Do diffusion models generalize on abstract rules for reasoning?

Binxu Wang (binxu_wang@hms.harvard.edu)†
Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University

Jiaqi Shang (jiaqishang@g.harvard.edu)†[1]
Program in Neuroscience, Harvard Medical School

Haim Sompolinsky (hsompolinsky@mcb.harvard.edu)
Center for Brain Science, Harvard University
Edmond and Lily Safra Center for Brain Sciences, Hebrew University

Abstract

Can diffusion models trained on a sample dataset acquire abstract relational reasoning ability? To explore this question, we train diffusion models on the RPM visual reasoning dataset. We find that diffusion models are capable of generating novel samples conforming to relational rules without directly memorizing training data. Moreover, the models successfully generate samples that conform to rules of similar structure unseen in training, suggesting generalization in the abstract relation space. Notably, the models exhibit ordered learning dynamics in rule acquisition, with local data structure learned earlier than global structure.

Keywords: visual reasoning, generative model, diffusion model, generalization, abstract rule, inference

Introduction

Humans excel at recognizing relations between objects and generalizing abstract relations like ‘constant’ across various contexts - for example, constant shape or size. A key goal in machine learning is to give machines similar capabilities in relational reasoning. Recently, diffusion models have shown impressive ability to generate realistic images and capture complex data distributions (Rombach, Blattmann, Lorenz, Esser, & Ommer, 2022). Can these models also emulate human-like generalization of abstract relations?

Characterizing generalization in diffusion models is complex because the underlying data distribution they should capture is often unknown. Traditional evaluations of these models (e.g. FID) usually focus on image diversity and realism (Heusel, Ramsauer, Unterthiner, Nessler, & Hochreiter, 2017). However, an important application is to have the generated images conform to specific relational rules. This study investigates whether diffusion models can learn and generalize the abstract relations defining data distributions. We utilize the Raven’s Progressive Matrices (RPM) task, a well-established benchmark for measuring abstract reasoning skills (Raven, 1936). We train diffusion models on RPM images with various relational rules and assess their ability to generate new images that follow both trained and novel rules. Our findings suggest that diffusion models can generalize abstract visual relations, prompting further research into their reasoning capabilities in vision and beyond.

Approach

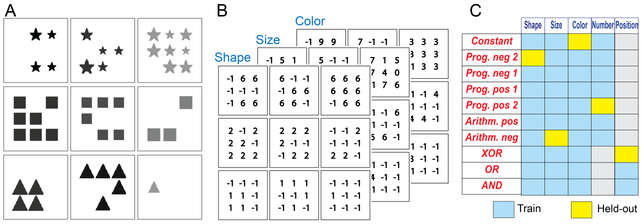

Figure 1: A) Example GenRAVEN dataset image and B) its attribute-value array representation. The underlying rule is constant shape. C) The 40 relational rules in the GenRAVEN dataset. 5 rules are held-out during training.

GenRAVEN Dataset We introduce the GenRAVEN dataset, comprising RPMs associated with 40 relational rules. Each RPM features a 3×3 matrix where each row follows a unique relational rule. Each rule is composed of an abstract relation (constant; progression ±2, ±1; arithmetic ±; XOR; OR; AND) applied to an attribute (shape, size, color, number, position). The dataset encodes each RPM by an integer array of $3 \times 9 \times 9$, indicating attribute values for each position in the panels, with -1 representing empty positions (Fig. 1B).

The dataset is designed so that the rule governing each row remains ambiguous when examining only the first two panels and only becomes evident when all three panels are considered. This design ensures that the rule governing the row cannot be directly deduced from the first two panels alone and the model must reason the entire matrix. We generate 4k random images per rule for training. To study the generalization of abstract relations such as constant to new attributes, we held out 5 rules (Fig 1C) during training and evaluate the model’s ability to generate images of these held-out rules.

Diffusion Model The Diffusion model has been the prominent approach for generative image modeling (Dhariwal & Nichol, 2021). Given the spatial nature of the task, we treat each RPM as a 9x9 image with 3 attribute channels and adapted existing diffusion methods for image generation. Specifically, we experimented with two network architectures, UNet (Karras, Aittala, Aila, & Laine, 2022) and Diffusion Transformer (DiT) (Peebles & Xie, 2023). We used deterministic samplers to generate samples: Heun’s 2nd order sampler with

[1] † Equal contribution.

