ACRE: Abstract Causal REasoning Beyond Covariation

Chi Zhang   Baoxiong Jia   Mark Edmonds   Song-Chun Zhu   Yixin Zhu

UCLA Center for Vision, Cognition, Learning, and Autonomy

{chi.zhang,baoxiongjia,markedmonds}@ucla.edu, sczhu@stat.ucla.edu, yixin.zhu@ucla.edu

Abstract

Causal induction, i.e., *identifying unobservable mechanisms that lead to the observable relations among variables, has played a pivotal role in modern scientific discovery, especially in scenarios with only sparse and limited data. Humans, even young toddlers, can induce causal relationships surprisingly well in various settings despite its notorious difficulty. However, in contrast to the commonplace trait of human cognition is the lack of a diagnostic benchmark to measure causal induction for modern Artificial Intelligence (AI) systems. Therefore, in this work, we introduce the Abstract Causal REasoning (ACRE) dataset for systematic evaluation of current vision systems in causal induction. Motivated by the stream of research on causal discovery in* Blicket *experiments, we query a visual reasoning system with the following four types of questions in either an independent scenario or an interventional scenario:* direct, indirect, screening-off, *and* backward-blocking, *intentionally going beyond the simple strategy of inducing causal relationships by covariation. By analyzing visual reasoning architectures on this testbed, we notice that pure neural models tend towards an associative strategy under their chance-level performance, whereas neuro-symbolic combinations struggle in backward-blocking reasoning. These deficiencies call for future research in models with a more comprehensive capability of causal induction.*

1. Introduction

> "There is something fascinating about science. One gets such wholesale returns of conjecture out of such a trifling investment of fact."
>
> — Mark Twain [64]

The history of scientific discovery is full of intriguing anecdotes. Mr. Twain is accurate in summarizing how influential science theories are distilled from sparse and limited investments. From only three observations, Edmond Halley precisely predicted the orbit of the Halley comet and its next visit, which he did not live to see. From a few cathode
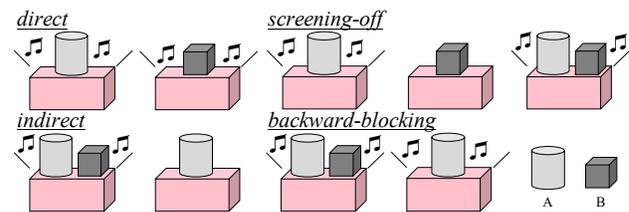

Figure 1. Abstract causal reasoning tasks administered to human participants [22, 61]. The Blicket machine possesses various activation patterns in these four cases. One needs to discover the hidden causal relations to answer two types of questions: whether object A / B is a Blicket, and how to make the machine stop / go.

rays, Joseph Thomson proved and derived the existence of electrons. From merely crossbreeding of pea plants, Gregor Mendel established the laws of Mendelian inheritance much beyond pea plants. Out of many other possible conjectures, pioneering scientists picked the most plausible ones.

The above examples of causal induction are only a few acclaimed cases of omnipresent causal reasoning scenarios in science history and our daily life. In fact, despite the notorious complexity in causal discovery, humans, even young toddlers, can felicitously identify and, sometimes, intervene in the unobservable mechanisms from only a trifling number of samples of observable events [19, 58].

This captivating commonplace trait of human cognition and its paramount connection to human learning mechanism motivate us to ask a counterpart question for modern Artificial Intelligence (AI) systems:

> *At what level do current visual reasoning systems induce causal relationships?*

To answer this question, we propose the Abstract Causal REasoning (ACRE) dataset. ACRE is inspired by the established stream of research on *Blicket* detection originally administered to young toddlers [7, 19, 20, 21, 22, 23, 35, 42, 44, 58, 60, 61, 66, 67]. The original experiments designed by Gopnik and Sobel [21] introduced a novel setup for investigating children's ability of causal induction, in which children were given a special machine referred to as "Blicket detector." Its underlying mechanism is intuitive: A Blicket detector would activate, lighting up and making

noise, when a "Blicket" was put on it. The experimenter demonstrated a series of trials to participants by placing various (combinations of) objects on the Blicket detector and showing whether the detector was activated or not. At length, the participants were asked which object is a Blicket and how to make an (in)activated Blicket machine stop (go).

This line of work's intricate nature lies in how the context and query were designed to test abstract causal reasoning beyond the simple strategy of covariation; see an illustration in Fig. 1. As a base test on causal discovery by covariation, Sobel *et al.* [61] show that children can correctly associate cause and effect using *direct* evidence. They also show that with only *indirect* evidence asserting the Blicketness of object B, children still made accurate predictions [22]. However, one must go beyond the simple covariation strategy to discover the hidden causal relations in the *screening-off* case and the *backward-blocking* case. Specifically, in the screening-off setting (Fig. 1 Top), object B (non-Blicket) is screened-off by A (Blicket) from probabilistically activating the machine [22]. The backward-blocking setting (Fig. 1 Bottom) is even more intriguing as object B, not independently tested, has undetermined Blicketness despite the fact that every appearance of it is associated with an activated machine [61]. See Section 3 for details and the supplementary for a symbolic summary.

The proposed ACRE dataset is built following a similar querying manner in the Blicket experiments to study how well existing visual reasoning systems can learn to derive ample causal information from scarce observation. In particular, inspired by the recent endeavors of visual reasoning in controlled environments [17, 32, 70], we adopt the CLEVR universe [32] in ACRE's design and add a Blicket machine to signal its state of activation, intentionally simplifying visual information processing and emphasizing causal reasoning. Following attempts made in abstract spatial-temporal reasoning benchmarks [31, 55, 72], we provide the visual reasoning system with sets of panel images as context and use image-based queries to ease language understanding, echoing the setup and the learning theories in developmental literature [19, 20, 21, 22, 23].

Specifically, each problem in ACRE consists of 10 panels: 6 for context and 4 for query. The 6 context panels are divided into two sets, the first of which serves as an introduction to the Blicket mechanism that some objects activate the machine, and others do not. This simpler set of panels resembles the introductory trials administered to children in human experiments [22, 61]. Instead of bringing in the concept of Blicket[1], in queries, we only ask a visual reasoning system to predict the state of the Blicket machine given the objects in the queries. Half of the queries concern the inde-

pendent scenarios, wherein a single object is presented, and the system is challenged to reason about whether this object is one of the causes that could activate the Blicket machine. The remaining half of the queries are for interventional scenarios, wherein we intervene in an existing context panel and ask what the state of the Blicket machine would be under the intervention. Each query is independent such that statistical bias [22, 61] and potential cheating for abstract reasoning [31, 72] are minimized. In summary, ACRE includes 30,000 abstract causal reasoning problems, supports all 4 types of reasoning queries (direct, indirect, screening-off, and backward-blocking), and is fully annotated with object attributes, bounding boxes, and masks. We further design two Out-Of-Distribution (O.O.D.) generalization splits in ACRE to evaluate models' generalizability.

