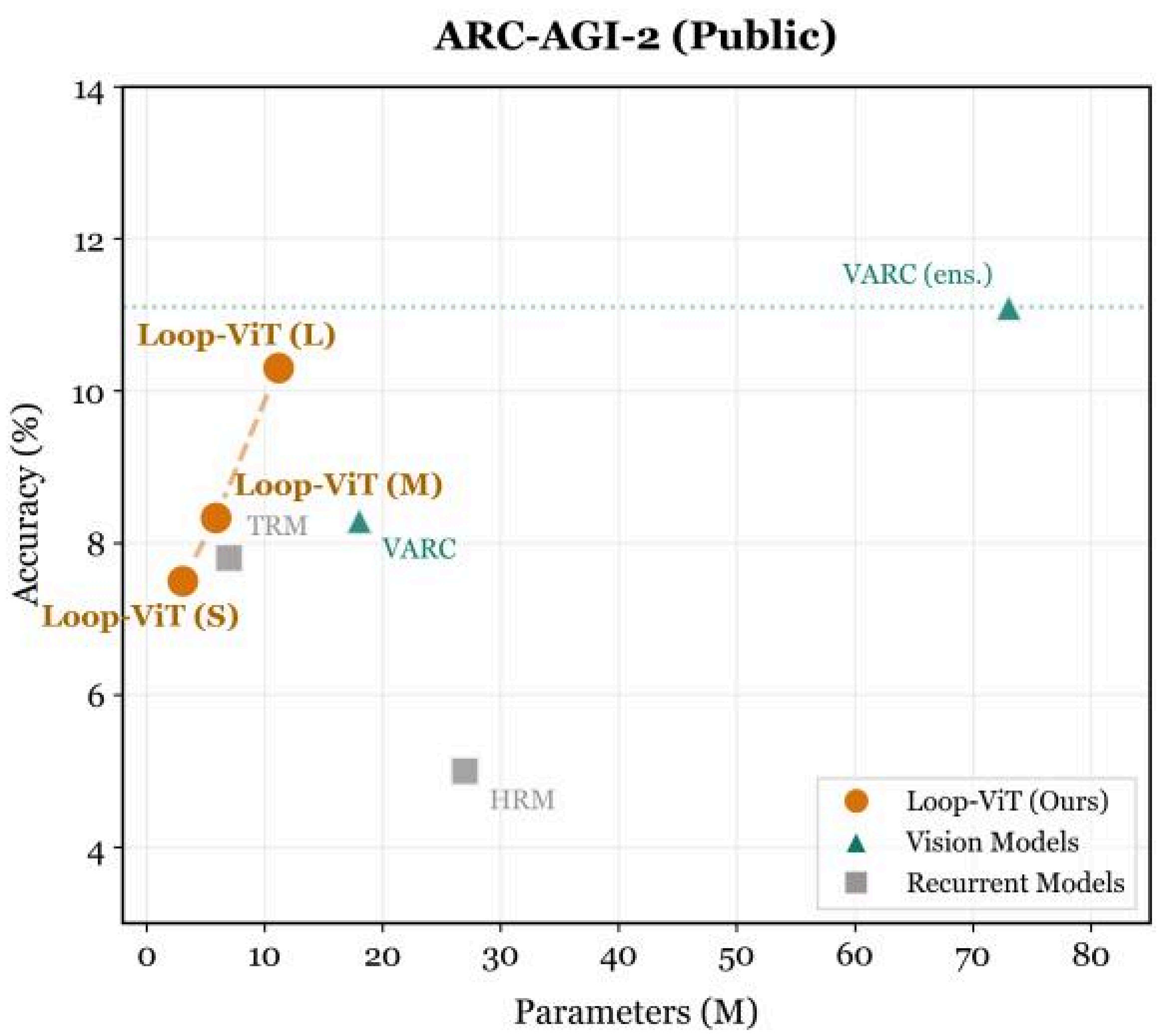
Thinking in Loops: Scaling Visual ARC with Looped Transformers

