DEPTH VS RECURSION: OUTPERFORMING TRANSFORMERS IN JIGSAW RECONSTRUCTION

Timofey Brayko, Artemii Miasoedov*
Innopolis University
{t.brayko, a.miasoedov}@innopolis.university,

Rustam Lukmanov
Research Center of the Artificial Intelligence Institute
Innopolis University
r.lukmanov@innopolis.ru

ABSTRACT

Chain-of-Thought (CoT) has demonstrated that explicit reasoning steps enhance large language model performance, yet this typically requires computationally expensive token sequences. In this work, we investigate Tiny Recursive Models (TRM), which internalize reasoning via iterative refinement of a latent "thought" vector. We benchmark accuracy of TRM against standard encoder-only Transformer (EOT) on the task of Jigsaw Puzzle reconstruction, a domain requiring robust global spatial reasoning. While both architectures perform comparably on trivial grids up to 3×3, EOT performance collapses as complexity increases. In contrast, TRM maintains robust scaling with tight parameter budget. Furthermore, TRMs exhibit "abrupt learning" phase transitions during training, suggesting that latent recursion enables a qualitative leap in reasoning depth unattainable by simply stacking transformer layers. Code available at: https://github.com/Tima-Toma/depth-vs-recursion.

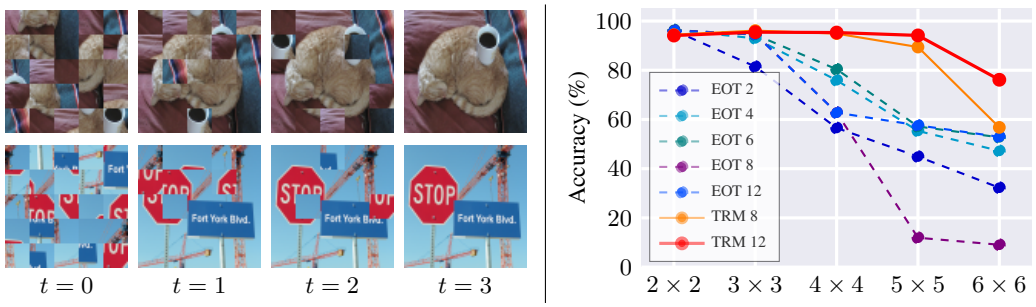


Figure 1: **TRM performance. Left:** Intermediate predictions of puzzle reconstruction across steps. **Right:** Systematic comparison showing the scalability gap against Transformer architectures.

1 INTRODUCTION

Chain-of-Thought (CoT) has established that explicit, step-by-step "thinking" can significantly improve the reasoning capabilities of large language models (LLMs) (Wei et al., 2022). Despite its success, CoT is often limited by its reliance on long, computationally expensive token sequences. In response, recent work proposes a recursive paradigm (Jolicoeur-Martineau, 2025; Wang et al., 2025) that internalizes the reasoning process within the model embedding space. Instead of outputting explicit text, these recursive models iteratively refine a candidate solution alongside a latent "thought" vector, effectively decoupling reasoning depth from model size. The **Tiny Recursion**

*All authors contributed equally.

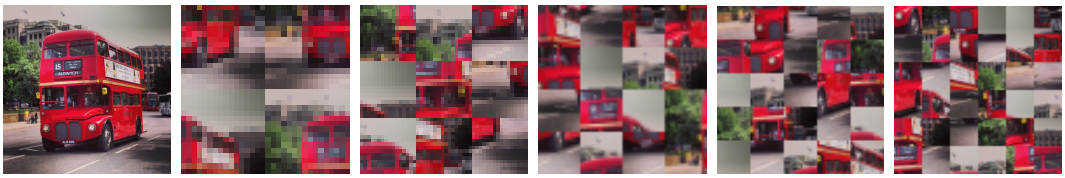
Model (TRM) (Jolicoeur-Martineau, 2025) exemplifies this approach, outperforming larger LLMs equipped with standard reasoning techniques.

While Sudoku is an effective benchmark, a robust evaluation requires testing recursive architectures across varied levels of complexity and spatial reasoning. Jigsaw puzzles are an ideal candidate, as they demand a combination of local feature matching and global spatial planning. We therefore benchmark TRM against a standard **encoder-only Transformer** (EOT) on the task of jigsaw puzzle reconstruction. By systematically evaluating performance across increasing puzzle grid sizes, we assess whether recursion in latent space offers a substantive advantage over traditional single-pass processing in complex visual reasoning tasks.