In experiments, we use the ACRE dataset to analyze current visual reasoning systems' ability in causal induction. Despite remarkable results in other visual reasoning tasks, we notice that pure neural networks [8, 28, 55, 68, 77] favor a covariation-based reasoning strategy and thus can only achieve performance marginally above the chance level. As the first attempt in the exploration to empower visual reasoning systems for causal induction, we resort to neuro-symbolic models [26, 39, 43, 50, 51, 70, 71, 74, 76] that combine neural visual processing [27] and symbolic causal reasoning [18, 49, 53, 62, 78, 79], which turn out to struggle in backward-blocking cases in abstract causal reasoning.

To sum up, this paper makes three primary contributions:
- We propose the Abstract Causal REasoning (ACRE) dataset to probe current visual reasoning systems' capacity in causal induction. The dataset is inspired by the Blicket experiments and contains 30,000 problems. ACRE covers all 4 types of causal reasoning queries (direct, indirect, screening-off, and backward-blocking) with additional O.O.D. generalization splits.
- We benchmark and analyze state-of-the-art visual reasoning models in ACRE. Experimental results show that neural models tend to capture statistical correlations in observation but fail to induce the underlying causal relationships demonstrated in the trials.
- We propose neuro-symbolic combinations that improve on pure neural networks. However, our analysis shows that even with the inductive bias in causality, they still fail to distinguish a true cause from superficial covariation in backward-blocking cases. Taken together, these deficiencies call for future research in models with a more comprehensive capability of causal induction.

2. Related Work

Abstract Visual Reasoning To date, the computer vision and AI community's efforts in abstract visual reasoning primarily focus on the specific task of Raven's Progressive Matrices (RPM) [5, 52], commonly known as Intelligence

---

[1]While the notion of "Blicket" is not necessary for a visual reasoning system to solve the task, we use the term throughout this paper to simplify expressions and facilitate understanding of the core ideas.

Context Trials

Independent Queries:
What would be the machine's state given the object?

direct
A: activated

indirect
A: activated

Interventional Queries on the Fourth Trial:
What would be the machine's state given the intervention?

backward-blocking
A: undetermined
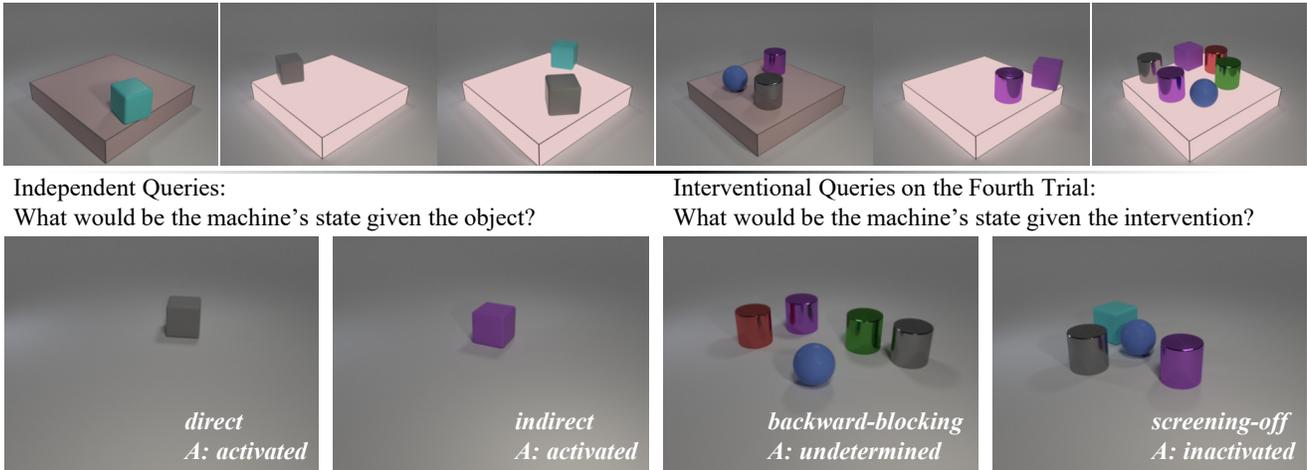
screening-off
A: inactivated

Figure 2. A sample problem in ACRE. Of the 6 context trials, we devote the first set of 3 panels for an introduction to the Blicket machinery and allow more complex configurations in the second set of panels. Queries are either on independent objects or interventional combinations for an existing trial. In this example, the first query tests causal reasoning from *direct* evidence, as the gray cube is independently tested and always associated with an activated machine. The second query requires comparing the fourth and fifth trial to realize that the Blicket machine is activated by the cube, not the cylinder, based on *indirect* evidence. As such, we infer that the red and green cylinders in the sixth trial may not activate the machine because the purple cube can already do so; despite their association with an activated machine only, their Blicketness is *backward-blocked* in the interventional trial. The cyan cube is *screened-off* by the gray cube's Blicketness from probabilistically activating the machine. Of note, the screening-off and the backward-blocking case cannot be solved by covariation.

Quotient (IQ) tests, that studies how visual reasoning systems can induce the hidden spatial-temporal transformation from limited context and apply it to derive a missing panel. Santoro *et al*. [55] extended the relational module [56] to take panel-based representation and introduced the Wild Relational Network (WReN). Zhang *et al*. [72] proposed to incorporate structural annotations in a neural modular manner. Methods considering contrast at data-level [29] or module-level [73] were later shown to improve performance significantly. Zheng *et al*. [77] formulated the problem as teacher-student learning, Wang *et al*. [68] used a multiplex graph model to capture the hidden relations, and Spratley *et al*. [63] revisited ResNet models combined with unsupervised learning. More recently, Zhang *et al*. [74] disentangled perception and reasoning from a monolithic model, wherein the visual perception frontend predicts objects' attributes, later aggregated by a scene inference engine to produce a probabilistic scene representation, and the symbolic logical reasoning backend abduces the hidden rules.

The proposed ACRE dataset complements the spectrum of abstract visual reasoning tasks by challenging visual reasoning systems with causal induction from a limited number of trials and adding missing dimensions of causal understanding into the prior spatial-temporal task set.

