DARR: A Dual-Branch Arithmetic Regression Reasoning Framework for Solving Machine Number Reasoning

Chengtai Li[1,2*], Yee Yang Tan[3*], Yuting He[1,4], Jianfeng Ren[1,5†], Ruibin Bai[1,5], Yitian Zhao[2], Heng Yu[1], Xudong Jiang[6]

[1]The Digital Port Technologies Lab, School of Computer Science, University of Nottingham Ningbo China
[2]Cixi Institute of Biomedical Engineering, Ningbo Institute of Industrial Technology, Chinese Academy of Sciences
[3] School of Computer Science, University of Nottingham Malaysia
[4] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[5] Beacons of Excellence Research and Innovation Institute, University of Nottingham Ningbo China
[6] School of Electrical & Electronic Engineering, Nanyang Technological University
{scxcl2, scxyh2, jianfeng.ren, ruibin.bai, heng.yu}@nottingham.edu.cn, hfyyt11@nottingham.edu.my, yitian.zhao@nimte.ac.cn, exdjiang@ntu.edu.sg

Abstract

Abstract visual reasoning (AVR) is a critical ability of humans, and it has been widely studied, but arithmetic visual reasoning, a unique task in AVR to reason over number sense, is less studied in the literature. To facilitate this research, we construct a Machine Number Reasoning (MNR) dataset to assess the model's ability in arithmetic visual reasoning over number sense and spatial layouts. To solve the MNR tasks, we propose a Dual-branch Arithmetic Regression Reasoning (DARR) framework, which includes an Intra-Image Arithmetic Regression Reasoning (IIARR) module and a Cross-Image Arithmetic Regression Reasoning (CIARR) module. The IIARR includes a set of Intra-Image Regression Blocks to identify the correct number orders and the underlying arithmetic rules within individual images, and an Order Gate to determine the correct number order. The CIARR establishes the arithmetic relations across different images through a '3-to-1' regressor and a set of '2-to-1' regressors, with a Selection Gate to select the most suitable '2-to-1' regressor and a gated fusion to combine the two kinds of regressors. Experiments on the MNR dataset show that the DARR outperforms state-of-the-art models for arithmetic visual reasoning.

Code — https://github.com/Yang8823/DARR-MNR

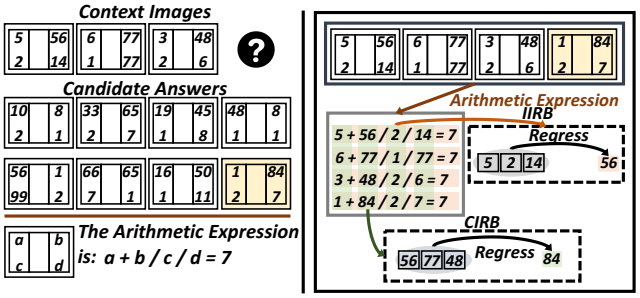

Figure 1: **Left:** A sample question in the MNR dataset. Given three context images, the task is to select the correct answer from eight candidates that follow the same arithmetic expression. **Right:** Illustration of the proposed Dual-branch Arithmetic Regression Reasoning (DARR), where Intra-Image Reasoning Block (IIRB) aims to explicitly uncover the arithmetic expression through regression reasoning within each sample image, and Cross-Image Reasoning Block (CIRB) aims to uncover the arithmetic expression by regressing the arithmetic relations across images.

Introduction

Abstract Visual Reasoning (AVR) is part of the long-standing efforts to develop artificial general intelligence (Fei et al. 2022). It is essential for evaluating machine intelligence, identifying the problems that a general AI system should solve, and providing insights into the limitations of current systems (Hernández-Orallo et al. 2016). The developed techniques have been applied beyond AVR, *e.g.*, relational networks for semantic segmentation (Mou, Hua, and Zhu 2019), relational reasoning in deep reinforcement learning (Zambaldi et al. 2019), contrastive AVR mechanisms for scene graph parsing (Huang et al. 2020), and many others (Małkiński and Mańdziuk 2023).

AVR spans a variety of tasks, including Raven's Progressive Matrices (RPMs) (Hu et al. 2021), odd-one-out (Mańdziuk and Żychowski 2019), Bongard problems (Nie et al. 2020), arithmetic visual reasoning (Zhang et al. 2020) and many emerging ones (Małkiński and Mańdziuk 2023; Li et al. 2024). Among these, arithmetic visual reasoning is a unique challenge for a solver to reason about number sense (Dehaene 2011), a special intuition that helps humans make sense of numbers, identify mathematical relations, discover underlying patterns, and infer general knowledge on numbers. Indeed, arithmetic reasoning is one of the key abilities of humans, while lacking to some extent in existing machine intelligence (Zhang et al. 2020).

Recently, Zhang et al. (2020) developed an MNS dataset

*These authors contributed equally.
†Corresponding author.

to assess visual number sense by discovering the underlying arithmetic expression and generating the missing number in the question image. However, a tiny variation in the expression may yield a totally different number. Thus, the problem is solved by a brute-force search for a specific sequence of arithmetic operations in (Zhang et al. 2020). such a search algorithm could hardly assess the reasoning ability, as a reasoning model should analyze the mathematical meaning behind numbers and comprehend the arithmetic expressions behind the observed images, rather than merely searching a perfect fit by iterating all possible enumerations. To deeply comprehend number sense, a Machine Number Reasoning (MNR) dataset is constructed in this paper. Given three context images following a common arithmetic expression, the target is to identify the image with the same expression from eight candidates, as shown in Fig. 1. This task is different from previous ones in two aspects. 1) Instead of brute-force search, the MNR task requests the model to deduce an arithmetic expression from images, assess the expression similarity between candidate and context images, and determine the one with the same expression as context images, which better assesses the reasoning ability over numbers. 2) The MNR dataset combines visual number sense and abstract spatial reasoning, while the latter is missing in the MNS dataset. In addition, more difficult rules are included to challenge the model's reasoning ability, *e.g.*, large variations in shape and spatial layout, and longer expression.

The MNR dataset poses challenges to existing reasoning models. Firstly, the underlying rules are complicated. For each task, both the order of numbers and the expression of numbers need to be deduced. It is challenging to reason out the exact expression of the ordered numbers. Secondly, numerous potential expressions can be deduced from a single image in the MNR dataset, as the evaluation result of the expression is unknown in advance and needs to be determined along with the number order and the expression itself. It is also difficult to discern the unique common expression across context images from a large feasible search space.

