Abstract—Computational learning approaches to solving visual reasoning tests, such as Raven’s Progressive Matrices (RPM), critically depend on the ability to identify the visual concepts used in the test (i.e., the representation) as well as the latent rules based on those concepts (i.e., the reasoning). However, learning of representation and reasoning is a challenging and ill-posed task, often approached in a stage-wise manner (first representation, then reasoning). In this work, we propose an end-to-end joint representation-reasoning learning framework, which leverages a weak form of inductive bias to improve both tasks together. Specifically, we introduce a general generative graphical model for RPMs, GM-RPM, and apply it to solve the reasoning test. We accomplish this using a novel learning framework Disentangling based Abstract Reasoning Network (DAReN) based on the principles of GM-RPM. We perform an empirical evaluation of DAReN over several benchmark datasets. DAReN shows consistent improvement over state-of-the-art (SOTA) models on both the reasoning and the disentanglement tasks. This demonstrates the strong correlation between disentangled latent representation and the ability to solve abstract visual reasoning tasks.

I. INTRODUCTION

Raven’s Progressive Matrices (RPM) [1], [2], [3] is a widely acknowledged metric in the research community to test the cognitive skills of humans. RPM is primarily used to assess lateral thinking, i.e., the ability to systematically process the results and find solutions to unseen problems without drawing on prior knowledge [4]. The vision community has often employed Raven’s test to evaluate the abstract reasoning skills of an AI model [5], [6], [7], [8], [9]. Figure 1 illustrates a RPM question, given a $3 \times 3$ matrix, where each cell contains a visual geometric design except for the last cell in the bottom row. An AI model must pick the best-fit image from a list of six to eight choices to complete the matrix. Solving the question requires figuring out the underlying rules in the matrix as shown in Figure 1 where the correct rule is “color is constant in a row”, which eliminates other choices leaving choice “A” as the correct answer. An AI model needs to infer the underlying rules in the top two rows or columns to fill the missing piece in the last row. The visual IQ score of the above model obtained via solving abstract reasoning tasks can provide ground to compare AI against human intelligence.

Earlier computation models depended on handcrafted heuristics rules on propositions formed from visual inputs to solve RPM [10], [11]. The lack of success in the previous approaches and the inclusion of large RPM datasets [12], PGM [13], RAVEN [14] facilitated the employment of neural networks to solve abstract reasoning tasks [5], [13], [14], [12], [15]. Until now, these abstract reasoning methods have employed existing deep learning methods such as CNN [16], ResNet [17] to improve reasoning but largely ignoring to learn the visual attributes as independent components. Even though these models have improved abstract reasoning tasks, the performance is still sub-optimal compared to humans. These setbacks to the model performance are caused due to the lack of adequate and task-appropriate visual representation. The model should learn to separate the key attributes needed for reasoning as independent components.

These critical attributes, aka disentangled representations, [18], [19] break down the visual features to their independent generative factors, capable of generating the full spectrum of variations in the ambient data. We argue that a better-disentangled model is essential for the better reasoning ability of machines. A recent study [20] via impossibility theorem has shown the limitation of learning disentanglement independently. The impossibility theorem states that without any form of inductive bias, learning disentangled factors is impossible. Since collecting label information of the generative factors is challenging and almost impossible in real-world datasets, previous works have focused on some form of semi-supervised or weakly-supervised methods. Few of the prior works in disentanglement using inductive bias involve [21] that uses a subset of ground truth factors, [22] that formed pair of images with common visual attributes on a subset of factors, and [23] where the factors in a pair of images are ranked on the subset of factors. Our work improves upon the model’s reasoning ability by using the inductive reasoning present in the spatial
features. Utilizing the underlying reasoning, i.e., rules on visual attributes in RPM induces weak supervision that helps improve disentanglement, leading to better reasoning.

[12] investigated the dependency between ground truth factor of variations and reasoning performance. We take a step further and consider jointly learning disentangled representation and learning to reason (critical thinking). Unlike the above-proposed model, i.e., (working in a staged process to improve disentangling or improve downstream accuracy), we work on the weakness of both components and propose a novel way to optimize both in a single end-to-end trained model. We demonstrate the benefits of the interaction between representation learning and reasoning ability. Our motivation behind using the same evaluation procedure by [12] is as follows: 1) the strong visual presence, 2) information of the generative factors help in demonstrating the model efficacy on both reasoning accuracy and disentanglement (strong correlation), 3) possibility of comparing the disentangled results with state-of-the-art disentangling results.