Causal Reasoning Equipping visual reasoning systems with causal reasoning capability has been an emerging topic in computer vision research [12, 41, 46]. Recent causal reasoning datasets [1, 70] established video-based benchmarks[2] for either trajectory prediction in counterfactual scenarios or visual question answering with explanatory, predictive, and counterfactual questions. Nevertheless, causal induction in prior computer vision research relies heavily on covariation. For instance, psychological research [2, 3, 15, 16, 34] points out that the key to solving these two problems is intuitive physics, with covariation-based causal reasoning in associating collision with object dynamics. Moreover, Edmonds *et al*. [9, 10, 11] further demonstrate that covariation would result in catastrophic failures when the visual features are similar but the underlying causal mechanisms dramatically differ. These results necessitate causal induction beyond covariation: Asymmetries in learning under various causal structures [65] refute parsimonious associative learning [59].

With a particular emphasis on causal induction beyond the simple causal reasoning strategy of covariation [36], we design ACRE with diversified causal queries, requiring a visual reasoning system to induce the hidden causal mechanism from only limited observation. From a cognitive standpoint, it is argued that Bayesian networks [48, 49] and the theory-theory [11, 19, 20, 23, 25, 60] play vital roles in abstract causal reasoning. However, how young toddlers induce accurate Bayesian representation and form a correct theory during such a short exposure remains unclear [14].

---

[2]Of note, these prior works do not echo Michotte's theory of perceived causality that humans possess a "causal detector" akin to how we perceive colors [45], as they fail to show humanlike causal perception [57, 80].
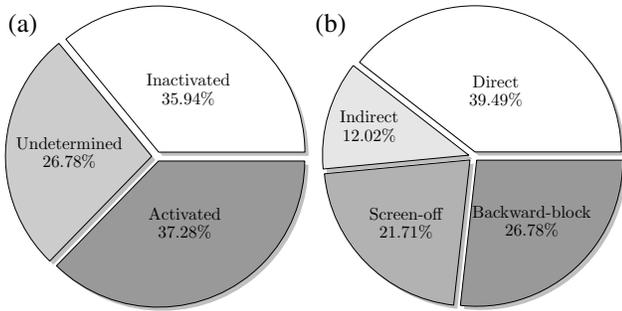
Figure 3. Distributions of (a) labels and (b) query types in ACRE.

3. Building ACRE

The ACRE dataset is designed to be light in visual recognition and heavy in causal induction. Specifically, we ground every panel in a fully-controlled synthetic environment by adopting the CLEVR universe [32] where all objects, including the Blicket machine, are placed on a tabletop with three-point lighting. All potential Blicket objects are of the same size and come with 3 possible shapes (cube, sphere, or cylinder), 2 possible materials (metal or rubber), and 8 possible colors (gray, red, blue, green, brown, cyan, purple, or yellow). For the context panels, we set all objects on a pink Blicket machine at the center of the scene and signal its state of activation by lighting it up. For the query panels, we directly put all objects on the tabletop. In both cases, objects are randomly distributed across the scenes. To avoid confusion during reference, every object is uniquely identifiable by its shape, material, and color. Other than the constraints, every object's attributes are randomly sampled from the aforementioned space. Collectively, every ACRE problem contains 5 to 8 unique objects. We keep other scene configurations the same as the original setup in CLEVR [32] and generate images by Blender [6]. Every image is also fully-annotated with object attributes, bounding boxes, and masks; see Fig. 2 for a sample problem in ACRE and refer to the supplementary for more examples.

ACRE Context Every ACRE problem contains 10 panels, of which 6 serve as context panels. Following the original design [22, 61], we further divide the 6 panels into 2 sets and use the first simpler set as the familiarization set. Specifically, for the first set of 3 panels, we randomly sample 2 objects and assign one to be a Blicket and another to be a non-Blicket. Both objects are independently tested on the Blicket machine and then placed together on it. These 3 simple trials reveal the nature of a Blicket detector: The machine will be activated when a Blicket is placed on it. For the second set of panels, we allow more random sampling; in particular, we sample another group of objects that is disjoint with the first one and partition it into 3 potentially overlapping subgroups, corresponding to the configurations for each of the rest panels. Either one or two of them are associated with an activated Blicket machine.

ACRE Query The Blicket machine's activation pattern in context panels supports all 4 types of queries (direct, indirect, screening-off, and backward-blocking) and provide sufficient clues for determining Blicketness for each object; see illustrations in Figs. 1 and 2. Based on explanations of the Blicket mechanism [20, 22, 23, 61], we detail query categorization in the following. Intuitively, an object is a Blicket if it is independently and always associated with an activated machine, which can be determined based on direct evidence; the same reasoning strategy is applicable to resolve non-Blickets. An object is also considered a Blicket if the machine activates when we place it together with other objects but not alone, and the other objects fail to activate the machine. In these cases, the Blicketness is resolved by indirect evidence; no direct observation is available. An object is considered a non-Blicket when putting it with other potential Blickets together will activate the machine, but it fails to do so by itself; this derivation is referred to as screening-off reasoning. In addition to being a Blicket or non-Blicket, an object's Blicketness could also be undetermined, which occurs when the object is not directly tested, but can activate the machine together with other potential Blickets; this is referred to as backward-blocking reasoning. Note that the Blicketness of an individual object may be undetermined, but together with other undetermined ones, they can form a set that activates the machine; as such queries happen in the indirect setting, we also refer to them as indirect reasoning.

The rich causal relations embedded in the context panels afford us to probe a reasoning system's causal induction capability. In particular, we design 4 queries in each ACRE problem, 2 for independent scenarios and another 2 for interventional scenarios, similar to the questions administered in human experiments [22, 61]. In the independent scenario, we randomly sample one object from those tested in the trials. In the interventional scenario, we pick a trial with an inactivated machine and add a set of objects randomly picked from those in the context panels. The reasoning system is then asked to tell the status of the Blicket machine after placing the objects on it, either inactivated, undetermined, or activated. To avoid statistical bias [22, 61] or potential cheating [31, 72], all queries in a problem are independent.

Generalization Splits ACRE comes with additional O.O.D. splits to measure model generalization in causal induction; we focus on compositionality and systematicity in systematic generalization [13, 24, 38, 69]. In the compositionality split, we assign different shape-material-color combinations to the training and test set and ensure the training set contains every shape, material, and color, similar to the Compositional Generalization Test (CoGenT) in CLEVR [32]. In the systematicity split, we vary the distribution of an activated Blicket detector in the context panels, with the machine lighting up 3 times in the training set and 4

times during testing. Note the strategy for causal induction remains the same irrespective of the distribution change.