Wen-Jie Shu, Xuerui Qiu, Rui-Jie Zhu, Harold Haodong Chen, Yexin Liu, Harry Yang

HKUST, Bitdeer AI, CASIA

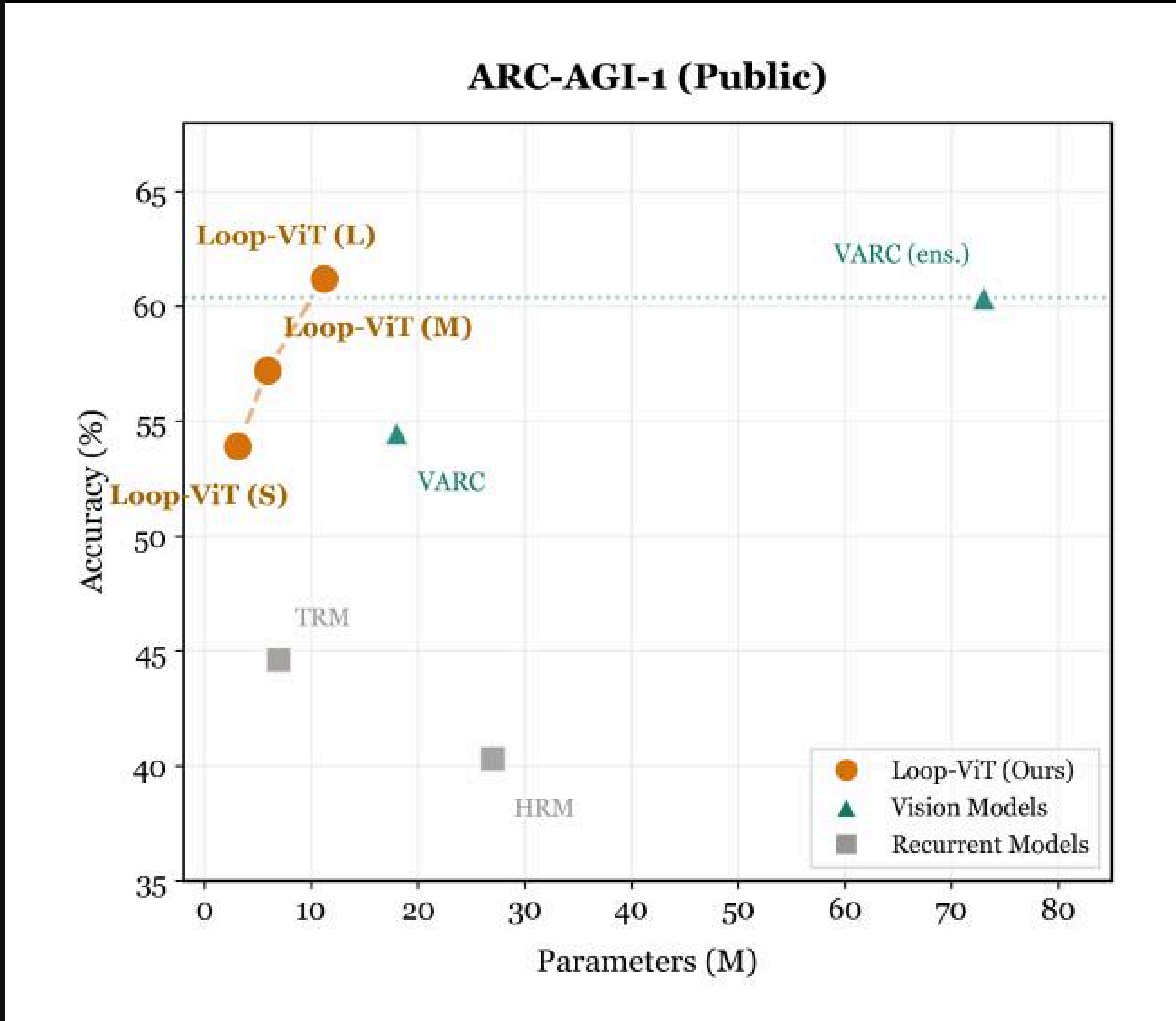
The Need for Iterative Reasoning

- Standard vision models use a feed-forward, 'one-shot' approach, mimicking fast but error-prone 'System 1' thinking.
- Inspired by NLP, Looped Transformers introduce iterative refinement, or 'Latent Chain-of-Thought', to vision.
- This allows a model to hypothesize, check, and correct itself within a fixed parameter budget.