In summary, the contributions of our work are threefold:
1) We propose a general generative graphical model for RPM, GM-RPM, which will form the essential basis for inductive bias in joint learning for representation + reasoning.
2) Building upon GM-RPM, we propose a novel learning framework named Disentangling based Abstract Reasoning Network (DAREN) composed of two primary components – disentanglement network, and reasoning network. It learns to disentangle factors and uses the representation to detect the underlying relationship and the object property used for the relation. To our knowledge, (DAREN) is the first joint learning framework that separates the underlying generative factors and solve reasoning.
3) We show that DAREN outperforms all state-of-the-art baseline models in reasoning and disentanglement metrics, demonstrating that reasoning and disentangled representations are tightly related; learning both in unison can effectively improve the downstream reasoning task.

II. RELATED WORKS

Visual Reasoning. Solving RPM have recently gained much attention due to their high correlation with human intelligence [3], [13], [14]. Initial works based on rule-based heuristics such as symbolic representations [1], [7], [8], [9] or relational structures [6], [24], [25] failed to comprehend the reasoning tasks due to their underlying assumptions. These assumptions include access to the symbolic representations of images, domain expertise on the underlying operations, and comparisons that help solve the task. [26] proposed a systematic way of automatically generating RPM using first-order logic to try to understand and solve these tasks fully. Growing interest introduced two RPM dataset [13], [14], which led to significant progress in solving reasoning tasks [27], [28], [5], [13].

Disentanglement. Recovering independent data generating ground truth factors is a well-studied problem in machine learning. In recent years there is renewed interest in unsupervised learning of disentangled representations [29], [30], [31], [32], [20], [19], [33]. Nevertheless, this research area has not reached a major consensus on two major notions: i) no widely accepted formalized definition [18], [19], [20], [33], ii) no single robust evaluation metrics to compare the models [34], [30], [35], [36], [37]. However, the key fact common in all models is the recovery of statistically independent [19] learned factors. A majority of the research follows the definition presented in [18], which states that the underlying generative factors correspond to independent latent dimensions, such that changing a single factor of variation should change only a single latent dimension while remaining invariant to others. Recent work [20] showed that it is impossible to learn disentangled representation without the presence of inductive bias, prompting the shift to semi-supervised [38], [39], [40] and weak-supervised [41], [22], [42], [43] disentangling models.

Recent reasoning works [27], [28] have focused on Raven [14], PGM [13] datasets for evaluation. In contrast to the above, we focus on learning both disentangled representation and solving abstract visual reasoning. Our work is inspired by the large-scale study in [12], suggesting dependence between learning disentangled representations and solving visual reasoning. Our proposed framework leverages join learning, which improves both the reasoning and the disentanglement performance. Since we quantify disentanglement score along with reasoning accuracy, the datasets used in [12] are well suited compared to Raven [14], PGM [13] which are not adapted for quantitative evaluation of disentanglement.

III. PROBLEM FORMULATION AND APPROACH

We begin by describing the problem of RPM in the domain of visual reasoning task in Section III-A, where we elaborate on the process of what constitutes valid RPM. Next, we propose our general generative graphical model for RPM, GM-RPM, which will form the essential basis for inductive bias in joint learning for representation and reasoning in Section III-B. Finally, in Section III-C, we describe our learning framework a.k.a Disentangling based Abstract Reasoning Network (DAREN) based on a variational autoencoder (VAE) and a reasoning network for joint representation-reasoning learning.

A. Visual Reasoning Task

The Raven’s matrix denoted as $M$, of size $M \times M$ contains images at all $i,j$ location except at $M_{MM}$. The aim is to find the best fit image $a^*$ at $M_{MM}$ from a list of choices denoted as $A$. For our current work, we follow the procedure by [12] to prepare RPM. Similar to prior work, we have fixed $M = 3$, where $M = \{x_{1,1}, \ldots, x_{3,3}\}$ in row-major order and $M_{33}$ is empty that needs to be placed with the correct image from the choices. We also set the number of choices $|A| = 6$, where $A = \{a_1, \ldots, a_6\}$. We improve upon the prior work by formulating an abstract representation for the matrices $M$ by defining a structure $S$ on the image attributes $(o)$ and relation types $(r)$ applied to the image attributes:

$$S = \{(r, o) : r \in R \text{ and } o \in O\}.$$
The set \( R \) consists of relations proposed by [1] that are constant in a row, quantitative pairwise progression, figure addition or subtraction, distribution of three values, and distribution of two values. We assume images are generated from underlying ground truth data generative factors \( (K) \) that constitute RPM. These image factors \( (O) \) consist of the object type, size, position (XY-axis), and color. The structure \( S \) is a set of tuples, where each tuple is formed by randomly sampling a relation from \( R \) and image attribute from \( O \). For instance, if \( S = \{ \text{constant in a row, color}, \text{quantitative pairwise progression, size}\} \), every image in each row of \( M \) will have the same (constant) value for attribute color, and progression relation instantiated on size of images from left to right. This set of \( S \) can contain a max of \( |R| \times |O| \) tuple, where the problem difficulty rises with the increase in the size of \( R, O \) or \(|S|\) or any combination of them.