In total, ACRE contains $30,000$ problems, evenly partitioned into an Independent and Identically Distributed (I.I.D.) split, a compositionality split, and a systematicity split. The dataset covers all 4 types of queries, and the label distribution is adjusted to be roughly uniform; see Fig. 3 for the label distribution and query type distribution in the dataset. Please refer to the supplementary for detailed distributions of labels and query types for each split.

4. Reasoning Systems on ACRE

This section details the deep neural models adopted to benchmark the proposed ACRE dataset and the neuro-symbolic combinations explicitly designed to incorporate inductive bias for causal induction.

4.1. Deep Neural Models

As ACRE shares the inductive nature with RPM, we test several established models designed for it [55, 68, 77]. We also test methods commonly used for linguistic or visual modeling [8, 28]. Each context-query pair is independently fed into the network and treated as a classification problem.

CNN-MLP   We concatenate context panels with the query panel in the channel dimension and use a 4-layer standard CNN architecture to extract features. The CNN architecture interleaves batch normalization and ReLU activation between convolution layers. The final convolved features are passed to a 2-layer Multilayer Perceptron (MLP) with a dropout layer with a rate of 0.5 in between the two layers.

ResNet-MLP   In this model, we replace the CNN backbone in CNN-MLP with ResNet-18 [28].

CNN-LSTM   We use a standard LSTM model [30] to further process visual features. Specifically, we independently extract image features of each panel using a CNN, append a one-hot position tag to each feature map, and pass them sequentially into the LSTM module. The final hidden state is further processed by a linear layer to produce logits.

CNN-BERT   A visual BERT [8] model is also tested. We compute image features using a CNN and follow practices in BERT: For the sequence of image features, we prepend `<CLS>`, separate the context panels and the query panel with `<SEP>`, and add position and segment embeddings. Output from `<CLS>` is then used for classification.

WReN   We adopt the WReN model proposed by Santoro *et al.* [55], which applies the relational module [56] on panel-based image representations.

LEN   LEN [77] stems from WReN but takes into account the row-wise and column-wise compositions in RPM, deeper features, and the multi-choice setting. We adapt the original LEN design to the proposed ACRE by removing branches for the column-wise composition and making the prediction on each query independent.

MXGNet   A similar strategy in LEN is used to make MXGNet [68] compatible with ACRE. The two sets of context trials are treated as rows in the model.

4.2. Neuro-Symbolic Models

In preliminary experiments, we notice that pure neural models tend to capture statistical correlation rather than modeling the hidden causal relations out of the context trials. To overcome this issue, we propose neuro-symbolic combinations and explicitly incorporate various forms of causal inductive bias for the abstract causal reasoning task.

Specifically, we draw inspirations from recent advances in neuro-symbolic literature [26, 39, 43, 50, 51, 70, 71, 74, 76] and decompose our model into a neural perception frontend and a causal reasoning backend. By design, the frontend is responsible for parsing each context trial to form an object-based representation, whereas the backend takes the symbolic output from the frontend and performs causal induction; see an overview of the method in Fig. 4.

Neural Perception Frontend   As the first attempt to solve ACRE problems, we disentangle our neural perception frontend and independently pretrain the model to parse each scene. Specifically, we use Mask-RCNN [27] with ResNet-50 FPN [28, 40] backbone. The perception model is tasked with predicting the Blicket machine's state, object masks, and object attributes (shape, material, and color) for each object in the scene. Both context and query panels in the training set of each split are used to train the frontend.

Causal Reasoning Backend   Due to its efficiency and accuracy, we use a score-based continuous optimization method, denoted as **NS-Opt**, to simultaneously learn a generalized Structural Equation Model (SEM) and derive the hidden causal relations [78, 79]. In particular, denoting the existence of object j in panel i as $X_{i,j} \in \{0, 1\}$, we can arrange the symbolic parsing results from the neural perception frontend into a data matrix $X \in \{0, 1\}^{6 \times n}$, where n equals the number of unique objects in all context panels plus the Blicket machine. A generalized SEM assumes that the state of object j is related to states of its parents via a function and can be represented as

$$X_j = f_j(X_{\mathrm{pa}(j)}) = g_j(X), \tag{1}$$

where $X = [X_1 | X_2 | \ldots | X_j | \ldots]$, and $\mathrm{pa}(j)$ denotes the parents of object j. The parent finding process is further generalized in $g_j(\cdot)$ and put into optimization constraints.

Following [79], we formulate causal discovery as an optimization problem

$$\begin{array}{ll} \underset{g:g_j, \forall j \in [n]}{\text{minimize}} & \frac{1}{n} \sum_j \ell(X_j, g_j(X)) \\ \text{subject to} & h(W(g)) = 0, \end{array} \tag{2}$$

where $W(g)_{k,j} = \|\partial_k g_j\| \, \forall k, j \in [n]$, and $h(W) = \mathrm{Tr}(e^{W \circ W} - I)$. We use $[n]$ to denote an integer set from
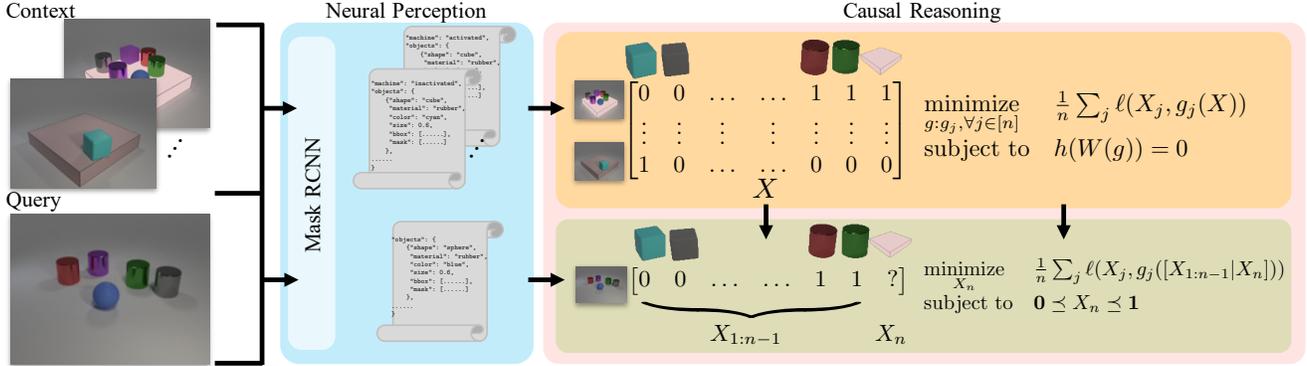
Figure 4. An illustration of the proposed neuro-symbolic combination (NS-Opt) for ACRE. The neural frontend is responsible for scene parsing. In particular, we use a Mask RCNN to detect objects and classify their attributes as well as the Blicket machine's state. The parsed results are arranged into data matrices and sent into the causal reasoning backend for optimization. A generalized SEM is learned from context trials during reasoning, which is further used to infer the state of the Blicket machine for each query.

