Visual Abstract Reasoning as a Step Towards Artificial General Intelligence

Amresh Kumar Singh

Department of Computer Science, Maharaja Ganga Singh University, Bikaner, India

Email: aksingh@mgsubikaner.ac.in

Abstract

Artificial General Intelligence (AGI) aims to develop systems that can learn new tasks efficiently, reason over structure, and generalize beyond the training distribution. Visual abstract reasoning is a strong candidate capability for AGI because it requires inferring latent rules from limited examples and applying them to novel inputs. Classic Raven-style matrix completion problems, along with modern benchmarks such as RAVEN and PGM, expose the gap between pattern recognition and rule-based reasoning by requiring relational, structural, and compositional inference. In parallel, the Abstraction and Reasoning Corpus (ARC) targets skill acquisition efficiency and emphasizes broad generalization with minimal supervision. This paper presents a practical view of visual abstract reasoning as an AGI-relevant subproblem. We synthesize capabilities required for AGI-oriented visual reasoning, review benchmark datasets and representative model families, present a reasoning block architecture that bridges perception and rule inference, and summarize evidence benchmark RAVEN and PGM results to highlight typical generalization gaps. We conclude with directions for improving systematic generalization via object-centric representations, compositional modules, and hybrid neural symbolic learning.

Keywords: Artificial General Intelligence, visual abstract reasoning, Raven Progressive Matrices, compositional generalization, computer vision

1 Introduction:

Deep learning has transformed perception, enabling high accuracy in tasks such as image classification, detection, and language modeling. Yet strong performance on perception benchmarks does not necessarily imply reasoning capability. Many models succeed by exploiting superficial correlations and fail when the test distribution shifts in a structured way. This limitation becomes apparent in abstract reasoning benchmarks, where the objective is to infer latent rules governing visual patterns rather than to match texture statistics.

AGI is often discussed as the ability to acquire skills efficiently across broad task scopes. Chollet argues that intelligence should be evaluated primarily through skill acquisition efficiency rather than raw task performance, and proposes ARC as a benchmark aligned with this view [2]. Visual abstract reasoning benchmarks align with this objective because they require learning latent rules that vary across instances and applying them to new cases.

Raven Progressive Matrices are a long-standing psychological test of fluid reasoning. Modern machine learning research has created large scale RPM like benchmarks such as RAVEN [1] and PGM [3], enabling systematic evaluation of reasoning architectures. RAVEN connects perception and reasoning using hierarchical structure annotations and multiple figure configurations [1]. PGM explicitly tests generalization regimes, including held-out rule combinations and extrapolation settings that make shortcut learning visible [3].

This paper asks how visual abstract reasoning supports progress toward AGI and what ingredients are missing in current approaches. We provide a capability-centric framing, a consolidated review of benchmarks and model families used in top venues, a practical reasoning block architecture, and a summary of benchmark results from the literature.

2 Related Work:

Recent progress in visual abstract reasoning has been supported by two parallel developments. First, several benchmarks have been designed to isolate reasoning and generalization, moving beyond recognition-centric evaluation. Second, a range of model families have been proposed to inject relational structure, compositionality, and interpretability into neural systems. In the following, we briefly summarize key datasets and representative architectures that motivate our AGI-oriented perspective.

2.1 Benchmarks and datasets for reasoning

RAVEN is a widely used RPM style dataset designed for relational and analogical visual reasoning [1]. PGM was introduced to probe abstract reasoning and systematic generalization in neural networks and includes controlled generalization regimes that separate memorization from rule learning [3]. ARC focuses on skill acquisition efficiency across tasks and emphasizes learning from few examples [2]. Diagnostic datasets such as CLEVR evaluate compositional visual reasoning with reduced dataset bias and detailed program supervision [13], and CLEVRER extends this idea to temporal and causal reasoning [15]. VCR targets higher level visual commonsense reasoning from images [20].