2 METHODOLOGY

2.1 TASK AND DATASET

Prior work on recursive models (Wang et al., 2025; Jolicoeur-Martineau, 2025) has primarily focused on LLMs solving complex logic games that remain challenging for state-of-the-art models, such as Sudoku and mazes (Seely et al., 2025). We argue that, while valuable, these benchmarks are abstract and purely symbolic. Instead, we focus on jigsaw puzzle reconstruction, a task that evaluates visually grounded reasoning. This task poses a dual challenge that remains non-trivial for current vision-language models (VLMs) (Lyu et al., 2025; Dirauf et al., 2026): successful reconstruction demands both **local visual perception** to match textures and **global reasoning** to organize the pieces into a coherent structure.



Initial 2×2 3×3 4×4 5×5 6×6

Figure 2: Shuffled jigsaw puzzles ranging from 2×2 to 6×6. Patch size is fixed to 16px.

Crucially, jigsaw puzzles support scalable data generation with controlled complexity, ranging from trivial 2×2 to complex 6×6 grids. This enables training on diverse real-world semantics and allows us to systematically probe the "breaking point" of both EOT and TRM architectures. Prior work has used MNIST and CIFAR datasets (Severo et al., 2025; Zhang et al., 2025), which are low-resolution and therefore ill-suited to large grids. To address this, we use the **COCO** dataset (Lin et al., 2014).

2.2 PERMUTATION FORMULATION

We define the puzzle reconstruction task as recovering the original spatial configuration of an unordered set of image patches $X = \{x_1, \ldots, x_N\}$. Let $\pi \in S_N$ denote a permutation such that $x_{\pi(i)}$ is the correct patch for the i-th position. A naive approach is to predict π by independently assigning an index in $\{1, \ldots, N\}$ to each position. However, this method allows N^N possible outputs, the vast majority of which are invalid since indices can be duplicated or omitted.

Algorithm 1 Fisher–Yates Shuffle

for $i = 1$ **to** $N - 1$ **do**
  $j \leftarrow$ uniform random integer in $\{0, \ldots, N - i\}$
  swap $\text{array}[i]$ and $\text{array}[i + j]$
end for

To constrain the model output strictly to the set of valid permutations, we adopt the **Fisher–Yates representation** (Fisher & Yates, 1938; Severo et al., 2025). Rather than predicting absolute positions directly, we predict a sequence of swap operations. Specifically, for each patch at index

$i \in \{1, \ldots, N-1\}$, the model predicts a relative index $j_i \in \{i, \ldots, N\}$. This value specifies which element from the remaining patches (to the right of the current position) should be swapped into position i, ensuring that the output is always a valid permutation by construction.

2.3 EXPERIMENTAL SETUP

To ensure that performance differences are attributable specifically to the recursive mechanism, both architectures share an identical bidirectional encoder-only Transformer backbone (Vaswani et al., 2017; Devlin et al., 2019). The backbone uses SwiGLU activations (Hendrycks & Gimpel, 2023; Shazeer, 2020), RMSNorm in a post-norm configuration (Zhang & Sennrich, 2019), and no bias terms. Patch encoding is handled by a lightweight CNN encoder that captures local visual features. The baseline is a bidirectional encoder-only Transformer that processes the sequence of patch embeddings in a single forward pass. It uses full self-attention to predict the Fisher–Yates permutation directly from the input.

TRM uses the same backbone but processes the input iteratively (Jolicoeur-Martineau, 2025). At each step t, the network receives the initial patch embeddings, the current latent solution $\mathbf{y}^{(t-1)}$, and a latent "thought" vector $\mathbf{z}^{(t-1)}$. The TRM processes the input through a "deep supervision" loop consisting of N_{sup} macro-steps. Within each macro-step t, the model iterates n times for "thinking time" before updating the current solution. After these reasoning steps, the model updates $\mathbf{y}$.

$$\mathbf{z}_0^{(0)} \sim \mathcal{N}(0, 1) \tag{1}$$

$$\mathbf{y}^{(0)} \sim \mathcal{N}(0, 1) \tag{2}$$

$$\mathbf{z}_n^{(t)} = f(\mathbf{y}^{(t-1)} + \mathbf{z}_{n-1}^{(t)} + \mathbf{X}) \tag{3}$$

$$\mathbf{y}^{(t)} = f(\mathbf{y}^{(t-1)} + \mathbf{z}_N^{(t)}) \tag{4}$$

3 RESULTS

While both architectures achieve comparable accuracy on simple 2×2 puzzles ($\sim$95%), TRM significantly outperforms the vanilla EOT as combinatorial complexity increases. On 5×5 puzzles, EOT accuracy degrades to 57.61%, whereas TRM maintains robust scaling, achieving 94.15% accuracy. Notably, TRM achieves this with the same $\sim$0.7M parameter budget as the baseline, although it requires approximately 20$\times$ more computation.