1 to n, $\|\cdot\|$ the L^2 function norm, and $\circ$ the Hadamart product. Using the binary cross entropy loss as $\ell(\cdot, \cdot)$ for each object j, the optimization problem regularizes the generalized SEM to reconstruct the observation, while constraining the relations among the variables to be a causal Directed Acyclic Graph (DAG): $W(g)$ can be regarded as the adjacency matrix among variables, and $h(\cdot)$ a metric for acyclicity. We use an MLP for each $g_j(\cdot)$ and optimize the problem by Augmented Lagrangian; see [78, 79] for details.

With a learned generalized SEM representing the hidden causal relations in the context trials, we treat each query as another optimization problem. Specifically, we construct a partial data vector for each panel from the symbolic representation parsed by the neural perception frontend. Denoting the Blicket machine as object n, the query vector can be represented as $X_{1:n-1}$. Treating X_n as the probability of the Blicket machine being activated, the query optimization is formulated as

$$\begin{aligned} \underset{X_n}{\text{minimize}} \quad & \tfrac{1}{n}\sum_j \ell(X_j, g_j([X_{1:n-1}|X_n])) \\ \text{subject to} \quad & \mathbf{0} \preceq X_n \preceq \mathbf{1}. \end{aligned} \quad (3)$$

We solve it using L-BFGS-B [4] and set thresholds on X_n to predict the final state of the Blicket machine.

We also test a constraint-based method [18, 62] and the well-known Rescorla–Wagner (RW) model [53] for the causal reasoning backend. The constraint-based method (denoted as **NS-PC**) first uses the state-of-the-art PC algorithm [62] to test conditional independence and search for an underlying causal DAG among objects and the Blicket machine. It then finds the parent nodes for the Blicket machine and estimates its conditional probability table, which can be directly read out for each query configuration. For the RW model (denoted as **NS-RW**), we simply treat co-occurrence of an object with an activated Blicket machine as its Blicketness. A query configuration's state is predicted based on the maximum Blicketness of all objects in it.

5. Experiments

5.1. Experimental Setup

ACRE is equally partitioned into 3 splits, *i.e.*, the I.I.D. split, the compositionality (comp) split, and the systematicity (sys) split. Each of the splits contains $10,000$ problems. We further divide each split into 10 folds, with 6 folds for training, 2 folds for validation, and 2 folds for testing. All models are trained on the training sets, with hyperparameters tuned on the validation sets. Results are reported for the best models on the test sets. In particular, we report two metrics: query accuracy and problem accuracy. The former measures how a model performs on *each* query, and the later whether a model correctly answers *all* 4 queries in a problem instance. Note that based on the label distribution shown in Fig. 3, a simple strategy of always predicting activation will yield around 37.3% query accuracy and 1.87% problem accuracy, and a completely random guess would yield 33.3% query accuracy and 1.19% problem accuracy. All neural models, including the neural perception frontend in the neuro-symbolic models, are implemented in PyTorch [47] and optimized using Adam [33]. All experiments were run on an Nvidia Titan XP GPU.

5.2. Performance on the I.I.D. Setting

The first portion of Table 1 reports how various models perform under the I.I.D. setting of ACRE. Surprisingly, existing state-of-the-art methods for the abstract spatial-temporal reasoning task [68, 77] do not fare much better (even worse in certain cases) than a simple CNN-MLP model. In particular, MXGNet performs slightly worse than a random guess, only correctly answering 1% of problems. With a relational module, WReN is on par with the CNN-MLP model. CNN-LSTM and ResNet-MLP achieve similar performance, with the LSTM-based reasoning model performing better in problem accuracy. Of all pure neural mod-

Method		MXGNet	LEN	CNN-MLP	WReN	CNN-LSTM	ResNet-MLP	CNN-BERT	NS-RW	NS-PC	**NS-Opt**
I.I.D.	Qry.	33.01%	38.08%	40.86%	40.39%	41.91%	42.00%	43.56%	46.61%	59.26%	**66.29**%
	Pro.	1.00%	2.05%	3.25%	2.30%	3.60%	3.35%	3.50%	6.45%	21.15%	**27.00**%
Comp.	Qry.	35.56%	38.45%	41.97%	41.90%	42.80%	42.80%	43.79%	50.69%	61.83%	**69.04**%
	Pro.	1.55%	2.10%	2.90%	2.65%	2.80%	2.60%	2.40%	8.10%	22.00%	**31.20**%
Sys.	Qry.	33.43%	36.11%	37.45%	39.60%	37.19%	37.71%	39.93%	42.18%	62.63%	**67.44**%
	Pro.	0.60%	1.90%	2.55%	1.90%	1.85%	1.75%	1.90%	4.00%	29.20%	**29.55**%

Table 1. Performances of models on the I.I.D. split, the compositionality split (Comp.), and the systematicity split (Sys.) in ACRE. We report 2 evaluation metrics: query accuracy (Qry.) and problem accuracy (Pro.). Please refer to the experimental setup for details.

els, the BERT model achieves the best in query accuracy, slightly overtaken by CNN-LSTM in problem accuracy.

Among the 3 neuro-symbolic models tested, NS-RW strictly follows the covariation strategy in solving the causal discovery problems. We notice that such a simple causal reasoning method can only handle less than half of ACRE queries and less than 10% of ACRE problems, verifying and necessitating our efforts to create a benchmark for causal induction beyond covariation. NS-PC serves as an oracle model for causal discovery as it adopts independence tests and search methods. However, our experiments show that NS-PC is inferior to the optimization-based NS-Opt method. We believe such a result is due to the sparse and limited observation in ACRE problems, making it difficult to perform reliable independence tests. This challenge further perplexes the underlying mechanisms on how humans, even toddlers, derive the hidden relations so quickly and accurately from scarce observation. The proposed NS-Opt method successfully handles two-thirds of the queries but still has much room to improve on problem accuracy.