General reasoning datasets beyond RPM benchmarks include GQA for real world visual reasoning based on scene graphs and functional programs [21], NLVR2 for grounded reasoning about natural language and real photographs [22], and VQA v2 which reduces language priors by pairing questions with complementary images [23]. For temporal

reasoning in controllable video settings, CATER provides diagnostic sequences that test composition of object motion and interaction [24]. Beyond vision, bAbI provides prerequisite toy tasks for multi step reasoning diagnosis [25], and CLUTRR is a diagnostic benchmark for inductive relational reasoning from text with systematic generalization splits [26].

2.2 Model families for abstract visual reasoning

Relational Networks introduced a simple module that reasons over pairs of object features by aggregating pairwise interactions [4]. Graph Networks provide a general framework for relational inductive biases, representing entities as nodes and relations as edges [12]. In RPM benchmarks, CoPINet uses perceptual contrasting to compare candidates and context panels [7], LEN targets abstract reasoning under distracting features to learn robust rule related representations [8], and stratified rule aware networks attempt to model rule structure more explicitly [9]. Multi layer relation networks deepen relational processing and improve RPM performance [10].

Transformers are central across vision and language. The Transformer architecture formalized attention for sequence modeling [5], and ViT adapted Transformers to images using patch embeddings [6]. Transformers can model global interactions but do not guarantee rule discovery. Object centric learning with Slot Attention learns a set of slots that bind to objects [14]. Modular computation approaches such as Neural Module Networks [16] and Neural Programmer Interpreters [17] emphasize compositional structure. Broader perspective papers discuss remaining gaps in deep learning for AI and emphasize the need for structured representations and reasoning mechanisms [18,19].

3 Visual Reasoning Tasks

We present two representative examples of visual reasoning tasks to clarify the problem setting and the kind of inference required. First, a Raven style matrix completion problem is shown in Figure 1, where a 3 by 3 matrix contains one missing cell and the solver must infer the underlying rule from the relationships among the visible panels, then select the correct completion from the answer options. This typically requires identifying structured patterns such as attribute changes, spatial transformations, object relations, or logical combinations across rows and columns, rather than relying on local visual similarity. Second, an ARC grid transformation problem is shown in Figure 2, where the task is to infer a transformation rule from a small number of input output examples and apply the same rule to a new test input grid to produce the correct output. ARC problems often demand abstraction and compositional reasoning because the rule may involve grouping, symmetry, object extraction, color mapping, or multi step operations, and the system must generalize from very limited supervision. Together, these two examples highlight how

visual reasoning benchmarks go beyond standard recognition by requiring models to discover latent rules and apply them consistently to unseen instances.

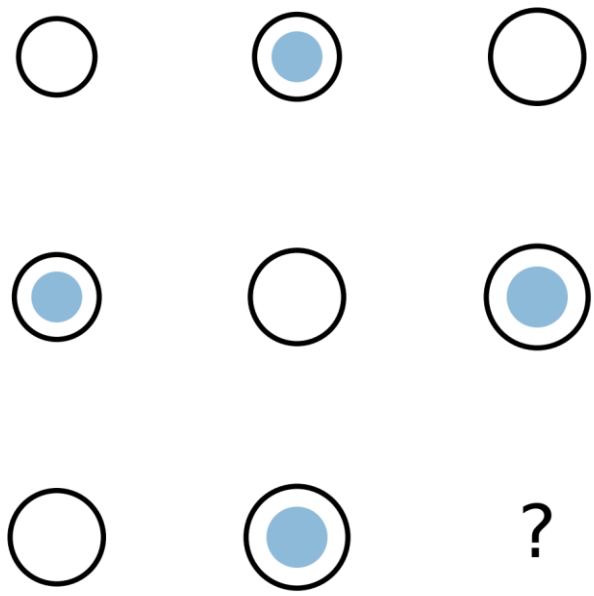

Figure 1: A typical 3×3 Raven's Progressive Matrices problem in which the final cell is missing. To solve it, the model or participant must discover the visual relationship governing the rows and columns, then choose the option that best completes the matrix