**Generating RPM.** Using \( S \), multiple realizations of the matrix \( M \) are possible depending on the randomly sampled values of \((r,o)\). We use \( o \) to denote the image attributes in \( S \), in the example above \( S = \{ \text{color, object type} \} \). In the generation process, we sample values for attributes in \( o \) that adhere to their associated relation \( r \) and the values for image attributes in \( O = O \setminus o \), that are not part of \( S \), are sampled randomly for every image. Next, we sample images at every \( M_{ij} \) where the image attribute values matches with the values sampled above for \( o \cup \sigma \). For the matrix \( M \) to be a valid RPM the sampled values for \( \sigma \) must not comply with the relation set \( r \in S \) across all rows in \( M \). However, in the above example where \( \sigma = \{ \text{position, size} \} \), a valid \( M \) can also have the same values for position or size (or both) in a row as long as they do not adhere to any relations in \( S \) for more than one row. The above is an example of a distractor, where the attributes in \( \sigma \) during the sampling process might satisfy some \((r,o)\) in \( S \) for any one row in \( M \) but not for all \( M \) rows. These randomly varying values in \( \sigma \) add a layer of difficulty towards solving RPM. The result of the above generation steps produces a valid RPM matrix \( M \). The task of any model trained to solve \( M \) has to find \( r \) that is consistent across all rows or columns in \( o \) and discard the distracting features \( \sigma \). In the rest of the paper, we focus on the row-based relationship in RPM.

**B. Inductive Prior for RPM (GM-RPM)**

While previous works have made strides in solving RPM [6], [7], [12], the gap in reasoning and representation learning between those approaches and the human performance remains. To narrow this gap, we propose a minimal inductive bias in the form of a probabilistic graphical model described here that can be used to guide the joint representation-reasoning learning. Figure 2 defines the structure of the general generative graphical model for RPM. This model describes an RPM \( M = \{ x_1, \ldots, x_{MM} \} \), where \( x_{ij}, i,j = 1, \ldots, M, \)

1 We interchangeably use \( o \) to denote the subset of image attributes that adhere to rules of RPM as well as the multi-hot vector \( o \in \{0,1\}^K \) whose non-zero values index those attributes.

2 Our solution could trivially be extended to address columns or both rows and columns.

**Fig. 2:** Generative model for RPM. See Sec. III-B for details.

denote the images in the puzzle, with the correct answer at \( x_{MM} \), defined by rule \( r \) and factors \( o \).

Latent vectors \( z_{ij} \in \mathbb{R}^{K+N} \) are the representations of the \( K \) attributes, to be learned by our approach, and some inherent noise process encompassed in the remaining \( N \) dimensions of \( z_{ij} \), \( z_{ij,n} \in \mathbb{R}^N \), which we refer to as nuisances. Ideally, some \( K \) factors in \( z_{ij} \) should be isomorphic to the attributes themselves in this simple RPM setting, after an optimal model is learned. We index those \( K \) relevant factors with a hierarchical indexing model, illustrated in Figure 3.

The latent attribute selection vector \( o \in \{0,1\}^K \) determines which, among the \( K \) possible, factors are used in the puzzle. This vector is embedded over a larger attribute-noise selection vector \( o_{KN} \in \{0,1\}^{K+N} \). In \( o_{KN,J} = 0 \), \( j = K+1, \ldots, K+N \) indicate the factors corresponding to nuisance, determination of which is a part of the inference process, as defined below in (2).

This latent vector gives rise to ambient images through some stochastic nonlinear mapping \( x \sim p(x|f(z(\Theta))) \), where \( Z = \{ z_{ij} \}_{i,j \in M \times M} \in \mathbb{R}^{M \times M \times (K+N)} \) is the mapped latent tensor for RPM, parameterized by \( \Theta \) which is to be learned,

\[
p(M|Z, \Theta) = \prod_i \prod_j p(x_{ij}|f(z_{ij}|\Theta)), \tag{1}
\]

The RPM inductive bias comes from the way (prior) \( z_{ij} \) is formed, given the unknown rule \( r \). Specifically,

\[
z_{ij} = o_{KN} \odot [o \odot z_{ij,o} + \sigma \odot z_{ij,n}] + o_{KN} \odot \left[ 0_K \odot z_{ij,n} \right], \tag{2}
\]

3 We drop RPM indices, where obvious.
where $z_{ij,o} \in \mathbb{R}^K$ is the latent representation of the factors that are used in rule $r$, $z_{ij,\pi} \in \mathbb{R}^K$ is the latent representation of the complementary, unused factors$^4$.