5.3. Performance on the O.O.D. Settings

The second and third portions of Table 1 depict the models' performance on the O.O.D. settings, *i.e.*, compositionality and systematicity. Comparing both query accuracy and problem accuracy in the compositionality split with those in the I.I.D. setting, we notice that models' performances have no significant changes. Considering the fact that the training set and the test set in the compositionality split contain completely different object attribute combinations, it is likely that neural models indeed have emerged a certain level of causal reasoning, though not perfect, to solve the problems, rather than entirely relying on statistical visual features from the training set. However, their underlying representation of causal knowledge is still elusive; future work is in need to discover their precise mechanisms.

Even if neural models emerged a causal reasoning strategy, such a strategy is not systematic, as demonstrated in the comparison between the systematicity split and the I.I.D. split. Note that only the distributions of an activated machine are different in the training set and the test set of the systematicity split, while the solutions can be derived in the same way. We note that except for the NS-PC model and the NS-Opt model, all other models experience per-

formance drop; some of them are even worse than always predicting "activated". This observation echoes the recent empirical results that pure neural models still struggle to systematically generalize [13, 37, 54].

Across the 3 splits, we also notice the conspicuously large gap between the query accuracy and problem accuracy. We hypothesize that the result indicates the existence of the bucket effect, which we verify in the next section.

5.4. A Closer Look at Queries

The drastic difference between query accuracy and problem accuracy motivates us to perform a closer inspection on how models perform on each type of queries; see Table 2 for a summary of our experimental results.

In general, we notice that neural models tend to capture causal relations by covariation. Most of them excel in query types directly solvable by this strategy, achieving the best performance in direct queries or indirect queries or both across the different splits. This effect is particularly significant in CNN-based reasoning models (CNN-MLP and ResNet-MLP) that even reach 87% accuracy for indirect queries by learning from only target labels. However, in contrast to the satisfactory performance on covariation-based reasoning, they are unable to handle the screening-off queries and the backward-blocking queries, which go beyond co-occurrence. Specifically, the best-performing neural model (CNN-BERT) embarrassingly fails on screening-off queries in the systematicity split, while CNN-based reasoning models also struggle in these settings. Among the relation-module-based models (MXGNet, LEN, and WReN), LEN and WReN are relatively stable across the different types of queries. However, with a multiplex graph, MXGNet shows different dynamics, learning best in the backward-blocking queries but counter-intuitively underperforming in the direct and indirect queries. It is also worth noting that causal reasoning that supports backward-blocking for MXGNet does not consistently enable screening-off reasoning. A converse observation is found in CNN-LSTM: The model shines in screening-off reasoning but fades in backward-blocking in 2 of the splits. Taking these results together, we hypothesize that pure neural visual reasoning systems have not yet mastered causal induction to a comparable level humans displayed in the developmental studies [22, 61].

Performance differences in queries among the neuro-

Method		MXGNet	LEN	CNN-MLP	WReN	CNN-LSTM	ResNet-MLP	CNN-BERT	NS-RW	NS-PC	**NS-Opt**
I.I.D.	D.R.	27.73%	49.07%	55.56%	51.04%	48.20%	54.87%	52.24%	88.88%	84.46%	**91.64**%
	I.D.	29.63%	45.11%	56.31%	41.04%	36.76%	48.37%	44.50%	**99.29**%	29.33%	69.25%
	S.O.	14.88%	33.68%	44.88%	29.75%	53.23%	42.29%	42.59%	7.21%	78.31%	**85.37**%
	B.B.	**59.09**%	23.91%	9.71%	35.61%	24.91%	21.12%	32.15%	1.66%	20.50%	11.98%
Comp.	D.R.	36.93%	47.58%	57.59%	55.29%	56.58%	62.79%	54.07%	91.74%	89.50%	**92.50**%
	I.D.	55.99%	52.51%	64.38%	66.94%	65.10%	70.01%	46.88%	**99.80**%	28.66%	76.05%
	S.O.	0.00%	18.01%	31.66%	8.44%	19.69%	30.52%	40.57%	4.07%	85.28%	**88.33**%
	B.B.	**52.35**%	33.63%	15.26%	35.99%	29.27%	8.54%	28.79%	0.67%	15.21%	13.48%
Sys.	D.R.	15.24%	46.22%	70.79%	53.56%	42.57%	65.19%	55.97%	92.44%	89.76%	**94.73**%
	I.D.	5.42%	47.90%	87.61%	71.35%	37.61%	85.07%	68.25%	**99.89**%	57.08%	88.38%
	S.O.	42.58%	30.91%	11.57%	16.80%	63.28%	9.57%	0.00%	0.20%	73.93%	**82.76**%
	B.B.	**56.38**%	24.89%	3.60%	31.62%	8.70%	13.38%	45.59%	0.46%	24.88%	16.06%

Table 2. A closer look at how models perform on each type of queries on different splits of ACRE: direct (D.R.), indirect (I.D.), screening-off (S.O.), and backward-blocking (B.B.).

symbolic models potentially point out an Achilles' heel for solving abstract causal reasoning problems. NS-RW's inferior performance is expected as the model only considers covariation and will surely fail the screening-off and backward-blocking queries, despite its success in direct and indirect queries. NS-RW's results also serve as a sanity check for queries in ACRE that nearly all of the direct and indirect queries can be solved by covariation (except for a minimum number of interventional cases) and nearly none of the screening-off and backward-blocking queries can (except for a minimum number of coincidences). Comparing NS-PC and NS-Opt, we notice that both models achieve fair performance on direct queries and the screening-off queries. However, the latter fares significantly better in indirect queries. We argue that the strict independence tests and search methods used in PC make the model less robust against noise, especially under the sparse-and-limited-data scenario. What is evident in both models, and more significant in NS-Opt, is their inability in differentiating the superficial correlation with an activated machine and the undetermined Blicketness within. This close inspection also indicates that adequately addressing the issue can further improve the general causal reasoning performance. By comparing the low accuracy of NS-Opt and pure neural networks in backward-blocking, we hypothesize that a potential solution to causal reasoning would be to combine the best of both worlds in learning and symbolic reasoning, keeping both the learnability of neural methods and the interpretability of symbolic methods.

