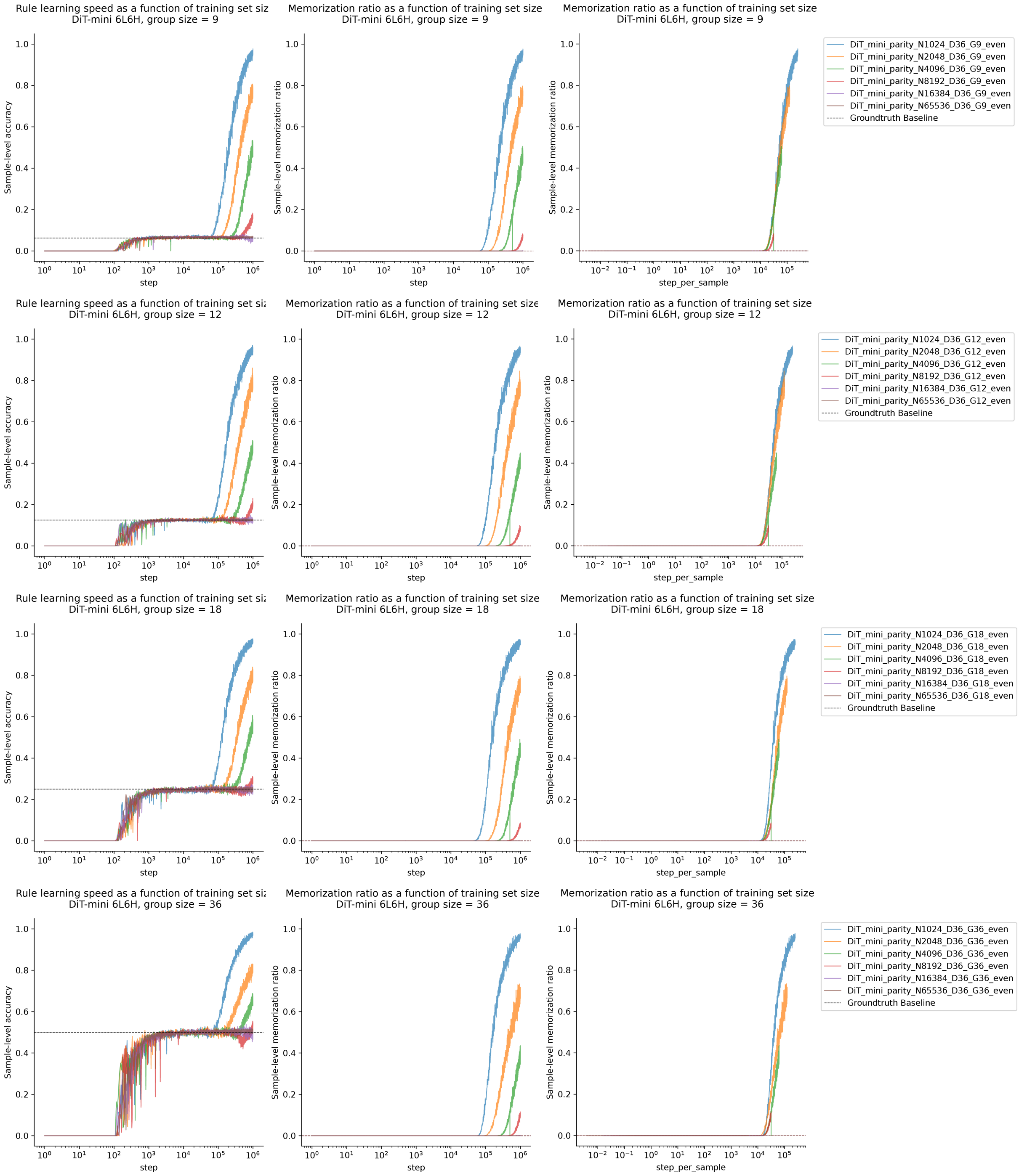

Figure 7: **Learning dynamics of rule acquisition and memorization across dataset size,** $G = 9, 12, 18, 36$**.** Similar format as Fig.6 and 4.

