Human-like compositional visual inference through neural diffusion on syntax trees

Sylvia Blackmore (sylvia.blackmore@yale.edu)
Psychology Department, Wu Tsai Institute, 100 College Street
New Haven, CT 06511 USA

Sarah Feng (sarah.feng@yale.edu)
Psychology Department, 100 College Street
New Haven, CT 06511 USA

Steve Chang (steve.chang@yale.edu)
Psychology Department, Wu Tsai Institute, 100 College Street
New Haven, CT 06511 USA

Ilker Yildirim (ilker.yildirim@yale.edu)
Psychology Department, Wu Tsai Institute, 100 College Street
New Haven, CT 06511 USA

Abstract

Compositionality—the ability to decompose experiences into constituent parts and flexibly recombine them—is fundamental to human intelligence. Despite the vast combinatorial space created by even basic elements, humans efficiently navigate potential configurations during visual inference. We present a neuro-symbolic approach framing visual compositional inference as inverse graphics through guided program synthesis, implemented as neural diffusion on syntax trees. Our model represents images as programs, using a conditional neural network and value model to enable efficient beam-search through program space. Validated against human behavioral data, the model achieved human-like performance across trial types. This framework provides a computational account of visual inference as search through compositional state space.

Keywords: inverse graphics; goal-conditioned world models; neural diffusion; neural replay

Introduction

Compositionality is fundamental to human intelligence (Frankland & Greene, 2020, Kurth-Nelson et al., 2023). The capacity to break down experiences into constituent parts and dynamically recombine them is present across cognitive domains. In the visual domain, the brain processes complex scenes by combining object representations with spatial rules. Despite the vast state space of possible combinations, humans rapidly learn rich visual representations and efficiently guide search through the compositional state space. We apply a neuro-symbolic approach to model compositional visual inference that explains both human structural capacity and search efficiency. Computationally, we treat visual inference as inverse graphics through program synthesis, implemented as guided neural diffusion on syntactically valid syntax trees (Kapur et al., 2024).

Task

In an MEG inference task, subjects were trained to construct shapes out of elementary building blocks, then presented with silhouettes constructed from these blocks and asked to infer the underlying blocks and their spatial relationships (Schwartenbeck et al., 2023) (Figure 1A). This task casts visual compositional inference as hypothesis testing. Our framework allows for a principled derivation of the dynamics of hypothesis testing through program synthesis in an inverse graphics task.

Computational Model

Our approach implements a neural diffusion model operating directly on syntax trees, enabling iterative program refinement while maintaining syntactic validity (Kapur et al., 2024) (Figure 1B). Visual images are represented as programs written in a Domain-Specific Language governed by a Context-Free grammar. Given an initial program z_0 (and corresponding image x_0, noise is added to the syntax tree through syntactically valid mutations. Denoising is cast as recovering the target program z corresponding to image observation x. A conditional neural network models the distribution of programs at the previous step, giving policy $q_\phi(z_{t-1}|z_t, x_t; x_0)$. An edit path between target and mutated trees trains a value model $\upsilon_\phi(X_A, X_B)$, to predict edit distance between images. The integration of policy and value model enables efficient search through program space, as only nodes with promising values are expanded.

Results

The model learned to perform the visual inference task at a human-like level After training, the model reached 77.08% ± 3.99% accuracy, approaching human performance levels of 82.38% ± 3.28% (Figure 2A). It recapitulated a qualitative performance difference across trial-types, where trials both humans and the model performed better on trials where they had to infer the relationship between blocks that were not