To tackle these challenges, we propose an end-to-end **D**ual-branch **A**rithmetic **R**egression **R**easoning (**DARR**) framework, consisting of an **I**ntra-**I**mage **A**rithmetic **R**egression **R**easoning (**IIARR**) module to explore the feasible arithmetic expressions through regression reasoning over numbers within an image, and a **C**ross-**I**mage **A**rithmetic **R**egression **R**easoning (**CIARR**) module to contrast the numbers across images to uncover the common underlying expression. Specifically, in IIARR, the encoded feature maps are first divided into patches of equal size, and one patch is randomly selected as the target, while the rest are assembled into an ordered patch sequence. To encapsulate the underlying expression, an **I**ntra-**I**mage **R**egression **B**lock (**IIRB**) is designed to regress the target patch using the remaining ones. Upon converging, the IIRB well embeds the arithmetic expression for the ordered numbers. The patch sequence is permuted to enumerate feasible number orders in the expression. To identify the one following the correct order, an **O**rder **G**ate (**OG**) is designed. In such a way, the correct number order and the arithmetic expression are explicitly embedded into the IIARR.

The proposed CIARR is designed to uncover the arithmetic relations across images and, hence, implicitly model the underlying expression. In particular, a **C**ross-**I**mage **R**egression **B**lock (**CIRB**) is designed to utilize the context images to regress the candidate image, thereby discovering the relations among the ordered numbers at corresponding positions across images. Considering the fact that the ordered numbers in context images and the correct candidate image obey a common expression, the regressed relation, hence, implicitly embeds the common expression into the CIRB. To further improve the robustness of irrelevant context features during regression, in addition to a '3-to-1' regressor that directly utilizes the three context images to regress the target, we also develop a set of '2-to-1' regressors that select two context images to regress the target, to mitigate the potential distortion from irrelevant context features. A **S**election **G**ate (**SG**) is then designed to select the more suitable '2-to-1' regressor, and two types of regressors are combined through a **G**ated **F**usion (**GF**) scheme. In such a way, the proposed CIARR implicitly but effectively encapsulates the common expression by establishing the mapping between ordered numbers across images.

Our contribution can be summarized as follows. 1) To facilitate the research on arithmetic visual reasoning, we construct the Machine Number Reasoning dataset, and propose a Dual-branch Arithmetic Regression Reasoning framework for solving MNR tasks. 2) The proposed IIARR explicitly models the underlying expressions by directly regressing a target image patch by using the rest patches, with an Order Gate to determine the correct number order. 3) The proposed CIARR implicitly captures the common expression by extracting the relations among the ordered numbers at respective positions across images through CIRB, and suppresses irrelevant features through the Selection Gate. 4) Extensive experiments show that the MNR dataset is challenging for abstract arithmetic reasoning while our DARR significantly outperforms state-of-the-art models on this dataset.

Related Work

Abstract Visual Reasoning. In the field of image reasoning (Zhang et al. 2024b,a; Song et al. 2023), AVR has recently attracted significant attention (He et al. 2025, 2024). Models for solving AVR tasks often contain two modules: one for visual perception and the other for analogical reasoning (Małkiński and Mańdziuk 2023). WReN (Barrett et al. 2018) encodes panels through a convolutional neural network (CNN) and employs a relation network to infer relations across panels. RelBase (Spratley, Ehinger, and Miller 2020) utilizes a two-stage encoder to extract visual features first and extract relations based on the features. SCL (Wu et al. 2020) utilizes three convolutional blocks for perception, and a scattering transformation to derive the relations. MRNet (Benny, Pekar, and Wolf 2021) utilizes multi-scale convolutional layers for visual perception and explores relations by enforcing the rule consistency across rows and columns. STSN (Mondal, Webb, and Cohen 2023) presents object-centric attention for visual perception and a transformer for reasoning. PredRNet (Yang et al. 2023) extracts high-level abstract relations through a prediction

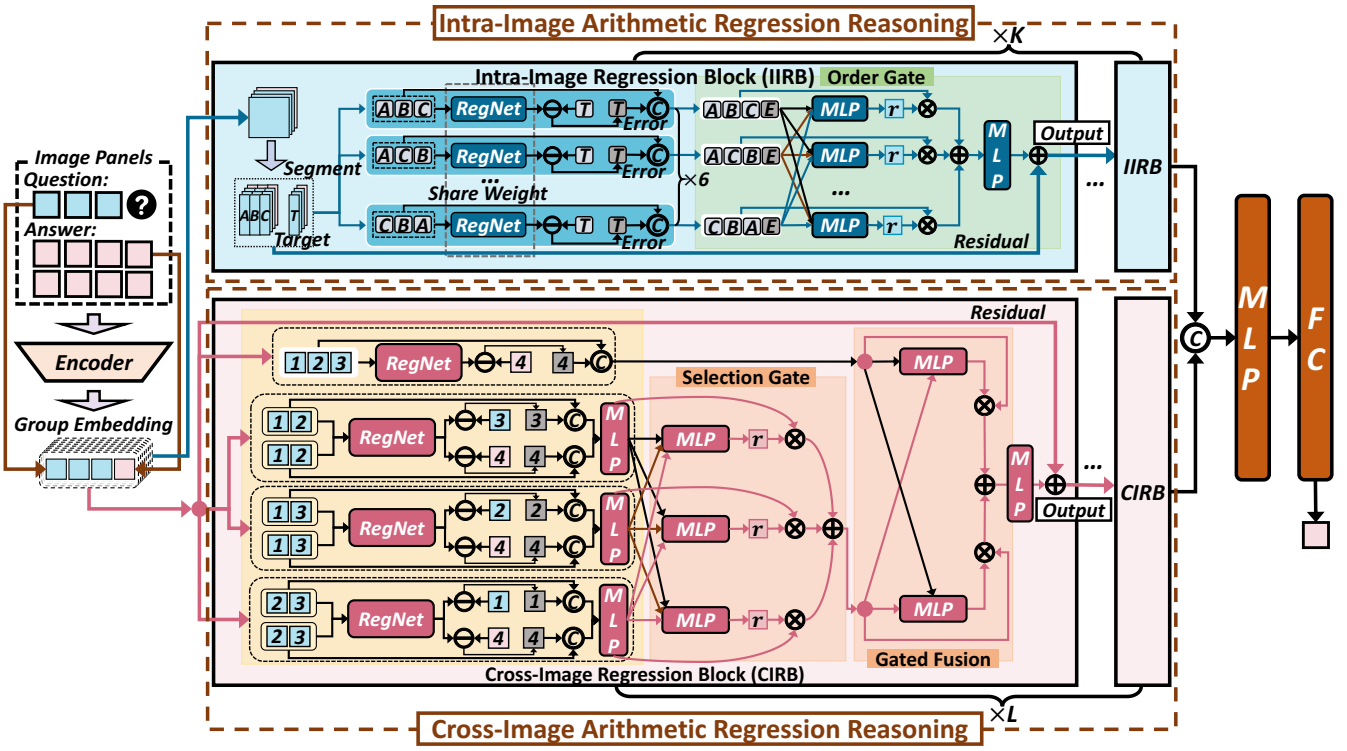

Figure 2: Overview of proposed DARR. It contains two main modules: IIARR and CIARR. The IIARR consists of K IIRBs, each containing a set of shared-weight regression networks and an Order Gate, to encapsulate arithmetic rules and number orders within an image. The CIARR contains L CIRBs, each containing one '3-to-1' regressor, three '2-to-1' regressors, a Selection Gate, and a Gated Fusion, to establish the relations across images and implicitly embed the underlying expression.

network. Recently, SCAR (Małkiński and Mańdziuk 2024) introduced a linear Structure-Aware dynamic Layer (SAL) that is adaptable for different types of tasks. Despite these advancements, existing methods still struggle to address the challenges posed by arithmetic visual reasoning.