Model	Params (M)	FLOPs (M)
EOT 2	0.7	74.9
EOT 4	1.2	84.4
EOT 6	1.8	93.9
EOT 8	2.3	103.4
EOT 12	3.3	122.5
TRM 8	0.7	979.6
TRM 12	0.7	1469.5

Table 1: **Complexity.** Model size and compute for 3×3 puzzle. Parameters and FLOPs measured in millions.

Model	Grid Size ($n \times n$)				
	2×2	3×3	4×4	5×5	6×6
EOT 2	96.18	81.49	56.54	44.92	32.27
EOT 4	**96.38**	92.95	75.93	55.34	47.25
EOT 6	96.34	94.45	80.40	57.14	52.74
EOT 8	96.08	94.60	62.76	57.61	52.95
EOT 12	96.14	94.74	62.82	11.94	9.09
TRM 8	94.46	**96.11**	94.88	89.43	56.78
TRM 12	94.10	95.47	**95.32**	**94.15**	**76.14**

Table 2: **Patch level accuracy.** Performance across grid sizes. Suffix denotes depth of model: EOT — number of layers, TRM — number of macro-steps.

Attempting to match the effective depth of TRM by increasing the number of EOT layers proved ineffective. Performance peaked at 6 layers, and further deepening the model to 12 layers led to training failure on complex grids (5×5 and 6×6). This confirms that recursive state refinement cannot be approximated simply by stacking layers (Jolicoeur-Martineau, 2025).

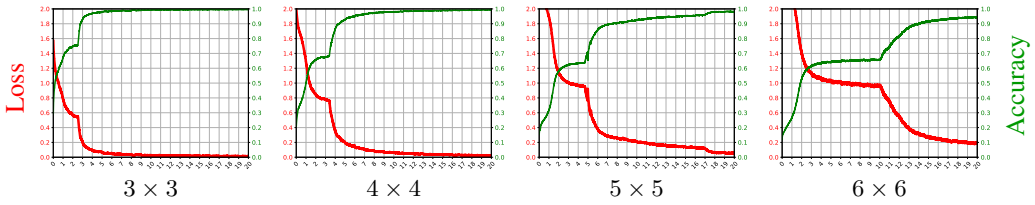
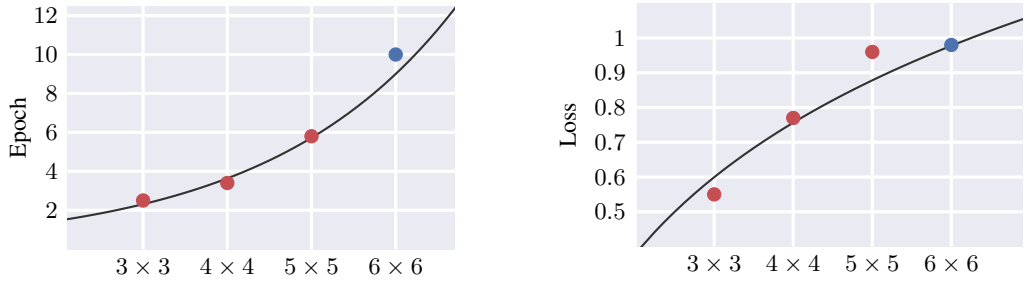


Figure 3: TRM training across increasing grid dimensions (3×3 to 6×6), with epochs along the X axis. Phase of abrupt learning is delayed as puzzle complexity increases.

Unlike standard Transformers, which settle into stagnant plateaus, TRM exhibits a distinct phenomenon of "abrupt learning" (Gopalani & Hu, 2025; Power et al., 2022; Naidu et al., 2025). After an initial stall, the model undergoes a delayed phase transition—a sudden drop in loss that signals the emergence of logical generalization. This breakthrough is tied to recurrence depth and task complexity: while the 8-step model faltered on complex 6×6 grids, increasing the depth to 12 reignited the phase transition, driving accuracy up to 76.14%.



(a) Exponential Growth ($y \sim ab^x$)

(b) Logarithmic Scaling ($y \sim a \log_b(x)$)

Figure 4: (a) Epoch and (b) loss of the "abrupt learning" phase transition across different grid sizes.