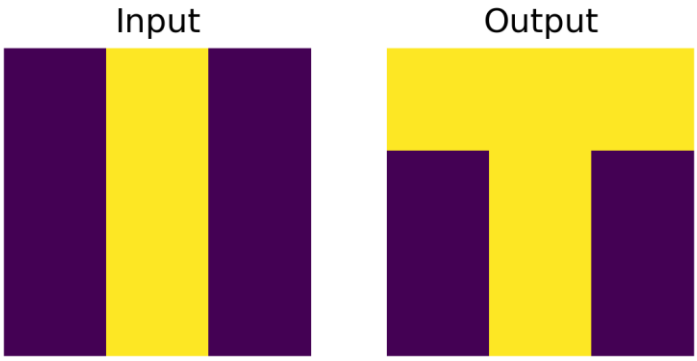

Figure 2: It illustrates a typical ARC task, consisting of example input-output grid pairs that demonstrate an underlying transformation pattern. The solver must infer this hidden rule from the examples and correctly apply it to produce the output for a new test input.

4 Methodology and Reasoning Block Architecture

RAVEN and PGM style problems can be framed as multiple choice classification. Given context panels that form an incomplete matrix and a set of candidate answers, the model selects the candidate that best satisfies the latent rule. ARC problems are closer to program induction: given a few input output examples, infer a transformation function and apply it to a new input [2].

4.1 Visual reasoning pipeline

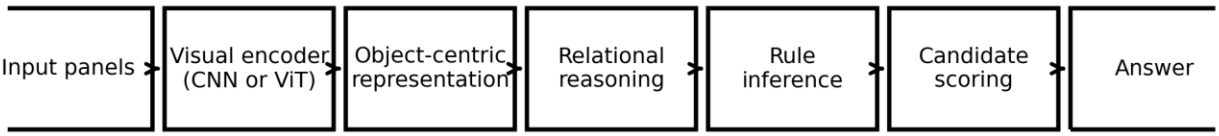

Figure 3: It presents a high-level visual reasoning pipeline in which the problem is solved through a sequence of distinct stages, from visual understanding to final decision-making. It separates low-level perception, relational analysis, rule inference, and candidate scoring so that the system can progressively interpret the matrix structure and select the most suitable answer.

A practical design principle is to separate stages that are often entangled in end to end models. First, a shared panel encoder extracts features. Second, an optional object centric stage forms object level slots or nodes. Third, a relational module computes interactions among objects and panels. Fourth, a rule inference module produces a compact rule representation from context. Finally, a candidate scoring module evaluates each answer option against the inferred rule.