Gate Mechanisms. Gate mechanisms control the flow of information within neural networks (Van den Oord et al. 2016), including spatial-wise, channel-wise, and layer-wise gates. Li et al. (2020) utilized fine-grained spatial-wise and channel-wise gates to fuse multi-level features for semantic segmentation. Yang et al. (2020) introduced a channel-wise gate to better cooperate or compete for information across different channels. Zhao et al. (2020) employed multi-level gate units to control the information transfer between the encoder and decoder, and suppress non-salient features. Yang et al. (2022) designed task-aware layer-wise gates to automate the selection of sub-models for specific tasks in continual object detection. In this paper, various gates are designed to select and fuse regression reasoning results.

AVR Datasets. AVR has various task formulations (Małkiński and Mańdziuk 2023), *e.g.*, RPMs such as PGM (Barrett et al. 2018), RAVEN (Zhang et al. 2019) and its variants (Hu et al. 2021; Benny, Pekar, and Wolf 2021); same-different tasks such as the SVRT dataset (Fleuret et al. 2011); odd-one-out tasks such as the CVR dataset (Zerroug et al. 2022); and arithmetic visual reasoning (Zhang et al. 2020). Among these, arithmetic visual reasoning is unique, requiring the model to reason explicit or implicit rules over numbers. In addition, Unicode Analogies Challenge (Spratley, Ehinger, and Miller 2023) increases the difficulty by including attributes with greater variability, and the ARC dataset (Chollet 2019) assesses the model's generalization ability from a few examples to a diverse set of problems. Very recently, Lu et al. (2023) introduced the MathVista dataset, which is targeted for evaluating Large Language Models and heavily integrated with textual contexts, which may lose focus on abstract visual reasoning.

Proposed DARR

Overview of Proposed DARR

As shown in Fig. 2, given three context images $\{\boldsymbol{C}_i\}_{i=1}^{3}$, the task is to identify the candidate among eight candidates $\{\boldsymbol{A}_j\}_{j=1}^{8}$ following the same rule as the context images. The rules in MNR tasks involve the number order and arithmetic expression. Since numerous number orders and arithmetic expressions can be formed, it is very challenging to determine the underlying rules from merely three context images.

To tackle the challenges, we propose a **D**ual-branch **A**rithmetic **R**egression **R**easoning (**DARR**) model. The first branch, **I**ntra-**I**mage **A**rithmetic **R**egression **R**easoning (**IIARR**) module, is designed to explicitly model the under-

lying rules. It first divides the encoded features into a set of patches of equal size. One patch is randomly selected as the target, while the remaining ones are arranged into an ordered patch sequence. Then, an **I**ntra-**I**mage **R**egression **B**lock (**IIRB**) is proposed to regress the target patch based using other patches in the same image. We enumerate all the possible permutations of the patch sequence to simulate the number orders, and design an **O**rder **G**ate (**OG**) to select the most suitable sequence with minimal regression error. The other branch, **C**ross-**I**mage **A**rithmetic **R**egression **R**easoning (**CIARR**) module, contains a set of **C**ross-**I**mage **R**egression **B**locks (**CIRB**s), each of which utilizes the context images to regress a target image, implicitly revealing the relations between ordered numbers in different images. Besides the common formulation of '3-to-1' regression, *i.e.*, utilizing three context images to regress the target, a set of '2-to-1' regressors are designed to mitigate the irrelevant features in context images and minimize their disturbance to the regression process. Furthermore, we design a **S**election **G**ate (**SG**) to select the most suitable '2-to-1' regressor with minimal regression error, and a **G**ated **F**usion (**GF**) scheme to combine the results of '3-to-1' regressor and '2-to-1' regressor. Formally, the goal of DARR is to find a mapping function between a candidate image and context images,

$$\hat{\boldsymbol{y}}_j = \mathcal{R}(\mathcal{E}(\{\boldsymbol{C}_i\}_{i=1}^3, \boldsymbol{A}_j), \tag{1}$$

where $\mathcal{E}$ is the perception module, $\mathcal{R}$ is the reasoning module, and $\hat{\boldsymbol{y}}_j$ is the predicted score for the candidate answer $\boldsymbol{A}_j$. The DARR aims to minimize Cross Entropy loss,

$$\mathcal{L} = -\sum_{j=1}^{8} \boldsymbol{y}_j \log(\hat{\boldsymbol{y}}_j), \tag{2}$$

where $\boldsymbol{y}_j$ denotes the one-hot labels for the j-th option.

Intra-Image Arithmetic Regression Reasoning

Intra-Image Regression Block. Eight ResNet blocks (Yang et al. 2023) are utilized as the perception encoder to extract image features $\{\boldsymbol{F}_i^c\}_{i=1}^3$ and $\{\boldsymbol{F}_j^a\}_{j=1}^8$ for for context images and candidate images respectively. We concatenate the context features with the features of a candidate image, and examine the probability that they share the same underlying arithmetic expression. For simplicity, we use one of the combinations $\boldsymbol{X}_1 = [\boldsymbol{F}_1^c, \boldsymbol{F}_2^c, \boldsymbol{F}_3^c, \boldsymbol{F}_1^a]$ as an example.

Specifically, we divide the feature maps into four patches of equal size. One patch $\boldsymbol{P}^t$ is randomly selected as the target and the other three $\{\boldsymbol{P}_i\}_{i=1}^3$ regress the target as,

$$\hat{\boldsymbol{P}}^t = \mathcal{F}_{\text{IIR}}(\{\boldsymbol{P}_i\}_{i=1}^3), \tag{3}$$

where $\mathcal{F}_{\text{IIR}}(\cdot)$ denotes the intra-image regression network. By minimizing the regression error $\boldsymbol{P}_\epsilon^t = \boldsymbol{P}^t - \hat{\boldsymbol{P}}^t$, we establish the relation between the target patch and the other three patches. As the patches carry feature attributes such as the numbers and their spatial layouts, such a regression relation $\mathcal{F}_{\text{IIR}}(\cdot)$ could establish the underlying expressions of the ordered numbers. Upon training convergence, the IIRB explicitly embeds the underlying expression. More importantly, the IIRB is shared among four images in $\boldsymbol{X}_1$. Such a network-sharing strategy explicitly enforces the rule consistency between the candidate and the context images, so that the same underlying rule is embedded into the IIRB for these four images. Although the regression error is not directly utilized as the training loss, it is implicitly related to the training label as follows: when the candidate does share the same underlying rule with the three context images, the regression error will be small across four images, and it will be large otherwise. The training label hence could help the IIRB converge and the inherent network design could help explicitly encapsulate the underlying rules over numbers.