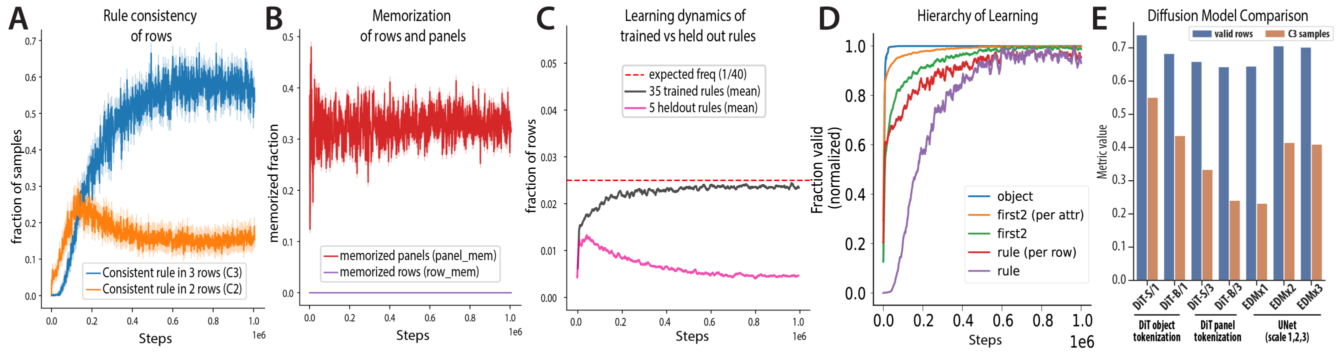

Figure 2: **Diffusion model learns to capture abstract relations**. **A.** Proportion of generated samples with consistent rules applying to three rows or two rows (only). **B.** Proportion of generated rows and panels that are memorized (found in the training dataset). **C.** Proportion of generated rows that conform to trained and held-out rules. The dashed line denotes the expected frequency $1/40$, if the model generates 40 rules uniformly. **D.** The fraction of generated samples that satisfy each criterion throughout training, normalized by their maximum values, highlighting their different learning rate. **E.** Model comparison regarding the validity of row and rule consistency across rows. **DiT-B** doubles the hidden size and attention head count of **DiT-S**. **EDM-x1,x2,x3**: UNet with 1,2,3 times width and depth.

18 steps for UNet model (Karras et al., 2022); and DDIM sampler with 100 steps for DiT (Song, Meng, & Ermon, 2020). We rounded the generated value of each attribute to the closest integer. Leveraging the in-painting capability of diffusion models (Lugmayr et al., 2022), we also challenged them with the RPM tasks: given 8 existing panels, let the diffusion model fill out the missing panel from noise.

For each row in a generated sample, we inferred the set of applicable rules. When no rule applies, the row is called invalid. We calculated the fraction of valid rows throughout training. We also calculated the fraction where the same rule applies to two or three rows in the sample, which we call the two or three-row consistent fraction (C2, C3).

Results

Diffusion model learns abstract relations without memorization Through diffusion training, the validity and rule consistency of the generated samples both increased (Fig. 2A). Using DiT-S/1 as our running example, after training, 55.1% of the samples had consistent rules applying to three rows (C3), and 73.9% of the rows are valid. This is substantially higher than the chance level (0.03% for C3; 16.9% for valid row) per uniform random sampling of attribute values. Importantly, among the 200k generated RPMs, no RPM or a single row exists in the training dataset, which suggests that the model has learned the underlying rules rather than memorizing specific examples (Fig. 2B). Nevertheless, 32% of the generated panels exist in the training set, showing that diffusion models could memorize local patterns of the training data, while creating novel rule-conforming combinations.

Further, the model extended to generate rows and samples that align with untrained, held-out rules more frequently than the chance would allow (Fig. 2C). This suggests its generalization on the next level: namely, by observing rows with ”constant size” and ”constant shape”, it learned to generate rows with ”constant color”, which provides evidence that the model captured the abstract relation of ‘constant’. Interestingly, the frequency of generating samples and rows following held-out rules exhibited non-monotonic dynamics, where initially the model learned to generate both trained and held-out rules similarly, however over time, the frequency of generating held-out rules decreased while that for trained rules increased. This suggests that early on the diffusion models tend to ”overgeneralize”, and later they learn to constrain the extent of generalization.

Hierarchy of learning dynamics across rule criteria To conform to relational rules, the generated samples must satisfy the following criteria: 1) **Object-consistency**: the attribute values at each location must either fail within predefined ranges or be marked by -1 if no object is present. 2) **First-two panel validity**: the values of each attribute in the first two panels must potentially comply with a rule. 3) **row-level rule validity**: across all three panels in a row, one attribute must satisfy a rule. 4) **Cross-row consistency**: The rules applied to the three rows within a matrix must be the same. Fig. 2D shows the fraction of generated samples that meet each of these criteria at different training epochs. We observe an order of learning, where the model initially learns more local data structures and subsequently adapts to more global ones.

Reasoning ability varies based on architecture. Regarding model architectures, transformer-based backbone (DiT) with object tokenization (DiT-S/1) achieved stronger rule consistency than the UNet across scales (EDM-x1 to EDM-x3); it also has better rule consistency than DiT with panel tokenization (DiT-S/3). This suggests the importance of object-level self-attention for rule inference. Scaling up the depth and width of UNet model improved performance, while scaling up the width of DiT model did not. The full comparison of rule validity and consistency across network architecture and scale is shown in Fig.2 E.

Diffusion models show RPM reasoning through impainting Finally, we tested the ability of diffusion models to complete RAVEN tasks through inpainting. We created novel row combinations for each rule and let the model fill-in the last panel. The panel generated by the model (DiT-S/1) was rule consistent at a rate of 28.7% and 8.2% for trained and held-out rules, surpassing the chance level per random sampling (2.2% and 3.5%). This performance highlights their ability to generate according to the abstract relations inferred from existing context and generalize to novel rules.

Conclusion and Future Direction

We demonstrate that diffusion models can effectively learn and generalize abstract relational rules, offering significant insights into their capabilities in complex reasoning tasks. The models not only accurately generate novel examples of trained rule but also extend abstract relation to generate examples of unseen rules. Additionally, the hierarchical learning dynamics we identified mark an initial step towards understanding how these models generalize. Looking forward, we aim to extend our findings to explore generalization behavior in more realistic reasoning tasks, particularly those based on pixel valued images rather than attribute values.

Acknowledgments

References

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, *34*, 8780–8794.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, *30*.

Karras, T., Aittala, M., Aila, T., & Laine, S. (2022). Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, *35*, 26565–26577.

Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., & Van Gool, L. (2022). Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 11461–11471).

Peebles, W., & Xie, S. (2023). Scalable diffusion models with transformers. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 4195–4205).

Raven, J. C. (1936). Mental tests used in genetic, the performance of related indiviuals on tests mainly educative and mainly reproductive. *MSC thesisUniv London*.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 10684–10695).

Song, J., Meng, C., & Ermon, S. (2020). Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.