4.2 Reasoning skills taxonomy

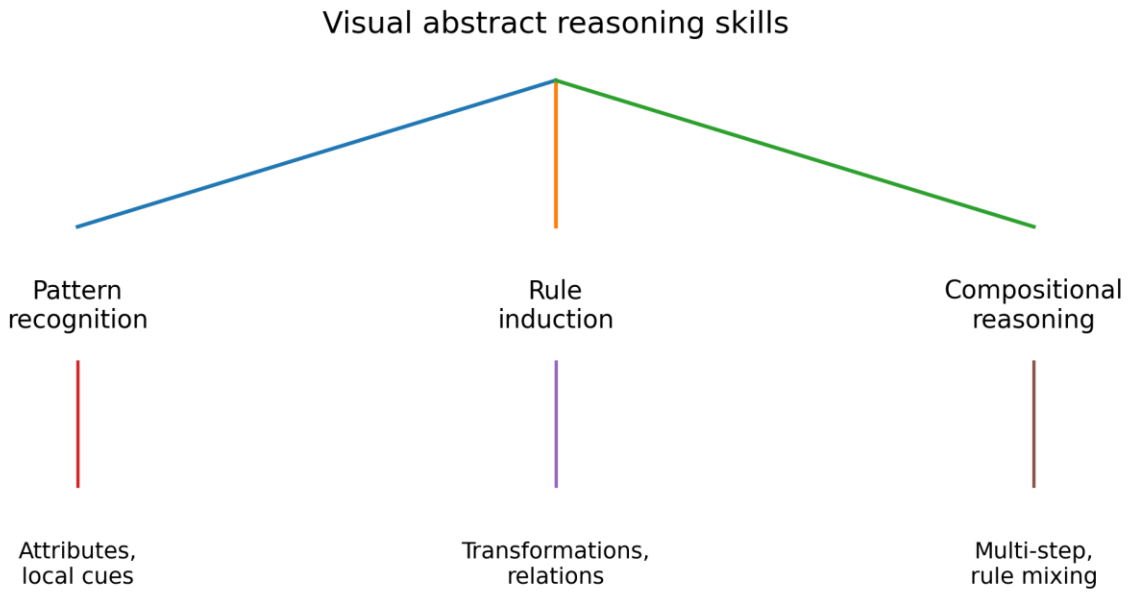

Figure 4: A capability-oriented view treats visual abstract reasoning as a set of separate skills that a model must learn, not a single score to optimize. Pattern recognition identifies visual regularities, rule induction discovers the underlying transformations, and compositional reasoning combines multiple rules to solve harder, novel cases, which is essential for AGI-oriented evaluation.

5 Benchmark findings

Now we summarize published benchmark findings that illustrate typical behaviors of current models. We use reported accuracy values from the literature on RAVEN and PGM, and interpret them through the lens of systematic generalization.

5.1 RAVEN results reported in CoPINet

CoPINet reports test accuracy on RAVEN for multiple baselines and configurations. In their Table 1, CoPINet achieves 91.42 percent mean accuracy on RAVEN, outperforming ResNet at 53.43 percent and ResNet plus DRT at 59.56 percent. The permutation invariant WReN variant reported there performs substantially lower at 17.62 percent mean accuracy. These results show that architectures with stronger inference structure can dramatically improve performance on the standard RAVEN split, but they do not by themselves guarantee broad generalization to new rule compositions [7].

Method	RAVEN mean test accuracy
WReN-NoTag-Aux	17.62%
CNN	36.97%

ResNet	53.43%
ResNet+DRT	59.56%
CoPINet	91.42%

Table 1: Reported mean test accuracy on RAVEN from CoPINet (NeurIPS 2019) [7].

5.2 PGM results and generalization regimes

PGM was designed to test generalization regimes explicitly. Barrett et al. report that a context blind ResNet performs near chance, while WReN achieves 62.6 percent test accuracy on the neutral split with distractors, and improves when distractors are removed [3]. CoPINet also reports overall PGM test accuracies for permutation invariant models: CoPINet achieves 56.37 percent mean test accuracy, compared with 49.10 percent for a WReN variant and 42.00 percent for ResNet [7]. Importantly, the PGM paper emphasizes that performance varies strongly across generalization regimes, and that predicting symbolic explanations improves generalization [3].

Source and setting	Reported result
PGM (neutral split, with distractors): WReN test accuracy [3]	62.6%
PGM (neutral split, distractors removed): WReN test accuracy [3]	78.3%
PGM (mean test acc, CoPINet Table 4): CoPINet [7]	56.37%
PGM (mean test acc, CoPINet Table 4): WReN-NoTag-Aux [7]	49.10%
PGM (mean test acc, CoPINet Table 4): ResNet [7]	42.00%

Table 2: Selected reported results on PGM from Barrett et al. (ICML 2018) [3] and CoPINet (NeurIPS 2019) [7].

Together, these findings motivate the need to measure and improve systematic generalization, not only IID accuracy. The literature consistently shows that adding relational inductive bias, contrasting mechanisms, and auxiliary supervision can improve performance, but robustness under controlled distribution shifts remains challenging [3,7,8,9].

6 Discussion

Visual abstract reasoning matters for AGI because it pressures models to infer latent structure and to generalize beyond training distributions. ARC emphasizes skill acquisition efficiency and broad generalization [2]. RAVEN and PGM provide controlled settings where compositional and out of distribution generalization can be tested explicitly [1,3].

A recurring obstacle is representation. When scenes are compressed into entangled vectors, object identity and relations can be difficult to recover. Object centric representations address this by separating scenes into object level slots or nodes, which can then be used by relational modules to infer rules [14]. Graph based reasoning provides a complementary route by explicitly encoding entities and relations and introducing relational inductive bias [12].