Order Gate. One of the key challenges of MNR is to determine the correct number order before building up the arithmetic expression. To tackle this, we enumerate all permutations of the patch sequence and perform regression reasoning over each sequence as shown in the previous section. To select the correct sequence, an Order Gate is designed. Specifically, after the IIRB, the prediction error is concatenated back with the other three patches as $\boldsymbol{Y} = [\{\mathcal{F}_P(\boldsymbol{P}_i\}_{i=1}^3), \boldsymbol{P}_\epsilon^t]$, where $\mathcal{F}_P(\cdot)$ is a permutation function. We have 6 different permutations for the three context images and hence 6 sets of features $\{\boldsymbol{Y}_i\}_{i=1}^6$. The gate coefficient for each set of features is derived as,

$$r_i^{\text{OG}} = \sigma(\sum_{j=1}^{6} \boldsymbol{W}_{i,j}^{\text{OG}} \boldsymbol{Y}_j), \tag{4}$$

where σ represents the Sigmoid activation function, and $\boldsymbol{W}_{i,j}^{\text{OG}}$ denotes the network parameter to capture the relations among the six sets of features. When the i-th permutation does not contain the correct number order, r_i^{OG} will be small, and large otherwise. In such a way, the correct number order could be identified along with the expression. The features after the Order Gate are then obtained as,

$$\hat{\boldsymbol{Y}} = \sum_{i=1}^{6} r_i^{\text{OG}} \boldsymbol{Y}_i + \boldsymbol{X}_1, \tag{5}$$

$\boldsymbol{X}_1$ are added back to form the input features $\hat{\boldsymbol{Y}}$ for the next IIRB. To handle complex rules, we stack K IIRBs as,

$$\hat{\boldsymbol{Y}}^k = \mathcal{F}_{\text{IIRB}}^k(\hat{\boldsymbol{Y}}^{k-1}), \tag{6}$$

where $\mathcal{F}_{\text{IIRB}}^k(\cdot)$ denotes the k-th IIRB, $\hat{\boldsymbol{Y}}^{k-1}$ and $\hat{\boldsymbol{Y}}^k$ are the input and output of the k-th IIRB, and $\hat{\boldsymbol{Y}}^0 = \boldsymbol{X}_1$.

Cross-Image Arithmetic Regression Reasoning

While IIARR module explicitly establishes the underlying expression within an image, the shared network design that explicitly enforces the rule consistency across images may be a hard constraint that four images may not satisfy. To better exploit the rule consistency across images, we directly contrast the candidate image with the three context images through arithmetic regression reasoning similar to that in IIARR. Instead of regressing patches, we regress the features of a candidate image using the features of context images. The intuition behind this is that if the target image shares the same underlying expression as the context images, the regression relation from the context images to the

target will implicitly reflect this underlying expression so that the regression error will be minimized, and the error will be large otherwise. An illustrative sample, together with a detailed explanation, is provided in Supplementary Material but omitted here due to the page limit.

Cross-Image Regression Block. We design multiple regressions in CIRB to represent potential interpretations of the underlying arithmetic rules. We still use $\boldsymbol{X}_1 = [\boldsymbol{F}_1^c, \boldsymbol{F}_2^c, \boldsymbol{F}_3^c, \boldsymbol{F}_1^a]$ as an input example. We first design a '3-to-1' regression for uncovering the underlying arithmetic rules across images, *i.e.*, we utilize the three context features $[\boldsymbol{F}_1^c, \boldsymbol{F}_2^c, \boldsymbol{F}_3^c]$ to regress the target image $\boldsymbol{F}_1^a$ as,

$$\hat{\boldsymbol{F}}_1^a = \mathcal{F}_{\text{CIR3}}(\boldsymbol{F}_1^c, \boldsymbol{F}_2^c, \boldsymbol{F}_3^c), \tag{7}$$

where $\mathcal{F}_{\text{CIR3}}(\cdot)$ denotes the cross-image regression network. By minimizing the regression error $\bar{\boldsymbol{F}}_1^a = \boldsymbol{F}_1^a - \hat{\boldsymbol{F}}_1^a$, the regression network $\mathcal{F}_{\text{CIR3}}(\cdot)$ implicitly captures the underlying arithmetic expression. This error is small if the target shares the same expression as context images, and large otherwise.

The '3-to-1' regression is effective in leveraging three context images to implicitly infer the common rules, but it has limitations when the derived features cannot accurately represent image attributes such as numbers and spatial positions, or there are abundant irrelevant features, making it more challenging to precisely infer the relation across images. To tackle this issue, we design a '2-to-1' regressor by randomly selecting two context images to regress the target,

$$\tilde{\boldsymbol{F}}_1^a = \mathcal{F}_{\text{CIR2}}(\mathcal{F}_S(\{\boldsymbol{F}_i^c\}_{i=1}^3), \tag{8}$$

where $\mathcal{F}_S(\cdot)$ is a function to randomly select two context images from the given three, and $\mathcal{F}_{\text{CIR2}}(\cdot)$ is the regression network. The regression error is then calculated as $\overline{\boldsymbol{F}}_1^a = \boldsymbol{F}_1^a - \tilde{\boldsymbol{F}}_1^a$. The '2-to-1' regressor considers fewer context images, but explores broader possible arithmetic rules, increasing the likelihood of identifying a valid common rule. The complementary use of both '2-to-1' and '3-to-1' regressors enhances the robustness and accuracy of the model. By minimizing the regression errors, the CIRB gradually captures the common underlying arithmetic rules across images.

Selection Gate. $\mathcal{F}_S(\cdot)$ yield three permutations and hence we have three '2-to-1' regressions respectively. Multiple potential rules may exist in these regressions, especially when the common rules are ambiguous or have multiple valid interpretations. To identify more suitable '2-to-1' regressions, we design a Selection Gate to select the '2-to-1' regressor that best captures the distinct underlying rules,

$$r_i^{\text{SG}} = \sigma(\sum_{j=1}^{3} \boldsymbol{W}_{i,j}^{\text{SG}} \boldsymbol{Z}_j), \quad i = 1, 2, 3, \tag{9}$$

where $\boldsymbol{Z}_i$ is the features for the i-th permutation after the regression, and r_i^{SG} is a scalar representing the importance of $\boldsymbol{Z}_i$. By focusing on the features that offer clearer, more consistent arithmetic relations, the Selection Gate ensures that the model extracts the most reliable information, thereby enhancing its ability to accurately identify the underlying rule. The features after the Selection Gate are obtained as,

$$\boldsymbol{Z}^{\text{SG}} = \sum_{i=1}^{3} r_i^{\text{SG}} \boldsymbol{Z}_i. \tag{10}$$