The moment of "abrupt learning" was delayed as task complexity increased: larger puzzles exhibited phase transitions at later epochs. To quantify this relationship, we plotted the puzzle size against both the epoch at which the phase transition occurred and the loss value at that moment. The timing of the phase transition is well approximated by an exponential function $y \sim ab^x$ with $R^2 = 0.98$, while the loss follows a logarithmic relationship $y \sim a \log_b(x)$ with $R^2 = 0.91$. Both curves were fitted using only the first three data points, with the fourth reserved for evaluation. Remarkably, the functions share the same base parameter $b = 1.57$, suggesting a unified scaling mechanism underlying the phase transition dynamics.

4 Conclusion

In this work, we demonstrate the superiority of TRMs over classical EOT on a complex visual reasoning task. By comparing the two architectures on high-dimensional jigsaw reconstruction, we show that latent recursion enables TRMs to solve tasks where size-matched static models fail to converge.

This performance gain comes with a computational trade-off. While TRM maintains the same compact parameter footprint (∼0.7M) as the baseline, it requires substantially more inference-time computation to iteratively refine its latent "thought" vector. However, since simply increasing the depth of a static Transformer yields diminishing returns, we argue that this trade-off is justified. TRM does not merely improve metrics; it enables a qualitative leap in reasoning capability that static architectures cannot replicate at this scale.

We hope that our experimental setup will serve as a valuable testbed for the community. This framework provides a rigorous environment to further explore the limits of recursive self-improvement without the prohibitive resource costs of massive language models.

ACKNOWLEDGEMENT

The study was supported by the Ministry of Economic Development of the Russian Federation (agreement No. 139-10-2025-034 dd. 19.06.2025, IGK 000000C313925P4D0002).

REFERENCES

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Richard Dirauf, Florian Wolz, Dario Zanca, and Björn Eskofier. Benchmarking content-based puzzle solvers on corrupted jigsaw puzzles. In Emanuele Rodolà, Fabio Galasso, and Iacopo Masi (eds.), *Image Analysis and Processing – ICIAP 2025*, pp. 286–298. Springer Nature Switzerland, 2026. ISBN 978-3-032-10185-3.

Ronald Fisher and Frank Yates. Statistical tables for biological, agricultural and medical research. *Science*, 88(2295):596–597, 1938. doi: 10.1126/science.88.2295.596. URL https://www.science.org/doi/abs/10.1126/science.88.2295.596.

Pulkit Gopalani and Wei Hu. What happens during the loss plateau? Understanding abrupt learning in transformers, 2025. URL https://arxiv.org/abs/2506.13688.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs), 2023. URL https://arxiv.org/abs/1606.08415.

Alexia Jolicoeur-Martineau. Less is more: Recursive reasoning with tiny networks, 2025. URL https://arxiv.org/abs/2510.04871.

Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Lecture Notes in Computer Science, Springer*, volume 8693 LNCS of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48. 13th European Conference on Computer Vision, ECCV 2014 ; Conference date: 06-09-2014 Through 12-09-2014.

Zesen Lyu, Dandan Zhang, Wei Ye, Fangdi Li, Zhihang Jiang, and Yao Yang. Jigsaw-puzzles: From seeing to understanding to reasoning in vision-language models, 2025. URL https://arxiv.org/abs/2505.20728.

Prudhviraj Naidu, Zixian Wang, Leon Bergen, and Ramamohan Paturi. Quiet feature learning in algorithmic tasks. In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025. URL https://openreview.net/forum?id=Rkl2ESuYCd.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022. URL https://arxiv.org/abs/2201.02177.

Jeffrey Seely, Yuki Imajuku, Tianyu Zhao, Edoardo Cetin, and Llion Jones. Sudoku-bench: Evaluating creative reasoning with sudoku variants, 2025. URL https://arxiv.org/abs/2505.16135.

Daniel Severo, Brian Karrer, and Niklas Nolte. Learning distributions over permutations and rankings with factorized representations, 2025. URL https://arxiv.org/abs/2505.24664.

Noam Shazeer. GLU variants improve transformer, 2020. URL https://arxiv.org/abs/2002.05202.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://papers.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Guan Wang, Jin Li, Yuhao Sun, Xing Chen, Changling Liu, Yue Wu, Meng Lu, Sen Song, and Yasin Abbasi Yadkori. Hierarchical reasoning model, 2025. URL https://arxiv.org/abs/2506.21734.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1e8a19426224ca89e83cef47f1e7f53b-Abstract.html.

Yongxing Zhang, Donglin Yang, and Renjie Liao. Symmetricdiffusers: Learning discrete diffusion on finite symmetric groups. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 5989–6014, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1199e1805888fabf6edff851ab86c057-Abstract-Conference.html.