The key in RPM is to define the priors on factors. The factors used in the rule, grouped as the tensor $Z_r = [z_{ij,r}]_{M \times M}$, follow a joint density over $r = 1, \ldots, M$ in row $i$:

$$p(Z_r) = \prod_{i} p(z_{i,\pi,r}) = \prod_{i} p(z_{i,1,\pi}, \ldots, z_{i,M,\pi,r}), \quad (3)$$

where $z_{i,\pi,r}$ is the matrix of size $K \times M$ or all latent representations in row $i$ of RPM. The factors not used in the rule, $\theta$, and actors representing the noise information have a different, iid prior (refer inner-most plate, “columns: $M$”, in Figure 2)

$$p(Z_{\theta}) = \prod_{i} \prod_{j} p(z_{ij,\pi}) = \prod_{i} \prod_{j} N(z_{m,j,0;0,1}). \quad (4)$$

We assume that all $K + N$ factors are independent

$$p(z_{ij,\pi}) = \prod_{k} p(z_{ij,\pi}^k), \quad p(z_{ij,\pi}^k) = \prod_{l} p(z_{ij,\pi}^k). \quad (5)$$

This gives rise to the full Generative Models (GM-RPM),

$$P(M; Z, Z_\theta, Z_\pi, Z_{\ast}, r, o, o_{KN}) = p(M|Z)p(Z_{\theta})p(Z_\pi)p(Z_{\ast}, o, o_{KN})p(Z_r)p(Z_{\pi}). \quad (5)$$

**Inference Model.** As described in Section III-B, the goal is to infer the value of the latent variables that generated the observations, i.e., to calculate the posterior distribution over $p(Z, o, r|M)$, which is intractable. An approximate solution for the intractable posterior was proposed by [44] that uses a variational approximation $q(Z, o, r|M; \phi)$, where $\phi$ are the variational parameters. In this work, we further define this variational posterior as

$$q(Z, o, o_{KN}; r|M; \phi) \propto q_{Z}(Z|Z', r, o, o_{KN}) q_{r}(r|Z', o) q_{o}(o|o_{KN}, Z') \prod_{ij} q_{o}(o_{ij}|x_{ij}), \quad (6)$$

where $Z'$ is an intermediate variable which is used to arrive at the final estimate of $Z$ using the Factor Consistency inference as described further in this section.

Currently, DAREN is designed for $r = \text{“constant-in-a-row”}$, i.e., $q_{r}(r|Z'; o) = \delta(r = r_{\text{const}})$ from (6). Next, we describe the sequential inference process for all the latents.

**Infer $Z'$:** Intermediate latent factors $Z'$ are first inferred independently for each element $x_{ij}$ using a general stochastic encoder $q_{o}(o_{ij}|x_{ij})$ of the VAE family:

$$q_{o}(Z'|M) = \prod_{ij} q_{o}(Z'_{ij}|x_{ij}). \quad (7)$$

Our framework accepts arbitrary choices of the VAE-family encoders, as discussed in Sec. III-C.

**Infer $o_{KN}$:** To infer $K$, we prune out the $N$ nuisance attributes from $Z' \in \mathbb{R}^{K+N}$ that have collapsed to the prior $\left(q_{o}(z'_{ij}|x_{ij}) = p(z'_{ij})\right)$. Thus the remaining latent dimensions form the relevant $K$ attributes. This is similar to computing the empirical variance $\forall(Z')$ to set the indices of $o_{KN}$ where the variance is above a threshold ($\epsilon = 0.05$).

$$q_{o}(o_{KN}|Z') = \delta(o_{KN} - o_{KN}(Z')) \quad (8)$$

$$o_{KN}(Z') = \mathbb{1}_{V(Z') > \epsilon}, \quad (9)$$

where $\mathbb{1}_{V(x) > \epsilon}$ is the multi-hot indicator vector whose entries are set to $1$ for the $K$ largest values of $a(x)$ for which $a(x) \geq b$ holds. These $K$ factors model the actual ground truth factors, while the remaining $N$ factors $z'_{ij,n}$ are considered nuisances.

**Infer $o$:** Next, we use the multi-hot vector $o_{KN}$ to set only the selected attribute for the given instance of RPM. For a given $M$ and $r = \text{constant in a row}$, $o_{KN,r}$ values remains the same for images in row $i$. We utilize KL divergence as a measure over all pairwise $z'_{ij,n}$ and set only on the indices with $l$-lowest divergence values to arrive at

$$q_{o}(o|o_{KN}, Z') = \delta(o - o_{KN}(o_{KN}, Z')) \quad (10)$$

$$o_{KN}(Z') = \mathbb{1}_{o_{KN}} \quad (11)$$

$$\delta_{KL}(k) = \frac{1}{M^2} \sum_{m,i,j=1}^{M} D_{KL}(q_{o}(z'_{ij,m}|x_{ij})||q_{o}(z'_{m,i}|x_{im})) \quad (12)$$

where, for $k = 1, \ldots, K$.