Another obstacle is compositionality. Humans solve RPM by composing primitives such as symmetry detection, counting, rotation, and logical operators. Neural systems often struggle with such composition, motivating modular computation and program like inference [16,17]. In practice, achieving both flexibility and interpretability may require hybrid neural symbolic methods that can represent rules explicitly while remaining learnable from data [9].

7 Conclusion and Future Work

Visual abstract reasoning provides a valuable step toward AGI because it requires rule induction, relational inference, and systematic generalization. RAVEN and PGM measure these capabilities in controlled settings [1,3], while ARC highlights the challenge of learning new tasks from a few examples [2]. Published results show that modern architectures can achieve high IID accuracy on standard splits, but generalization remains a central challenge across regimes [3,7].

Future work should prioritize improved object-centric representations, compositional modules that encourage reuse of rule primitives, and evaluation regimes that explicitly test out-of-distribution generalization rather than only IID accuracy. Progress along these directions can move visual reasoning systems closer to the flexibility expected from AGI.

References

[1] C. Zhang, F. Gao, B. Jia, Y. Zhu, and S. C. Zhu. RAVEN: A Dataset for Relational and Analogical Visual rEasoNing. CVPR, 2019.

[2] F. Chollet. On the Measure of Intelligence. arXiv:1911.01547, 2019.

[3] D. G. T. Barrett, F. Hill, A. Santoro, A. S. Morcos, and T. Lillicrap. Measuring Abstract Reasoning in Neural Networks. ICML, 2018.

[4] A. Santoro et al. A Simple Neural Network Module for Relational Reasoning. NeurIPS, 2017.

[5] A. Vaswani et al. Attention Is All You Need. NeurIPS, 2017.

[6] A. Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR, 2021.

[7] C. Zhang et al. Learning Perceptual Inference by Contrasting. NeurIPS, 2019.

[8] K. Zheng, Z. Zha, and W. Wei. Abstract Reasoning with Distracting Features. arXiv:1912.00569, 2019.

[9] S. Hu et al. Stratified Rule Aware Network for Abstract Visual Reasoning. AAAI, 2021.

[10] M. Jahrens et al. Solving Raven's Progressive Matrices with Multi Layer Relation Networks. arXiv:2003.11608, 2020.

[11] D. Wang et al. Abstract Diagrammatic Reasoning with Multiplex Graph Networks. OpenReview, 2020.

[12] P. W. Battaglia et al. Relational Inductive Biases, Deep Learning, and Graph Networks. arXiv:1806.01261, 2018.

[13] J. Johnson et al. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. CVPR, 2017.

[14] F. Locatello et al. Object Centric Learning with Slot Attention. NeurIPS, 2020.

[15] K. Yi et al. CLEVRER: Collision Events for Video Representation and Reasoning. ICLR, 2020.

[16] J. Andreas et al. Neural Module Networks. CVPR, 2016.

[17] S. Reed and N. de Freitas. Neural Programmer Interpreters. ICLR, 2016.

[18] Y. LeCun. A Path Towards Autonomous Machine Intelligence. OpenReview, 2022.

[19] Y. Bengio, Y. LeCun, and G. Hinton. Deep Learning for AI. Communications of the ACM, 2021.

[20] R. Zellers et al. From Recognition to Cognition: Visual Commonsense Reasoning. CVPR, 2019.

[21] D. A. Hudson and C. D. Manning. GQA: A New Dataset for Real World Visual Reasoning and Compositional Question Answering. CVPR, 2019.

[22] A. Suhr et al. A Corpus for Reasoning about Natural Language Grounded in Photographs. ACL, 2019.

[23] Y. Goyal et al. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. CVPR, 2017.

[24] R. Girdhar and D. Ramanan. CATER: A Diagnostic Dataset for Compositional Actions and Temporal Reasoning. ICLR, 2020.

[25] J. Weston et al. Towards AI Complete Question Answering: A Set of Prerequisite Toy Tasks. arXiv:1502.05698, 2015.

[26] K. Sinha et al. CLUTRR: A Diagnostic Benchmark for Inductive Reasoning from Text. EMNLP, 2019.