B Extended theory note

B.1 Connection of Self-Attention and Energy

There is a correspondence between the network architecture and the form of distribution which it can represent exactly at a particular depth. The best example of this is the well known correspondence between a Gaussian distribution and a linear score (Wang & Vastola, 2024). In our case, the difficulty of learning binary patterns can be in part explained by the misalignment between the score which a multi-headed attention can express and the true score equation 2 of our parity distribution.

Consider a multi-head attention operation. Let $x_1, \ldots, x_T \in \mathbb{R}^d$ be the token features (for our 6×6 binary images with patch size 1 we have $T = 36$). For head $h \in \{1, \ldots, H\}$, define queries and keys $q_i^{(h)} = W_q^{(h)} \mathbf{x}_i,\ k_j^{(h)} = W_k^{(h)} \mathbf{x}_j$, with head dimension d_h. We simplify our notation and define $W = \frac{1}{\sqrt{d_h}} W_q^\top W_k$.

Consider the multi-headed attention energy

$$E_{\text{multi-head attention}}(\{x_i\}_{i \in \mathcal{I}}) = \sum_{h=1}^{H} \sum_{i=1}^{T} \log \sum_{j=1}^{T} \exp(x_i^\top W_h x_j) \tag{4}$$

Taking the gradient, we find

$$\nabla_X E_{\text{multi-head attention}}(X) \;=\; -\sum_h \Big(W_h\, X\, A_h^\top \;+\; W_h^\top\, X\, A_h \Big), \tag{5}$$

Where $A_{ij} \;:=\; \frac{\exp\left(x_i^\top W\, x_j\right)}{\sum_{k=1}^{T} \exp\left(x_i^\top W\, x_k\right)}$. We combine tokens $X = [\, x_1, x_2, \ldots, x_T \,]$ as a matrix in $\mathbb{R}^{d \times T}$. Then $A \in \mathbb{R}^{T \times T}$ is the attention matrix with entries A_{ij}. In the DiT, it is not necessarily the case that the value matrix is tied to our key matrices, but for simplicity, we assume that it is.

For a full transformer block, we require also an MLP energy form. Consider the energy

$$E(x) \;=\; \sum_{i=1}^{m} \phi\big(w_i^\top x + b_{1,i}\big) \;+\; b_2^\top x, \tag{6}$$

where $x \in \mathbb{R}^d$, $w_i \in \mathbb{R}^d$, $b_2 \in \mathbb{R}^d$, and $b_{1,i} \in \mathbb{R}$. Then

$$\nabla_x E(x) = \sum_{i=1}^{m} w_i\, \phi'\big(w_i^\top x + b_{1,i}\big) \;+\; b_2 \tag{7}$$

Thus, we approximate the overall transformer block energy (ignoring LayerNorm, residuals, and positional terms) by

$$\begin{aligned} E_{\text{transformer block}}(X) &= E_{\mathrm{SA}}(X) + E_{\mathrm{MLP}}(X) & (8)\\ &= \sum_{h=1}^{H} \sum_{i=1}^{T} \log \sum_{j=1}^{T} \exp\big(x_i^\top W_h\, x_j\big) + \sum_{i=1}^{T} \Big(b^\top x_i + \sum_{k=1}^{m} \phi\big(w_k^\top x_i + b_{1,k}\big) \Big) & (9) \end{aligned}$$

Recall that the energy function equation 1 contains interaction terms of degree D. The second term is multiplicative across all coordinates and flips sign with any bit. By contrast, each coordinate of the attention-score is a sum of linear transforms of other tokens, i.e., it is assembled from pairwise inner products. Consequently, a single attention layer cannot represent the global monomial $\prod_{j=1}^{d} x_j$ exactly, while multiple layers can create higher-order interactions compositionally, which helps to explain the improvements seen from increasing the depth of the network.