**Infer $Z$ using Factor Consistency:** We describe the process of estimating $Z$ from the intermediate variable $Z'$ using $q_{Z}(Z|Z', r, o, o_{KN})$, for the chosen case of $r$; the goal here is to obtain consistent, denoised final estimates of the factors, given the intermediate noisy estimates $Z'$ and the estimated relevant factors $o$. Specifically,

$$\hat{z}_{ij} = \delta_{KN} \left[ \hat{\phi} \odot f_{avg}(z'_{ij,0}) + \hat{\eta} \odot z'_{ij,n} + \hat{\theta}_{KN} \odot 0_{K}\right], \quad (13)$$

and $z_{ij}(Z'|r, o, o_{KN}) = \delta(\hat{Z} - \hat{Z}).$

Since the relation $r$ acts on the latent vector $Z_{\theta}$, we apply the averaging strategy on it. For $r = r_{\text{const}}$, the averaging strategy is a variant of the method in Multi Level VAE [22] described for row $i$ as:

$$f_{avg}(z'_{ij,o}) = \frac{1}{M} \sum_{j=1}^{M} z'_{ij,o}. \quad (14)$$

Using (13) & (14), we obtain updated rule-attribute constrained latent representations $Z$ for each in $M$. The resulting $Z$ is given as an input to the decoder network to learn to reconstruct the original Raven’s matrix, which we denote $M$.

**C. DAREN**

Inspired by GM-RPM, we propose a novel framework named Disentangling based Abstract Reasoning Network (DAREN). Please refer to Figure 4 for an overview of DAREN. DAREN is composed of two primary components, a variational autoencoder (VAE) module and a reasoning module. Using Section III-B described above our variational autoencoder (VAE) learns $q_{o}(Z|M)$ (Z is the final estimate) and $p_{r}(M|Z)$, where the former is referred as encoder or inference model and the later as decoder or generative model. We assume that the factors vary independently, hence to drive to maximize statistical independence we append the VAE evidence lower bound objective with the Total Correlation (TC) term [30],

$$TC(Z) = \prod_{i} \prod_{j} D_{KL}(q(z_{ij})||p(z_{ij})) \quad (15)$$
Fig. 4: Illustration of DAReN. It consists of a VAE-based generative and reasoning network. Generative Model. The encoder encodes $M$ and $A$ to $Z'$ and $Z_a'$. The $K$ possible attributes are learned from $M$ by picking the factor indices with high variance, and the rest are kept as nuisances $(N)$, performed by the threshold function $T_i$. Next, with $S_L$, we further split the $K$-D latent representation to two groups of vectors $Z_a'$ and $Z_o'$ related to the active $(set bits of o)$ and inactive $(set bits of o)$ attributes. No operation is performed on the factor indices at the set bit of $\pi$ and the nuisance factors. The rule constraint is enforced on the set bit of $\pi$ to take the same value via an averaging strategy (G). The decoder recovers the updated latent representation of the correct choice image $a_o$ to reconstruct the image back $(M)$. Reasoning Model. We consider the latent representation $Z_{meta}$ to extract the standard deviation across factor index for the top two rows and all possible six rows. An MLP trained on the concatenated standard deviation of the top two rows with choice $a_o$ predicts the best fit image.

The form of the augmented ELBO objective is described as:

$$L_{o,q}(M) = E_{p(M)}[E_{q_o(Z_o|M,r,q_0|Z_{meta}|M,r,q_0|Z_{meta}|M)}[\log p_X(M|Z)]] - \lambda_1 [D_{KL}[q(Z_o|M,r)|p(Z_o|r)] + D_{KL}[q(Z_o|M,r)|p(Z_o|r)] + D_{KL}[q(Z_o|M)|p(Z_o)] - \lambda_2 J(C(Z)), \tag{16}$$

where hyperparameter $\lambda_1, \lambda_2$ controls the weight on the KL-divergence and the TC respectively.

Reasoning Module. The reasoning component of DAReN incorporates a relational structure to infer the abstract relationship on the attribute $o$ for images in $M$. The reasoning module receives disentangled representations, $Z$ (of $M$) and $Z_A$ (of choices $A$). We prepare $Z_{meta}$ of size $(M + |A| - 1) \times M$ by iteratively filling each choice as the missing piece.

Therefore, solving $M$ is equivalent to finding the correct row in $\{M, \ldots, M + |A| - 1\}$, that satisfies the same rules shared by the top $M - 1$ rows. We compute the variance for all $M$ representation in each row $i$, $\sigma(Z_{meta}(i))$ over each dimension in the latent representation. The above process is applied to all the $M + A - 1$ rows that include all $|A|$ probable last rows. Next, we concatenate the variance vector of the top $M - 1$ rows with each probable last row (Figure 4) to prepare $|A|$ choice variance vectors, $[\sigma(Z_{meta}(i))] \times M \times (K + N)$. We feed this concatenated variance vector to a three-layered MLP ($\psi$) with a ReLU and a dropout layer [45], the probability of $\pi$ is estimated as:

$$p_i = \psi([\sigma(Z_{meta}(i1)) \cdots \sigma(Z_{meta}(M-1)) \sigma(Z_{meta}(i))]) \tag{17}$$

where $i \in \{M, \ldots, (M + |A| - 1)\}$ corresponds to $|A|$ probable choices. The choice with the highest score is predicted as the correct answer. The above process is trained using a Cross Entropy loss function.