Gated Fusion. To fuse the results of '3-to-1' and '2-to-1' regressors, we design a Gated Fusion scheme to synthesize the strengths of both regressors, ensuring that the final features reflect a more comprehensive and accurate understanding of the underlying arithmetic rules, *i.e.*,

$$\hat{\boldsymbol{Z}} = \mathcal{F}_{\text{GF}}(\boldsymbol{Z}^{\text{SG}}, \boldsymbol{Z}^{\text{CIR3}}), \tag{11}$$

where $\boldsymbol{Z}^{\text{CIR3}}$ is the features after the '3-to-1' regressor. To handle complex arithmetic rules, we stack L CIRBs as,

$$\hat{\boldsymbol{Z}}^l = \mathcal{F}_{\text{CIRB}}^l(\hat{\boldsymbol{Z}}^{l-1}), \tag{12}$$

where $\hat{\boldsymbol{Z}}^{l-1}$ and $\hat{\boldsymbol{Z}}^l$ are the input and output of the l-th CIRB $\mathcal{F}_{\text{CIRB}}^l$, respectively. The proposed CIRB effectively combines '2-to-1' and '3-to-1' regression results through Gated Fusion, enabling the model to accurately capture and reason about the common arithmetic rules across images, ultimately improving the robustness of arithmetic reasoning.

Finally, $\hat{\boldsymbol{Y}}^K$ and $\hat{\boldsymbol{Z}}^L$ are concatenated and fused through an MLP $\mathcal{F}_{\text{MLP}}$, and a two-layer fully-connected layer $\mathcal{F}_{\text{C}}$ is utilized as the classifier to make the final prediction as,

$$\boldsymbol{y} = \mathcal{F}_{\text{C}}(\mathcal{F}_{\text{MLP}}(\hat{\boldsymbol{Y}}^K, \hat{\boldsymbol{Z}}^L)). \tag{13}$$

Construction of MNR Dataset

As shown in Fig. 1, each question image in the MNR dataset consists of three context images and eight candidate images, in which the numbers in the three context images and one of the candidate answers follow a common arithmetic expression. The task is to identify the candidate answer following the same underlying rule as the three context images.

Question Panel Generation. The three context images are generated using an And-Or Graph (AOG) (Zhang et al. 2020). We define three problem types, *i.e.*, `combination`, `composition`, and `partition`, as shown in Fig. 3. Each category consists of two major attributes: a spatial and geometric attribute layout, and an arithmetic algebra. The layout consists of two attributes. 1) Geometric shape, including `circle`, `triangle`, `hexagon`, `square`, and `rectangle`. 2) Spatial relations. In the `combination` category, the relations include `inclusion`, `overlap`, and `tangent`. In the `composition` category, the arrangements include `circle`, `square`, `triangle`, `cross`, and `line`. In the `partition` category, a shape can be partitioned into 2, 4, 6, and 8 parts. The algebra has two components, *i.e.*, arithmetic objects and interpretation. The former refers to numbers and arithmetic operators, and the latter corresponds to the basic style of human cognition (Nisbett et al. 2001): holistic where all numbers are calculated together, and analytic where numbers are grouped and calculated separately. The context images are then generated based on the same rule of problem type, spatial layout and algebra, differing from specific numbers in the expression.

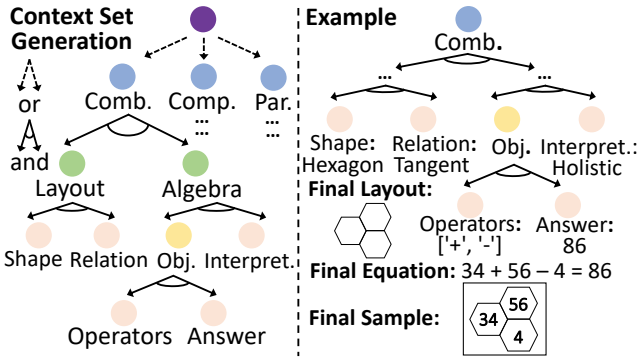

Figure 3: **Left:** AOG for context set generation with each leaf node representing a problem configuration. **Right:** An example of context set image generation in the MNR dataset.

The other seven wrong candidate answers are generated from modifying the underlying arithmetic expression by mutating the operators in the AOG of the underlying expression. For shorter analytic interpretation, the numbers of the expression are mutated following a Gaussian distribution, as empirical study shows that it may lead to clear clues to solve the problem if only mutating the operators in a short expression. More details on dataset generation are provided in the Supplementary Material.

Dataset Statistics. The MNR dataset consists of three types of problems: combination, composition, and partition, and each type includes two interpretations, holistic and analytic. The ratios for combination, composition, and partition are $3:3$, $9:5$, and $4:5$ for holistic and analytic interpretations, respectively. Additionally, for each interpretation of each problem type, there are five possible shapes that can be used to generate problems. In total, the MNR dataset consists of 145 problem configurations, with 1000 sample questions in each configuration. Each question contains 3 context images and 8 candidate answer images of size 80×80 pixels.

Evaluation Protocol. Two evaluation protocols are adopted. 1) **Protocol I**: For each configuration, the dataset is split into training, testing, and validation sets with a ratio of $6:2:2$. *e.g.*, 600, 200, and 200 samples per configuration for training, testing, and validation, respectively. This is the default protocol in this paper. 2) **Protocol II**: The dataset is split according to problem configurations. Of the 145 configurations, 87, 29 and 29 are selected as the training, validation, and test set respectively following a ratio of $6:2:2$. This protocol evaluates models in an out-of-distribution setting.

Experimental Results

Experimental Setup

The MNR dataset and three RAVEN datasets, RAVEN (RVN) (Zhang et al. 2019), I-RAVEN (I-RVN) (Hu et al. 2021) and RAVEN-FAIR (RVN-F) (Benny, Pekar, and Wolf 2021), are utilized for evaluation. Each RAVEN dataset consists of 70K question sets, where each contains 8 question images and 8 candidate answer images. The question images are generated in the same way for three datasets, but the answer images are generated through different schemes. The three RAVEN datasets have been widely utilized for evaluation abstract visual reasoning models. We strictly follow the standard evaluation protocol of the three RAVEN datasets as in (Małkiński and Mańdziuk 2024; Mondal, Webb, and Cohen 2023; Yang et al. 2023; Benny, Pekar, and Wolf 2021), *i.e.*, the dataset is randomly split into 10 folds, with 6 folds for training, 2 folds for validation and 2 folds for testing respectively. Detailed dataset descriptions are omitted here and readers may refer to the original paper for more details.