6. Conclusion

In this work, we present a new dataset for Abstract Causal REasoning (ACRE), aiming to measure and improve causal induction in visual reasoning systems. Apart from the inductive reasoning nature, the defining feature of the ACRE dataset is the requirement to perform causal reasoning beyond covariation. Inspired by the established stream of research on Blicket experiments, the ACRE dataset is grounded on a similar setting using the synthetic CLEVR universe [32]. To measure causal induction beyond covariation, we challenge a visual reasoning system with 4 types of queries in either independent scenarios or interventional scenarios: direct, indirect, screening-off, and backward-blocking. The first 2 types of queries can be answered by counting co-occurrence, while the last 2 types require in-depth causal representation. To better measure generalization in causal discovery, we further propose the compositionality and the systematicity O.O.D. split.

We devise an optimization-based neuro-symbolic method to equip a visual reasoning system with the causal discovery ability. In particular, we decompose the model into a neural perception frontend and a causal reasoning backend. The neural perception frontend parses a given trial using a Mask RCNN [27], whereas the causal reasoning backend performs continuous optimization for causal discovery [78, 79]. The context trials are leveraged to learn a generalized SEM, and the answer to a query trial is solved by finding the best value to fit the SEM. As the first attempt, we separately train the two components, leaving the problem of closing the loop between visual perception and causal discovery for future work [39, 74, 75].

Existing visual reasoning systems' causal induction capability has been benchmarked on ACRE. Specifically, we notice that pure neural models tend to perform causal reasoning by capturing the statistical correlation, achieving satisfactory results on direct and indirect queries but failing on screening-off and backward-blocking ones. For neuro-symbolic models, we notice that all of them struggle on backward-blocking and that the sparse and limited observation further adds to the complexity of the problem. Comparing performances of these 2 types of models on various queries, we hypothesize that further combining learning and symbolic reasoning would be a promising direction for causal induction and broader causal reasoning problems.

At length, we hope challenges in this causal reasoning task would call for attention into visual systems with human-level spatial, temporal, and causal reasoning ability.

References

[1] Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. Cophy: Counterfactual learning of physical dynamics. In *International Conference on Learning Representations (ICLR)*, 2020. 3

[2] Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences (PNAS)*, 110(45):18327–18332, 2013. 3

[3] Neil R Bramley, Tobias Gerstenberg, Joshua B Tenenbaum, and Todd M Gureckis. Intuitive experimentation in the physical world. *Cognitive Psychology*, 105:9–38, 2018. 3

[4] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, 16(5):1190–1208, 1995. 6

[5] Patricia A Carpenter, Marcel A Just, and Peter Shell. What one intelligence test measures: a theoretical account of the processing in the raven progressive matrices test. *Psychological review*, 97(3):404, 1990. 2

[6] Blender Online Community. Blender–a 3d modelling and rendering package, 2016. 4

[7] Claire Cook, Noah D Goodman, and Laura E Schulz. Where science starts: Spontaneous experiments in preschoolers' exploratory play. *Cognition*, 120(3):341–349, 2011. 1

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019. 2, 5

[9] Mark Edmonds. Decomposing human causal leanring: bottom-up associative learning and top-down schema reasoning. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2019. 3

[10] Mark Edmonds, James Kubricht, Colin Summers, Yixin Zhu, Brandon Rothrock, Song-Chun Zhu, and Hongjing Lu. Human causal transfer: Challenges for deep reinforcement learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2018. 3

[11] Mark Edmonds, Xiaojian Ma, Siyuan Qi, Yixin Zhu, Hongjing Lu, and Song-Chun Zhu. Theory-based causal transfer: Integrating instance-level induction and abstract-level structure learning. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 3

[12] Amy Fire and Song-Chun Zhu. Learning perceptual causality from video. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):1–22, 2015. 3

[13] Jerry A Fodor, Zenon W Pylyshyn, et al. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988. 4, 7

[14] Caren A Frosch, Teresa McCormack, David A Lagnado, and Patrick Burns. Are causal structure and intervention judgments inextricably linked? a developmental study. *Cognitive Science*, 36(2):261–285, 2012. 3

[15] Tobias Gerstenberg, Matthew F Peterson, Noah D Goodman, David A Lagnado, and Joshua B Tenenbaum. Eye-tracking causality. *Psychological science*, 28(12):1731–1744, 2017. 3

[16] Tobias Gerstenberg and Joshua B Tenenbaum. Intuitive theories. In *Oxford handbook of causal reasoning*, pages 515–548. Oxford University Press New York, NY, 2017. 3

[17] Rohit Girdhar and Deva Ramanan. Cater: A diagnostic dataset for compositional actions & temporal reasoning. In *International Conference on Learning Representations (ICLR)*, 2020. 2

[18] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019. 2, 6

[19] Alison Gopnik. Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. *Science*, 337(6102):1623–1627, 2012. 1, 2, 3

[20] Alison Gopnik, Clark Glymour, David M Sobel, Laura E Schulz, Tamar Kushnir, and David Danks. A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, 111(1):3, 2004. 1, 2, 3, 4

[21] Alison Gopnik and David M Sobel. Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child development*, 71(5):1205–1222, 2000. 1, 2

[22] Alison Gopnik, David M Sobel, Laura E Schulz, and Clark Glymour. Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental psychology*, 37(5):620, 2001. 1, 2, 4, 7

[23] Alison Gopnik and Henry M Wellman. Reconstructing constructivism: Causal models, bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, 138(6):1085, 2012. 1, 2, 3, 4

[24] Jonathan Gordon, David Lopez-Paz, Marco Baroni, and Diane Bouchacourt. Permutation equivariant models for compositional generalization in language. In *International Conference on Learning Representations (ICLR)*, 2019. 4

[25] Thomas L Griffiths and Joshua B Tenenbaum. Theory-based causal induction. *Psychological Review*, 116(4):661, 2009. 3

[26] Chi Han, Jiayuan Mao, Chuang Gan, Josh Tenenbaum, and Jiajun Wu. Visual concept-metaconcept learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2, 5

[27] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2017. 2, 5, 8

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 5

[29] Felix Hill, Adam Santoro, David Barrett, Ari Morcos, and Timothy Lillicrap. Learning to make analogies by contrasting abstract relational structure. In *International Conference on Learning Representations (ICLR)*, 2019. 3

[30] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 5

[31] Sheng Hu, Yuqing Ma, Xianglong Liu, Yanlu Wei, and Shihao Bai. Hierarchical rule induction network for abstract visual reasoning. *arXiv preprint arXiv:2002.06838*, 2020. 2, 4

[32] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 4, 8

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[34] James R Kubricht, Keith J Holyoak, and Hongjing Lu. Intuitive physics: Current research and controversies. *Trends in Cognitive Sciences*, 21(10):749–759, 2017. 3