IV. EXPERIMENTS

A. Datasets, Baselines, Experimental Setup

We study the performance of DAReN on six benchmark datasets used in disentangling work, (i) dsprites [46], (ii) modified dSprites [12], (iii) shapes3d [30], (iv-vi) MPI3D – (Real, Realistic, Toy) [47]. We use experimental settings similar to [12] to create RPM for the above datasets. The training procedure in [12], referred as Staged-WReN, is used as our baseline model. The Staged-WReN is a two-stage training process, where a disentangling generative network is first trained ($\sim 300K$ iterations), followed by training a Wild Relational Network (WReN) [13] on RPM using the representation (at $300K$) obtained from the trained encoder. Building on Staged-WReN, we propose an adapted baseline referred to as E2E-WReN, where we jointly train both the disentangling and the reasoning network from end-to-end. To train DAReN, we use a warm start by initializing only its VAE parameters with a partially trained Factor VAE [30] model for $\sim 100K$ on the above datasets (not RPM instances) followed by training DAReN for $200K$ iterations on the RPM generated from the datasets. We evaluate the model’s performance on fresh RPM samples. Refer to Appendix for details on the dataset, experimental setup, E2E-WReN, and additional results.

B. Evaluating Abstract Visual Reasoning Results

We compare our results in both the reasoning accuracy and the disentanglement scores against the SOTA methods. Both scores are important in the context of interpretable reasoning models, as the reasoning accuracy alone does not necessarily reflect the discovery of the underlying RPM factors. Both scores are shown in Table I.

Reasoning. We report the performance of reasoning accuracy in the column “Reasoning” in Table I. Our proposed model DAReN, compared against Staged-WReN and E2E-WReN, shows an improvement of $\sim 2 - 17\%$ and $\sim 0 - 2\%$ respectively. It is seen from Table I, our model outperforms the prior work, i.e. Staged-WReN by a margin of $\sim 2-17\%$, especially for datasets with color as a ground truth factor (all excluding dsprites). The VAE model DAReN can firmly separate color attributes from other factors compared to Staged-WReN.

The WReN module is similar for both E2E-WReN vs. Staged-WReN, where it learns a relational structure by aggregating all pairwise relations on the latent space within the images in $M$ and between $M$ and candidates in $A$.
However, a joint optimization of reasoning + representation (E2E-WReN) learns to solve reasoning task better than Staged-WReN. Despite the improvement via joint optimization, WReN performance is still sub-optimal in learning the underlying reasoning patterns. For each choice filled RPM, the pairwise relation tensor \( (6 \times 9 \times 9) \) contains intra-pairwise relations formed within \( \mathcal{M} : (8 \times 8) \) and rest are inter-pairwise relations between \( \mathcal{M} \) and candidates in \( \Lambda \). The intra-pairwise relations remain invariant across all six choices. Only the inter-pairwise score between \( \mathcal{M} \) and candidates in \( \Lambda \) play a key role in inferring the relationship. We verify by extracting the output of edge MLP, i.e., \( d \times 9 \) feature representations \( \mathbf{K}^{d} \), where \( d = 256 \) or 512, and computing the L2-norm on these feature vectors. The features of inter-pairwise relations given to MLP determine the correct answer. The above is verified on both the trained networks \( \text{Staged-WReN} \) and \( \text{E2E-WReN} \) over all the datasets. \( \text{DAReN} \) avoids forming redundant relations; instead, it works by matching the attributes to find the correct row that satisfies the rule in the top two rows. Our results on these datasets provide evidence of a stronger affinity between reasoning and disentanglement (discussed in the section below), which results from jointly learning both tasks.

In Figure 5, we present qualitative analysis of shapes3d and modified dsprites. The quality of the reconstructed images confirms that the learned distribution correctly captured the image content. In the right column, we present the distribution of per attribute accuracy over the hyperparameter sweep. The performance of object color in modified dsprites is sensitive to hyperparameter changes which is evident from large variance. We also see low performance for object shape due to its discrete nature; since our current latent representation is modeled towards real values, it fails to handle discrete representation.