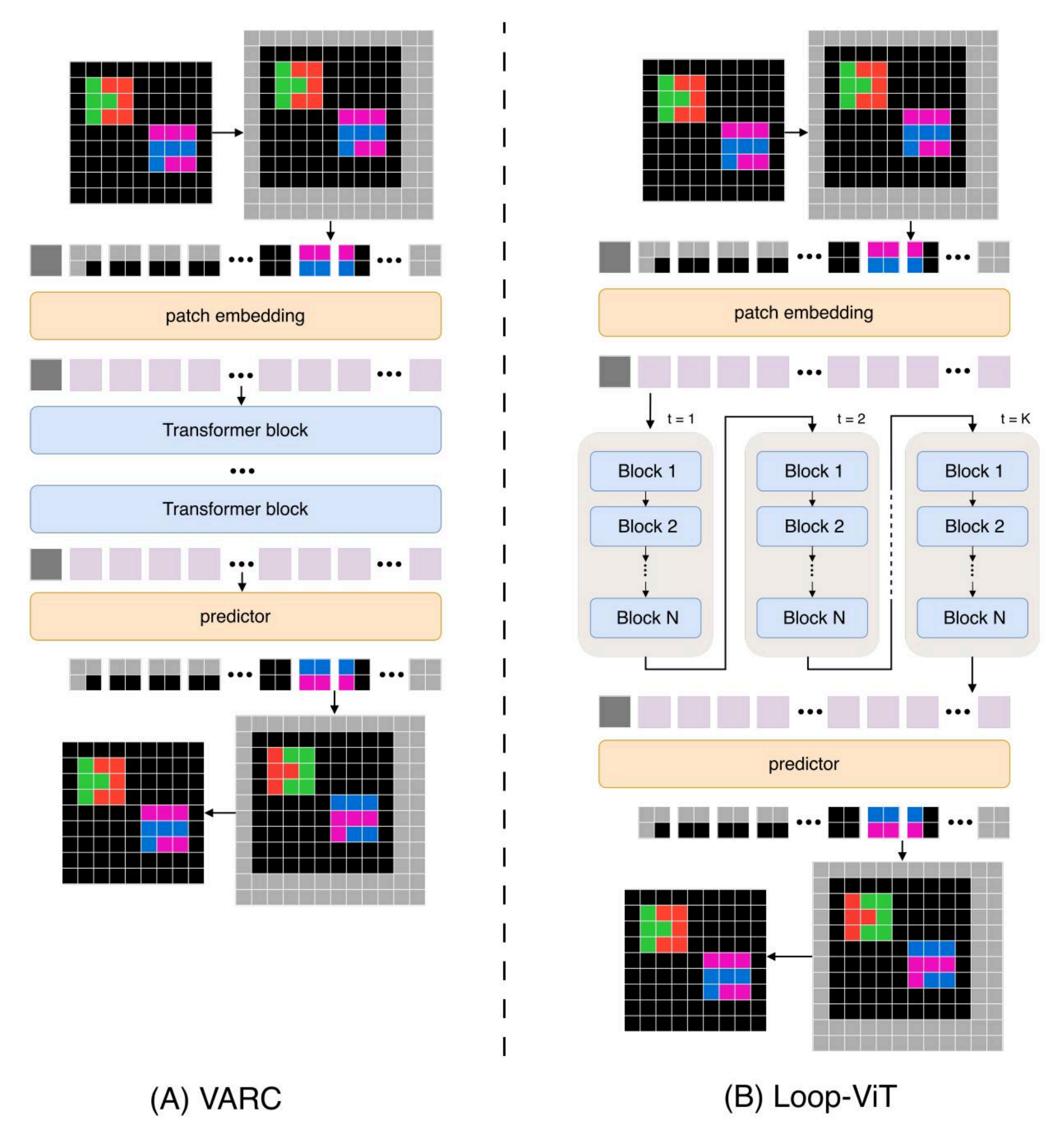
A New Efficiency Frontier for Visual Reasoning

- Loop-ViT establishes a new Pareto frontier for model parameters vs. accuracy on ARC-AGI.
- **Efficiency Victory:** Our 5.9M parameter model outperforms a 18M baseline (57.2% vs. 54.5%).
- **Scaling Compute:** A single 11.2M looped model surpasses a complex 73M parameter ensemble (61.2% vs. 60.4%).



Methodology: A Looped Vision Transformer

- We build on the VARC framework, which treats ARC as an image-to-image task on a unified 'canvas'.
- We replace the standard feed-forward backbone with a weight-tied (looped) transformer core.
- This core is executed for a fixed number of iterations (K), allowing the model to refine its prediction over time.



The Looped Inference Process

01 **Initialization**

Embed the input canvas into an initial hidden state h_0.

02 **Iterative Refinement**

For steps $t = 1, ..., K$, repeatedly apply the same transformer core F_θ to update the hidden state.