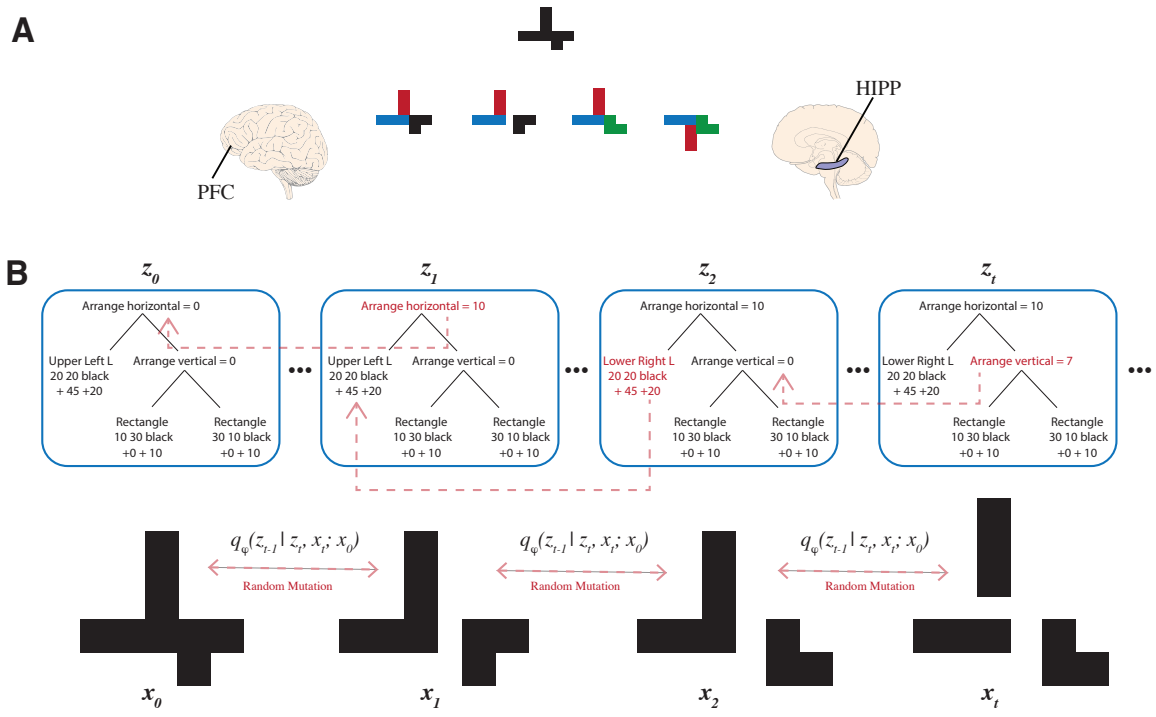

Figure 1: A. In MEG, humans performed a compositional visual inference task. Given a target silhouette, they had to reconstruct its underlying structure and composition. Neural replay in the hippocampal-prefrontal circuit was found to occur at the time of inference. B. Treated as an inverse graphics task, the neural diffusion model uses a conditional neural network conditioned on the current program z_t, current output x_t, and target output x_0 (here, the silhouette) to model the distribution of programs at the previous step: $q_\phi(z_{t-1}|z_t, x_t; x_0)$.

adjacent to each other in the silhouette (not-connected trials) than those where they were (connected trials). In addition, there was a strong positive correlation between human reaction times and the number of nodes expanded by the program ($r(46) = .77$, $p = .07$), suggesting similar computational approaches during search through compositional state space (Figure 2B).

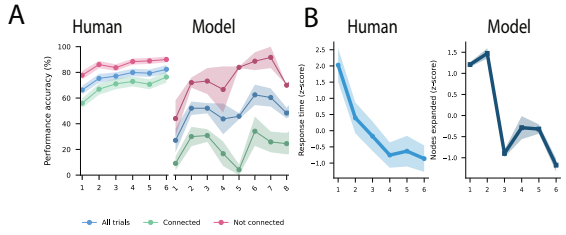

Figure 2: A. After training, the model approached human performance levels and recapitulated differences across trial-types. B. Correlation between human reaction times and the number of nodes expanded by the model $r(46) = .77$, $p = .07$).

Conclusions

Our neuro-symbolic approach, which implements visual compositional inference as inverse graphics through neural diffusion on syntax trees, offers a computational framework that explains both human structural capacity and search efficiency. The model achieved human-like performance and mirrored performance differences across trial types, with model search complexity correlating with human reaction times. This suggests humans may employ similar computational strategies when navigating compositional state spaces. While the neural mechanisms underlying compositional thought remain largely unexplored, our framework provides a computationally principled method to investigate these algorithms in the brain. By framing visual inference as a guided search through a compositional program space, we establish a foundation for future work connecting computational models with neural data, which we aim to do in ongoing investigation of MEG data.

References

Frankland, S. M., & Greene, J. D. (2020). Concepts and compositionality: in search of the brain's language of thought. *Annual review of psychology*, *71*(1), 273–303.

Kapur, S., Jenner, E., & Russell, S. (2024). Diffusion on syntax trees for program synthesis. *arXiv preprint arXiv:2405.20519*.

Kurth-Nelson, Z., Behrens, T., Wayne, G., Miller, K., Luettgau, L., Dolan, R., ... Schwartenbeck, P. (2023). Replay and compositional computation. *Neuron*, *111*(4), 454–469.

Schwartenbeck, P., Baram, A., Liu, Y., Mark, S., Muller, T., Dolan, R., ... Behrens, T. (2023). Generative replay underlies compositional inference in the hippocampal-prefrontal circuit. *Cell*, *186*(22), 4885–4897.