**Disentangling.** In Table I, “Disentanglement”, we report the disentanglement scores of trained models on four widely used evaluation metrics namely, the Factor-VAE metric [30], DCI [36], MIG [35], and SAP-Score [37]. \( \text{DAReN} \) improves the reasoning performance and also strongly disentangles the latent vector, in contrast to both \( \text{Staged-WReN} \) and \( \text{E2E-WReN} \). One primary reason for significant improvements by \( \text{DAReN} \) compared to \( \text{Staged-WReN} \) and even \( \text{E2E-WReN} \) is due to the extraction of the \( K \) underlying generative factors and the averaging strategy over the least varying index in \( K \). Generally, in an unsupervised training process, weak signals from the true generative factors often leak into nuisance factors. However, \( \text{DAReN} \) avoids such infusion of nuisance factors by separating \( o \) from \( n \).

### Table I: Performance (mean ± variance) of Reasoning accuracy and four widely used benchmark Disentanglement Metrics on the six benchmark datasets. Note: higher score implies better result. The best score for each dataset among the competing models is shown in bold red and second-best in blue. (Note: * values are taken from [12].)

| Dataset | Model | Reasoning | Disentanglement | DCI | MIG | SAP |
|---------|-------|-----------|-----------------|-----|-----|-----|
| DSprites | Staged-WReN | 97.4 ± 4.2 | 74.4 ± 7.3 | 52.9 ± 10.5 | 28.7 ± 11.5 | 40 ± 14 |
|          | E2E-WReN  | 99.6 ± 1.5 | 77.6 ± 5.0 | 58.1 ± 8.0 | 38.2 ± 7.7 | 60 ± 2.2 |
|          | DAReN     | 99.3 ± 0.5 | 79.2 ± 6.2 | 59.0 ± 6.4 | 39.0 ± 0.0 | 60 ± 2.0 |

**Fig. 5:** Top: Shapes3D. Bottom: Mod DSprites. In the order of left to right, Reconstructions from \( \text{DAReN} \) (representative samples of median reconstruction error), odd columns show real samples and even columns their reconstruct. Distribution of reasoning performance per generative attribute over all 35 trained models. Expected prior KL-divergence for individual dimensions (left plot represents the best performing model, right plot represents the lowest performing model).
REFERENCES

[1] P. A. Carpenter, M. A. Just, and P. Shell, “What one intelligence test measures: a theoretical account of the processing in the raven progressive matrices test,” Psychological review, vol. 97, no. 3, p. 404, 1990.

[2] J. C. Raven, “Standardization of progressive matrices, 1938,” British Journal of Medical Psychology, 1941.

[3] J. C. Raven and J. H. Court, Raven’s progressive matrices and vocabulary scales. Oxford psychologists Press Oxford, England, 1998.

[4] B. Klein, J. Raven, and S. Fodor, “Scrambled adaptive matrices (sam)—a new test of eductive ability,” Psychological Test and Assessment Modeling, vol. 60, no. 4, p. 451, 2018.

[5] D. Hoshen and M. Werman, “Iq of neural networks,” arXiv preprint arXiv:1710.01692, 2017.

[6] D. R. Little, S. Lewandowsky, and T. L. Griffiths, “A bayesian model of rule induction in raven’s progressive matrices,” in Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 34, no. 34, 2012.

[7] A. Lovett and K. Forbus, “Modeling visual problem solving as analogical reasoning,” Psychological review, vol. 124, no. 1, p. 60, 2017.

[8] A. Lovett, K. Forbus, and J. Usher, “A structure-mapping model of raven’s progressive matrices,” in Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 32, no. 32, 2010.

[9] A. Lovett, E. Tomai, R. Gargiulo, and J. Usher, “Solving geometric analogy problems through two-stage analogical mapping,” Cognitive science, vol. 33, no. 7, pp. 1192–1231, 2009.

[10] S. Bringsjord and B. Schimanski, “What is artificial intelligence? psychometric ai as an answer,” in IJCAI. Citeseer, 2003, pp. 887–893.

[11] A. Lovett, K. Forbus, and J. Usher, “Analog with qualitative spatial representations can simulate solving raven’s progressive matrices,” in Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 29, no. 29, 2007.

[12] S. van Steenkiste, F. Locatello, J. Schmidhuber, and O. Bachem, “Are disentangled representations helpful for abstract visual reasoning?” in Advances in Neural Information Processing Systems, 2019, pp. 14245–14258.

[13] D. Barrett, F. Hill, A. Santoro, A. Morcos, and T. Lillicrap, “Measuring inductive biases for factorial representation learning,” in International Conference on Machine Learning, 2018, pp. 511–520.

[14] C. Zhang, F. Gao, B. Jia, Y. Zhu, and S.-C. Zhu, “Raven: A dataset for relational and analogical visual reasoning,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5317–5327.

[15] X. Steenbrugg, T. Verbelen, S. Leroux, and B. Dhodt, “Improving generalization for abstract reasoning tasks using disentangled feature representations,” in NeurIPS2018, part of 32nd Conference on Neural Information Processing Systems, 2018, pp. 1–8.

[16] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,” Neural computation, vol. 1, no. 4, pp. 541–551, 1989.