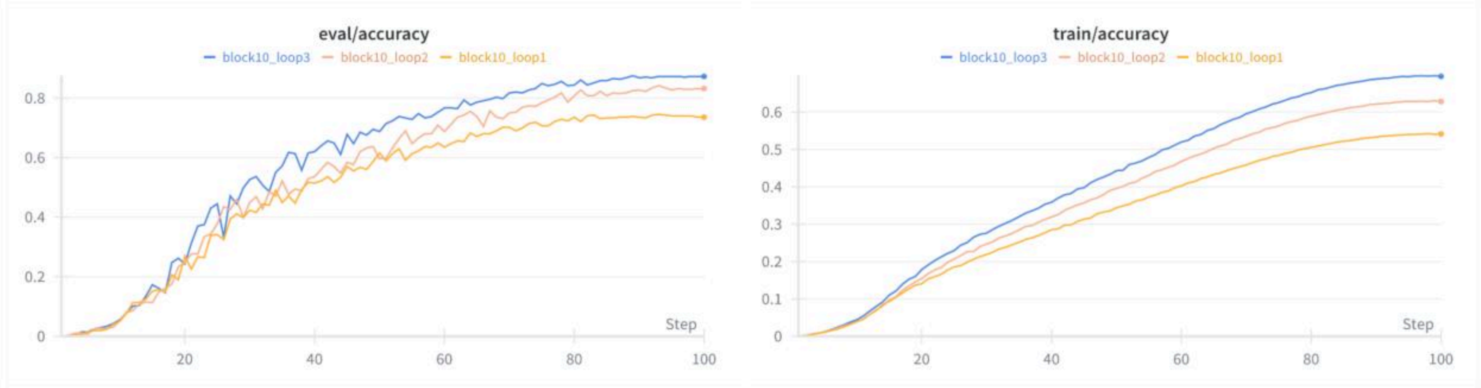
03 **Final Prediction**

Decode the hidden state at each step for an intermediate prediction. The final answer is the prediction from the last step, p_K.

$$h_t = F_\theta(h_{t-1}) \quad ; \quad p_t = D(h_t)$$

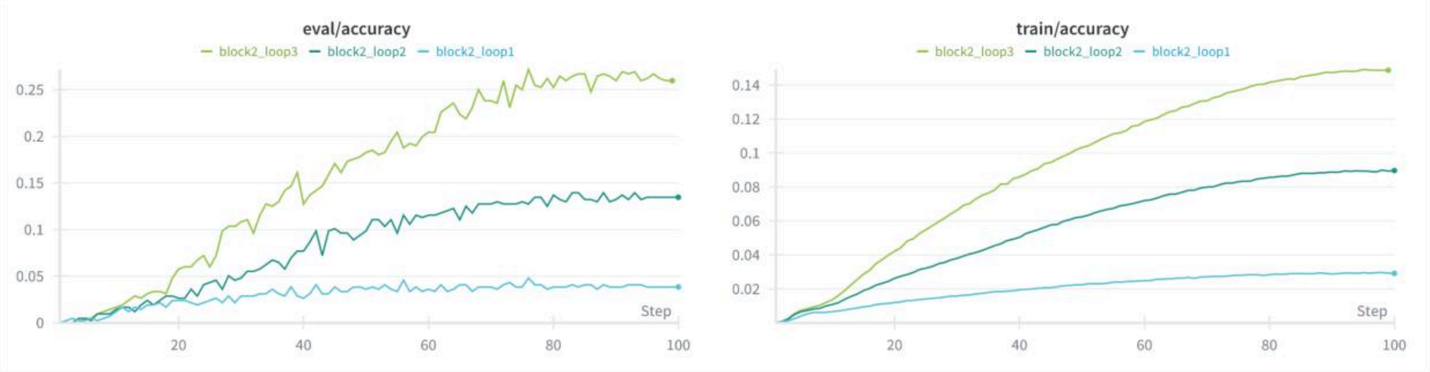
Training Dynamics: More Loops, Better Generalization

- Analysis of offline training reveals a key benefit of looping.
- Increasing the number of loop steps (K) not only improves evaluation accuracy (solid lines) but also reduces the gap between training (dashed lines) and evaluation performance.
- This suggests that iterative computation provides a powerful inductive bias for generalization.