[35] Tamar Kushnir and Alison Gopnik. Conditional probability versus spatial contiguity in causal learning: Preschoolers use new contingency evidence to overcome prior spatial assumptions. *Developmental psychology*, 43(1):186, 2007. 1

[36] David A Lagnado, Michael R Waldmann, York Hagmayer, and Steven A Sloman. Beyond covariation. In *Causal learning: Psychology, philosophy, and computation*, pages 154–172. Oxford University Press, 2007. 3

[37] Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of International Conference on Machine Learning (ICML)*, 2018. 7

[38] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017. 4

[39] Qing Li, Siyuan Huang, Yining Hong, Yixin Chen, Ying Nian Wu, and Song-Chun Zhu. Closed loop neural-symbolic learning via integrating neural perception, grammar parsing, and symbolic reasoning. *arXiv preprint arXiv:2006.06649*, 2020. 2, 5, 8

[40] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[41] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou. Discovering causal signals in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[42] Christopher G Lucas, Sophie Bridgers, Thomas L Griffiths, and Alison Gopnik. When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, 131(2):284–299, 2014. 1

[43] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *International Conference on Learning Representations (ICLR)*, 2019. 2, 5

[44] Andrew N Meltzoff, Anna Waismeyer, and Alison Gopnik. Learning about causes from people: Observational causal learning in 24-month-old infants. *Developmental psychology*, 48(5):1215, 2012. 1

[45] Albert Michotte. *The perception of causality*, volume 21. Routledge, 2017. 3

[46] Roozbeh Mottaghi, Mohammad Rastegari, Abhinav Gupta, and Ali Farhadi. "what happens if..." learning to predict the effect of forces in images. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016. 3

[47] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. *NIPS-W*, 2017. 6

[48] Judea Pearl. *Causality*. Cambridge university press, 2009. 3

[49] Judea Pearl and Thomas S Verma. A theory of inferred causation. In *Studies in Logic and the Foundations of Mathematics*, volume 134, pages 789–811. Elsevier, 1995. 2, 3

[50] Siyuan Qi, Baoxiong Jia, Siyuan Huang, Ping Wei, and Song-Chun Zhu. A generalized earley parser for human activity parsing and prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 2, 5

[51] Siyuan Qi, Baoxiong Jia, and Song-Chun Zhu. Generalized earley parser: Bridging symbolic grammars and sequence data for future prediction. In *Proceedings of International Conference on Machine Learning (ICML)*, 2018. 2, 5

[52] J. C. et al. Raven. Raven's progressive matrices. *Western Psychological Services*, 1938. 2

[53] Robert A Rescorla. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Current research and theory*, pages 64–99, 1972. 2, 6

[54] Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt, and Brenden M Lake. A benchmark for systematic generalization in grounded language understanding. *arXiv preprint arXiv:2003.05161*, 2020. 7

[55] Adam Santoro, Felix Hill, David Barrett, Ari Morcos, and Timothy Lillicrap. Measuring abstract reasoning in neural networks. In *Proceedings of International Conference on Machine Learning (ICML)*, 2018. 2, 3, 5

[56] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3, 5

[57] Brian J Scholl and Patrice D Tremoulet. Perceptual causality and animacy. *Trends in Cognitive Sciences*, 4(8):299–309, 2000. 3

[58] Laura Schulz, Tamar Kushnir, and Alison Gopnik. Learning from doing: Intervention and causal inference. *Causal learning: Psychology, philosophy, and computation*, pages 67–85, 2007. 1

[59] David R Shanks and Anthony Dickinson. Associative accounts of causality judgment. In *Psychology of learning and motivation*, volume 21, pages 229–261. Elsevier, 1988. 3

[60] David M Sobel and Natasha Z Kirkham. Blickets and babies: the development of causal reasoning in toddlers and infants. *Developmental psychology*, 42(6):1103, 2006. 1, 3

[61] David M Sobel, Joshua B Tenenbaum, and Alison Gopnik. Children's causal inferences from indirect evidence: Backwards blocking and bayesian reasoning in preschoolers. *Cognitive science*, 28(3):303–333, 2004. 1, 2, 4, 7

[62] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000. 2, 6

[63] Steven Spratley, Krista Ehinger, and Tim Miller. A closer look at generalisation in raven. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 3

[64] Mark Twain. *Life on the Mississippi*. Penguin, 1984. 1

[65] Michael R Waldmann and Keith J Holyoak. Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, 121(2):222, 1992. 3

[66] Caren M Walker and Alison Gopnik. Toddlers infer higher-order relational principles in causal learning. *Psychological science*, 25(1):161–169, 2014. 1

[67] Caren M Walker and Alison Gopnik. Discriminating relational and perceptual judgments: Evidence from human toddlers. *Cognition*, 166:23–27, 2017. 1

[68] Duo Wang, Mateja Jamnik, and Pietro Lio. Abstract diagrammatic reasoning with multiplex graph networks. In *International Conference on Learning Representations (ICLR)*, 2020. 2, 3, 5, 6

[69] Sirui Xie, Xiaojian Ma, Peiyu Yu, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. Halma: Humanlike abstraction learning meets affordance in rapid problem solving. *arXiv preprint arXiv:2102.11344*, 2021. 4

[70] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations (ICLR)*, 2020. 2, 3, 5

[71] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2, 5

[72] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3, 4

[73] Chi Zhang, Baoxiong Jia, Feng Gao, Yixin Zhu, Hongjing Lu, and Song-Chun Zhu. Learning perceptual inference by contrasting. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3

[74] Chi Zhang, Baoxiong Jia, Song-Chun Zhu, and Yixin Zhu. Abstract spatial-temporal reasoning via probabilistic abduction and execution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 5, 8

[75] Chi Zhang, Yixin Zhu, and Song-Chun Zhu. Metastyle: Three-way trade-off among speed, flexibility, and quality in neural style transfer. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 8

[76] Wenhe Zhang, Chi Zhang, Yixin Zhu, and Song-Chun Zhu. Machine number sense: A dataset of visual arithmetic problems for abstract and relational reasoning. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 2, 5

[77] Kecheng Zheng, Zheng-Jun Zha, and Wei Wei. Abstract reasoning with distracting features. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2, 3, 5, 6

[78] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2, 5, 6, 8

[79] Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020. 2, 5, 6, 8

[80] Yixin Zhu, Tao Gao, Lifeng Fan, Siyuan Huang, Mark Edmonds, Hangxin Liu, Feng Gao, Chi Zhang, Siyuan Qi, Ying Nian Wu, Josh B Tenenbaum, and Song-Chun Zhu. Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering*, 6(3):310–345, 2020. 3