The proposed method is compared with the following state-of-the-art methods. **WReN** (Barrett et al. 2018) encodes panels through a convolutional neural network, and employs multiple relation networks to infer pair-wise relations across images. **RelBase** (Spratley, Ehinger, and Miller 2020) utilizes a 4-layer encoder for feature extraction and 1D convolutional layers to learn sequential rules. **SCL** (Wu et al. 2020) utilizes scattering transformation to compute compositional representations of images extracted by three convolutional neural networks. **MRNet** (Benny, Pekar, and Wolf 2021) utilizes multi-scale convolutional layers for feature extraction and permutation-invariant operator DIST3 to uncover underlying rules. **STSN** (Mondal, Webb, and Cohen 2023) employs a generic transformer to perform reasoning over the representation extracted by an object-centric encoder with slot attention. **PredRNet** (Yang et al. 2023) employs four residual blocks as the perception module and reasons the underlying rules through a prediction network. **SCAR** (Małkiński and Mańdziuk 2024) employs a dynamic linear layer with weights computed using a sliding window mechanism for both perception and reasoning.

The input image size is configured to 80×80. The datasets are divided into training, validation, and test sets following the standard evaluation protocols, with the validation set used for hyperparameter tuning. No additional auxiliary supervision is applied during training. The default number of IIRBs is set to $K = 3$ and the default number of CIRBs is set to $L = 3$. The maximum number of training epochs is 50. The Adam optimizer is used with an initial learning rate of 0.001. The batch size is 64.

Comparisons on MNR Dataset

We have compared state-of-the-art models for abstract visual reasoning on the MNR dataset under both Protocol I and II, where **Comb.**, **Comp.** and **Par.** denote three problem types combination, composition, and partition, and **Hol.** and **Ana.** denote two types of interpretations holistic and analytic respectively. From Table 1, the following can be observed. 1) Under both Protocols, the proposed DARR outperforms all the compared models under all settings. Specifically, compared to the second-best method, PredRNet (Yang et al. 2023), the performance gains in terms of the average accuracy are 6.0% and 5.7% under Protocol I and II, respectively, which clearly demonstrates that our approach can effectively tackle the MNR problems under various problem configurations. 2) For combination analytic and partition analytics that other models struggle with, the DARR outperforms PredRNet by 9.7% and 8.5% under protocol I, and 8.0% and 7.3% under protocol II, respectively, which suggests that the DARR is particularly well-suited for han-

Methods	Avg.	Comb.		Comp.		Par.		FLOPs (G)	Params (M)
		Hol.	Ana.	Hol.	Ana.	Hol.	Ana.		
WReN (2018)	12.3/12.5	12.2/13.18	11.0/11.5	12.3/12.5	12.6/12.1	12.6/14.1	12.6/12.5	0.28	1.21
SCL (2020)	12.5/12.5	11.6/13.5	12.2/13.0	13.0/12.9	12.8/12.0	13.1/13.0	12.3/11.8	0.07	0.13
STSN (2023)	15.6/15.4	19.7/20.3	12.7/12.0	17.3/18.7	13.1/12.7	22.2/19.5	12.7/13.0	109.85	0.57
RelBase (2020)	32.6/32.0	43.4/40.6	26.2/26.3	37.5/37.5	25.7/26.3	45.9/44.4	26.9/25.3	12.10	1.91
MRNet (2021)	35.5/35.2	46.2/46.5	28.4/27.5	40.8/41.3	27.1/26.8	53.7/52.3	28.3/28.1	5.72	4.91
SCAR (2024)	48.6/45.5	51.3/52.6	46.8/41.2	41.5/43.1	49.8/43.3	56.6/56.7	47.2/40.9	2.58	0.41
PredRNet (2023)	49.0/48.1	54.1/54.5	45.8/44.5	45.5/45.0	47.2/46.2	59.8/57.5	45.9/45.0	2.62	1.26
Proposed DARR	**55.0/53.8**	**56.2/55.3**	**55.5/52.5**	**47.0/48.2**	**56.7/55.2**	**60.1/59.3**	**54.4/52.3**	5.00	2.07

Table 1: Comparisons with state-of-the-art models on the MNR dataset under Protocol I (left) and Protocol II (right), along with the computational complexity and model size. The proposed DARR consistently and significantly outperforms all the compared methods for all settings under both Protocol I and II.

dling complex arithmetic reasoning tasks that involve intricate arithmetic logical relations. 3) Although the DARR faces difficulty in the composition-holistic setting, where its performance gain is less pronounced compared to other settings, it still outperforms PredRNet by 1.5% and 3.2% under both Protocol I and II respectively. This highlights that despite the challenges posed by certain configurations, the DARR remains a robust and effective solution. 4) Overall, the performance of other models has not surpassed 50%. WReN (Barrett et al. 2018) and SCL (Wu et al. 2020) even approach random guess. This suggests that the dataset challenges the reasoning capabilities of compared models, pushing the boundaries of current methodologies. 5) The average performance of all models in Protocol II is lower than that in Protocol I, which indicates that Protocol II, as an out-of-distribution setting, is more challenging and requires more sophisticated reasoning. 6) Lastly, the dual-branch structure of DARR leverages both intra-image and cross-image relations and hence significantly surpasses state-of-the-art methods in terms of accuracy, while requiring competitive FLOPs and model size compared to MRNet (Benny, Pekar, and Wolf 2021), RelBase (Spratley, Ehinger, and Miller 2020) and STSN (Mondal, Webb, and Cohen 2023), showing its efficiency and effectiveness in arithmetic visual reasoning.

The MNR dataset consists of a huge amount of training samples. To evaluate the generalization performance of models on smaller training datasets, we conduct experiments with training set sizes of 10%, 20%, and 50%. From Table 2, the following can be observed. 1) The proposed DARR outperforms all the compared methods for all different training set sizes. Specifically, compared to the second-best method, PredRNet (Yang et al. 2023), the performance gains using 10%, 20%, 50% and 100% of training dataset are 1.2%, 1%, 7.2% and 6.0%, respectively. The results demonstrate that DARR has robust generalization capability, even when trained on smaller datasets. 2) When reducing the training samples, other models show a significant drop in performance, *e.g.*, SCAR (Małkiński and Mańdziuk 2024) has a performance drop of 32.3% when using only 10% of training samples. In contrast, DARR maintains an accuracy of 31.0%, which suggests that DARR has better generalization ability and is more efficient at learning from limited data. 3) When training samples are limited to just 10% and 20%, all the model struggles to reason accurately due to the small sample size, resulting in low performance across all models. The results indicate that the MNR dataset is challenging and there is still a long way to go for developing a robust and generalized model for arithmetic reasoning.

Methods	10%	20%	50%	100%
WReN (2018)	13.0	12.8	12.6	12.3
SCL (2020)	12.7	12.5	12.3	12.5
STSN (2023)	12.6	14.4	12.8	15.6
RelBase (2020)	29.0	29.7	30.0	32.6
MRNet (2021)	29.4	30.2	33.3	35.5
SCAR (2024)	16.3	15.5	43.6	48.6
PredRNet (2023)	29.8	31.7	44.2	49.0
Proposed DARR	**31.0**	**32.7**	**51.4**	**55.0**

Table 2: Evaluation of models' generalization performance by using different percentages of the training dataset.