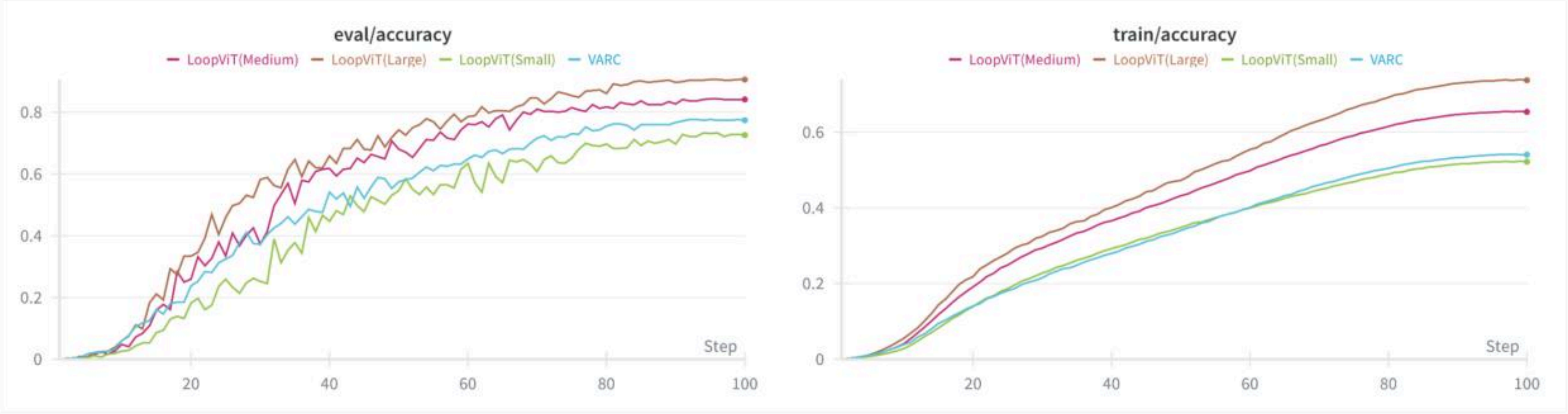
Training Dynamics: More Loops, Better Performance

- Experiments show that increasing loop iterations (K) from 1 to 3 consistently improves accuracy.
- For a fixed core depth (e.g., B=4), performance jumps from 41.0% (K=1) to 57.4% (K=3).
- This demonstrates that 'thinking time' is a more efficient scaling axis than just model width.



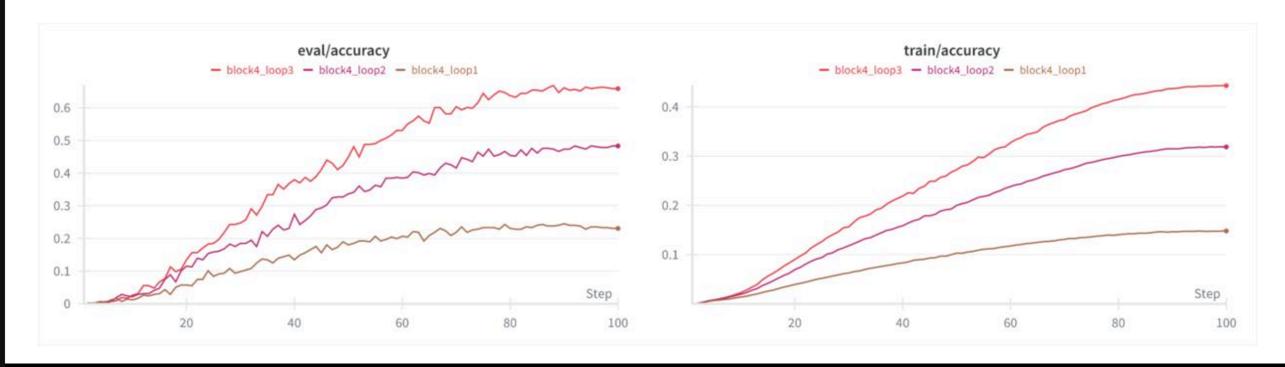
Main Results on ARC-AGI

- Loop-ViT significantly outperforms previous vision-based and recurrent models.
- A single 11.2M Loop-ViT surpasses the 73M VARC ensemble, demonstrating the power of iterative computation.



Key Findings: Efficiency via Recurrence

- **Efficiency:** A 5.9M Loop-ViT (K=6) surpasses the 18M VARC baseline.
- **Scaling:** A single 11.2M Loop-ViT (K=6) outperforms the 73M VARC Ensemble.
- **Consistency:** Gains hold on both ARC-1 and ARC-2 benchmarks, using the same model checkpoint.



Conclusion

- **Computational time** is a potent, underutilized resource in visual reasoning.
- **Progressive refinement** via looping yields better generalization and parameter efficiency than simply stacking more layers.
- Future work can explore adaptive computation and apply looped transformers to other visual tasks like inpainting and video understanding.