[17] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[18] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 8, pp. 1798–1828, 2013.

[19] K. Ridgeway, “A survey of inductive biases for factorial representation-learning,” arXiv preprint arXiv:1612.05299, 2016.

[20] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem, “Challenging common assumptions in the unsupervised learning of disentangled representations,” in Proceedings of the 36th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 4114–4124. [Online]. Available: https://proceedings.mlr.press/v97/locatello19a.html

[21] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, “Semi-supervised learning with deep generative models,” in Advances in Neural Information Processing Systems, 2014, pp. 3581–3589.

[22] D. Bouchacourt, R. Tomioka, and S. Nowozin, “Multi-level variational autoencoder: Learning disentangled representations from grouped observations,” in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[23] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, “Learning fine-grained image similarity with deep ranking,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1386–1393.

[24] K. McGreggor and A. Goel, “Confident reasoning on raven’s progressive matrices tests,” in Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.

[25] C. S. Mekik, R. Sun, and D. Y. Dai, “Similarity-based reasoning, raven’s matrices, and general intelligence,” in IJCAI, 2018, pp. 1576–1582.

[26] K. Wang and Z. Su, “Automatic generation of raven’s progressive matrices,” in Twenty-Fourth International Joint Conference on Artificial Intelligence, 2015.

[27] S. Hu, Y. Ma, X. Liu, Y. Wei, and S. Bai, “Hierarchical rule induction network for abstract visual reasoning,” arXiv preprint arXiv:2002.06838, vol. 2, no. 4, 2020.

[28] Y. Benny, N. Pekar, and L. Wolf, “Scale-localized abstract reasoning,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12557–12565.

[29] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner, “beta-vae: Learning basic visual concepts with a constrained variational framework,” in International Conference on Learning Representations, 2017.

[30] H. Kim and A. Mnih, “Disentangling by factorising,” in International Conference on Machine Learning, 2018, pp. 2649–2658.

[31] M. Kim, Y. Wang, P. Sahu, and V. Pavlovic, “Bayes-factor-vae: Hierarchical bayesian deep auto-encoder models for factor disentanglement,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2979–2987.

[32] M. Kim, P. Sahu, Y. Wang, and V. Pavlovic, “Relevance factor vae: Learning and identifying disentangled factors,” arXiv preprint arXiv:1902.01568, 2019.

[33] M. Tschannen, O. Bachem, and M. Lucic, “Recent advances in autoencoder-based representation learning,” arXiv preprint arXiv:1812.05609, 2018.

[34] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, “Understanding disentangling in beta-vae,” arXiv preprint arXiv:1804.05599, 2018.

[35] R. T. Q. Chen, X. Li, R. Grosse, and D. Duvenaud, “Isolating sources of disentanglement in variational autoencoders,” in Advances in Neural Information Processing Systems, 2018.

[36] C. Eastwood and C. K. Williams, “A framework for the quantitative evaluation of disentangled representations,” in International Conference on Learning Representations, 2018.

[37] A. Kumar, P. Sattigeri, and A. Balakrishnan, “Variational inference of disentangled latent concepts from unlabeled observations,” in International Conference on Learning Representations, 2018.

[38] F. Locatello, M. Tschannen, S. Bauer, G. Rätsch, B. Schölkopf, and O. Bachem, “Disentangling factors of variations using few labels,” in International Conference on Learning Representations, 2019.

[39] P. Sornson, C. Rother, and U. Köthe, “Disentanglement by nonlinear ica with general incompressible-flow networks (gm),” arXiv preprint arXiv:2001.04872, 2020.

[40] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen, “Variational autoencoders and nonlinear ica: A unifying framework,” in International Conference on Artificial Intelligence and Statistics. PMLR, 2020, pp. 6348–6359.

[41] H. Hosoya, “Group-based learning of disentangled representations with generalizability for novel contents,” arXiv preprint arXiv:1809.02383, 2018.

[42] R. Shu, Y. Chen, A. Kumar, S. Ermon, and B. Poole, “Weakly supervised disentanglement without compromises,” in International Conference on Machine Learning. PMLR, 2020, pp. 6348–6359.

[43] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” arXiv preprint arXiv:1312.6114, 2013.

[44] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from over-fitting,” The journal of machine learning research, vol. 15, no. 1, pp. 1929–1958, 2014.
[46] L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner, “dsprites: Disentanglement testing sprites dataset,” URL https://github.com/deepmind/dsprites-dataset/[Accessed on: 2018-05-08], 2017.

[47] M. W. Gondal, M. Wuthrich, D. Miladinovic, F. Locatello, M. Breidt, V. Volchkov, J. Akpo, O. Bachem, B. Schölkopf, and S. Bauer, “On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset,” Advances in Neural Information Processing Systems, vol. 32, pp. 15740–15751, 2019.