Comparisons on RAVEN Datasets

The comparison results with state-of-the-art approaches on the three RAVEN datasets (Zhang et al. 2019; Benny, Pekar, and Wolf 2021; Hu et al. 2021) are summarized in Tab. 3. Key observations include: 1) On all three RAVEN datasets, the proposed DARR outperforms all the compared methods. Specifically, compared to the second-best method, PredRNet (Yang et al. 2023), DARR achieves an average performance gain of 2.7%, demonstrating its ability to not only solve the MNR problem but also efficiently tackle other AVR challenges. 2) DARR achieves an accuracy exceeding 99.0% across all three RAVEN datasets, highlighting its exceptional reasoning capabilities. These results clearly demonstrate that IIARR effectively uncovers the intra-relationships within the images, enabling a deeper understanding of attribute patterns. Meanwhile, CIARR captures inter-relationships among different image samples, enabling comprehensive reasoning over underlying rules.

Models	Avg.	O-RVN	I-RVN	RVN-F
WReN (2018)	23.6	16.8	23.8	30.3
MRNet (2021)	83.9	84.0	81.0	86.8
RelBase (2020)	92.1	91.7	91.1	93.5
SCL (2020)	92.2	91.6	95.0	90.1
SCAR (2024)	93.8	92.8	94.7	93.9
STSN (2023)	93.6	89.7	95.7	95.4
PredRNet (2023)	96.5	95.8	96.5	97.1
Proposed DARR	**99.2**	**99.0**	**99.4**	**99.2**

Table 3: Comparison with state-of-the-art models on the original RAVEN (Zhang et al. 2019), I-RAVEN (Hu et al. 2021) and RAVEN-FAIR (Benny, Pekar, and Wolf 2021) datasets. Results of compared methods are obtained from their original papers.

Baseline	IIARR		CIARR		DARR
	w. OG	w/o OG	w. SG	w/o SG	
48.7	53.4	52.4	53.0	51.0	55.0

Table 4: Ablation study of major components. Both modules demonstrate significant performance gains.

Ablation Studies

Ablation of Major Components of DARR. We ablate the two major components of DARR, IIARR and CIARR, on the MNR dataset. The baseline model consists of the same encoder containing eight residual convolutional blocks with two fully-connected layers as the reasoner. As shown in Table 4, by adding the IIARR, a significant performance gain of 4.7% is achieved over the baseline model, demonstrating the capability of IIARR to explicitly model the underlying arithmetic expressions. The performance gain brought by CIARR is 4.3% over the baseline model, showing that CIARR can effectively discover the relations among the ordered numbers at corresponding positions across different images, and hence implicitly model the underlying rules. By adopting both IIARR and CIARR simultaneously, the accuracy is further boosted to 55.0%, demonstrating the contributions of both IIARR and CIARR. To further investigate the effectiveness of the gate mechanism, we also ablate OG and SG. If removing OG or SG, a performance drop of 1.0% and 2.0% can be observed from Table 4 respectively, demonstrating that the IIARR with OG effectively captures the number order, while the CIARR with SG is more robust in cross-image regression reasoning. The ablation results demonstrate the effectiveness of the two proposed modules.

Ablation of K and L. We further ablate the number of IIRBs K and the number of CIRBs L on the MNR dataset. From Table 5, the following observations can be made. 1) When K is increased from 1 to 3, there is a noticeable performance improvement from 52.4% to 55.0%, indicating that adding more IIRBs enhances the model reasoning capability up to a certain point. The performance for $K = 4$ drops to 53.6%, possibly due to over-fitting from using too many IIRBs. 2) When L is increased from 1 to 3 yields a performance boost of from 52.6% to 55.0%. However, when $L = 4$, the performance of CIRBs decreases to 53.8%, possibly due to overfitting. The results show that the optimal number of blocks is $K = 3$ and $L = 3$. These parameters are used as the default values for our model.

	Avg.	Comb.		Comp.		Par.	
		Hol.	Ana.	Hol.	Ana.	Hol.	Ana.
K=1	52.4	56.5	50.4	46.9	53.1	58.6	50.7
K=2	53.9	56.1	54.1	46.6	54.9	59.6	53.6
K=3	**55.0**	56.2	**55.5**	**47.0**	**56.7**	**60.1**	**54.4**
K=4	53.6	**57.5**	53.0	46.9	53.4	59.9	54.0
L=1	52.6	55.9	53.3	46.2	53.3	58.8	50.4
L=2	53.8	56.1	54.2	46.6	54.5	59.7	53.4
L=3	**55.0**	56.2	**55.5**	47.0	**56.7**	60.1	**54.4**
L=4	53.8	**56.8**	53.1	**48.3**	53.5	**61.1**	52.7

Table 5: Ablation study of hyperparameter K and L.

WReN	SCL	STSN	RelBase	MRNet	SCAR	PredRNet
12.3	12.5	15.6	32.6	35.5	48.6	49.0
50.7	**36.1**	**31.3**	**49.9**	**47.6**	**50.3**	**53.3**

Table 6: Ablation of DARR as a plug-and-play reasoner.

DARR as Plug-and-Play Reasoning Module. The proposed DARR could serve as a plug-and-play reasoning module. We replace the reasoning module of all the compared methods by DARR and summarize the comparison results in Table 6. As shown in Table 6, the proposed DARR consistently boosts the performance of all the compared models regardless of their specific perception module, demonstrating its effectiveness for abstract arithmetic visual reasoning and its generalization ability as a plug-and-play module.

Conclusion

In this paper, we propose an MNR dataset to facilitate the research on arithmetic visual reasoning over number sense and spatial layout. To tackle the challenges of solving MNR tasks, we propose a Dual-branch Arithmetic Regression Reasoning framework. The proposed IIARR module focuses on intra-image arithmetic reasoning by explicitly modeling the arithmetic rules within individual images, with the help of the Order Gate to determine the correct number order. The proposed CIARR module establishes the underlying numerical relations across images through '2-to-1' regressors and '3-to-1' regressors, and hence implicitly models the underlying arithmetic rules. Furthermore, we design a Selection Gate in CIARR to select the most suitable '2-to-1' regressor, and a Gated Fusion to fuse the results of two kinds of regressors. Extensive experiments on the MNR dataset and the three RAVEN datasets demonstrate that our DARR significantly outperforms existing state-of-the-art methods in machine number reasoning and abstract visual reasoning.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 72071116 and 61976037, in part by the Ningbo Municipal Bureau of Science and Technology under Grant 2022J171, 2022Z173, 2022Z217, 2023Z138, 2023Z223, 2023Z237 and 2024Z110, and in part by the Yongjiang Technology Innovation Project under Grant 2022A-097-G.

References

Barrett, D.; Hill, F.; Santoro, A.; Morcos, A.; and Lillicrap, T. 2018. Measuring abstract reasoning in neural networks. In *ICML*, 511–520. PMLR.

Benny, Y.; Pekar, N.; and Wolf, L. 2021. Scale-localized abstract reasoning. In *CVPR*, 12557–12565.

Chollet, F. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Dehaene, S. 2011. *The number sense: how the mind creates mathematics*. OUP USA.

Fei, N.; Lu, Z.; Gao, Y.; Yang, G.; Huo, Y.; Wen, J.; Lu, H.; Song, R.; Gao, X.; Xiang, T.; et al. 2022. Towards artificial general intelligence via a multimodal foundation model. *Nat. Commun.*, 13(1): 3094.

Fleuret, F.; Li, T.; Dubout, C.; Wampler, E. K.; Yantis, S.; and Geman, D. 2011. Comparing machines and humans on a visual categorization test. *PNAS*, 108(43): 17621–17625.

He, W.; Ren, J.; Bai, R.; and Jiang, X. 2024. Hierarchical perceptual and predictive analogy-inference network for abstract visual reasoning. In *ACM MM*, 4841–4850.

He, W.; Ren, J.; Bai, R.; and Jiang, X. 2025. Two-stage rule-induction visual reasoning on RPMs with an application to video prediction. *PR*, 160: 111151.

Hernández-Orallo, J.; Martínez-Plumed, F.; Schmid, U.; Siebers, M.; and Dowe, D. L. 2016. Computer models solving intelligence test problems: progress and implications. *AIJ*, 230: 74–107.

Hu, S.; Ma, Y.; Liu, X.; Wei, Y.; and Bai, S. 2021. Stratified rule-aware network for abstract visual reasoning. In *AAAI*, volume 35, 1567–1574.

Huang, H.; Saito, S.; Kikuchi, Y.; Matsumoto, E.; Tang, W.; and Yu, P. S. 2020. Addressing class imbalance in scene graph parsing by learning to contrast and score. In *ACCV*.

Li, C.; He, Y.; Ren, J.; Bai, R.; Zhao, Y.; Yu, H.; and Jiang, X. 2024. Regression residual reasoning with pseudo-labeled contrastive learning for uncovering multiple complex compositional relations. In *IJCAI*, volume 4, 3466–3474.

Li, X.; Zhao, H.; Han, L.; Tong, Y.; Tan, S.; and Yang, K. 2020. Gated fully fusion for semantic segmentation. In *AAAI*, volume 34, 11418–11425.

Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.-W.; Galley, M.; and Gao, J. 2023. Mathvista: evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Małkiński, M.; and Mańdziuk, J. 2023. A review of emerging research directions in abstract visual reasoning. *INFORM FUSION*, 91: 713–736.

Małkiński, M.; and Mańdziuk, J. 2024. One self-configurable model to solve many abstract visual reasoning problems. In *AAAI*, volume 38, 14297–14305.

Mańdziuk, J.; and Żychowski, A. 2019. DeepIQ: a human-inspired AI system for solving IQ test problems. In *IJCNN*, 1–8. IEEE.

Mondal, S. S.; Webb, T.; and Cohen, J. D. 2023. Learning to reason over visual objects. *arXiv preprint arXiv:2303.02260*.

Mou, L.; Hua, Y.; and Zhu, X. X. 2019. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. In *CVPR*, 12416–12425.

Nie, W.; Yu, Z.; Mao, L.; Patel, A. B.; Zhu, Y.; and Anandkumar, A. 2020. Bongard-logo: a new benchmark for human-level concept learning and reasoning. *NeurIPS*, 33: 16468–16480.

Nisbett, R. E.; Peng, K.; Choi, I.; and Norenzayan, A. 2001. Culture and systems of thought: holistic versus analytic cognition. *Psychological Review*, 108(2): 291.

Song, X.; Jin, J.; Yao, C.; Wang, S.; Ren, J.; and Bai, R. 2023. Siamese-discriminant deep reinforcement learning for solving jigsaw puzzles with large eroded gaps. In *AAAI*, volume 37, 2303–2311.

Spratley, S.; Ehinger, K.; and Miller, T. 2020. A closer look at generalisation in raven. In *ECCV*, 601–616. Springer.

Spratley, S.; Ehinger, K. A.; and Miller, T. 2023. Unicode analogies: an anti-objectivist visual reasoning challenge. In *CVPR*, 19082–19091.

Van den Oord, A.; Kalchbrenner, N.; Espeholt, L.; Vinyals, O.; Graves, A.; et al. 2016. Conditional image generation with pixelcnn decoders. *NeurIPS*, 29.

Wu, Y.; Dong, H.; Grosse, R.; and Ba, J. 2020. The scattering compositional learner: discovering objects, attributes, relationships in analogical reasoning. *arXiv preprint arXiv:2007.04212*.

Yang, B.; Deng, X.; Shi, H.; Li, C.; Zhang, G.; Xu, H.; Zhao, S.; Lin, L.; and Liang, X. 2022. Continual object detection via prototypical task correlation guided gating mechanism. In *CVPR*, 9255–9264.

Yang, L.; You, H.; Zhen, Z.; Wang, D.; Wan, X.; Xie, X.; and Zhang, R.-Y. 2023. Neural prediction errors enable analogical visual reasoning in human standard intelligence tests. In *ICML*, 39572–39583. PMLR.

Yang, Z.; Zhu, L.; Wu, Y.; and Yang, Y. 2020. Gated channel transformation for visual recognition. In *CVPR*, 11794–11803.

Zambaldi, V.; Raposo, D.; Santoro, A.; Bapst, V.; Li, Y.; Babuschkin, I.; Tuyls, K.; Reichert, D.; Lillicrap, T.; Lockhart, E.; et al. 2019. Deep reinforcement learning with relational inductive biases. In *ICLR*.

Zerroug, A.; Vaishnav, M.; Colin, J.; Musslick, S.; and Serre, T. 2022. A benchmark for compositional visual reasoning. *NeurIPS*, 35: 29776–29788.

Zhang, C.; Gao, F.; Jia, B.; Zhu, Y.; and Zhu, S.-C. 2019. Raven: A dataset for relational and analogical visual reasoning. In *CVPR*, 5317–5327.

Zhang, J.; Wang, X.; Yao, C.; Ren, J.; and Jiang, X. 2024a. Visual-linguistic cross-domain feature learning with group attention and gamma-correct gated fusion for extracting commonsense knowledge. In *ACM MM*, 4650–4659.

Zhang, W.; Zhang, C.; Zhu, Y.; and Zhu, S.-C. 2020. Machine number sense: a dataset of visual arithmetic problems for abstract and relational reasoning. In *AAAI*, volume 34, 1332–1340.

Zhang, Y.; Yu, Z.; Wang, T.; Huang, X.; Shen, L.; Gao, Z.; and Ren, J. 2024b. GenFace: a large-scale fine-grained face forgery benchmark and cross appearance-edge learning. *TIFS*, 19: 8559–8572.

Zhao, X.; Pang, Y.; Zhang, L.; Lu, H.; and Zhang, L. 2020. Suppress and balance: a simple gated network for salient object detection. In *ECCV*, 35–51. Springer.