COMPUTER SCIENCE

Mental bootstrapping enables human-level concept learning in self-supervised deep models

Lingxiao Yang[1], Muyang Lyu[2], Ying-Jie Wang[3], Xiaohua Xie[4], Zonglei Zhen[5], Jian-Huang Lai[4], Ru-Yuan Zhang[6,7,8,9]*

How agents acquire abstract concepts from sparse, diverse examples—often without explicit supervision—remains a central problem in cognitive science and artificial intelligence. Human studies suggest that this ability depends on mental bootstrapping, the gradual construction of complex concepts from simpler partial structures. Building on this idea, we develop a self-supervised framework that trains models on systematically simplified versions of abstract reasoning tasks containing incomplete but structured concept cues. This algorithm enables models to form internal abstractions under limited resources and later apply them to more complex problems. We evaluate the framework across 12 abstract visual reasoning datasets testing in-distribution concept induction, out-of-distribution generalization, and few-shot learning. To contextualize performance, we also measure human accuracy on the same tasks. Models trained on simplified problems generalize robustly, reaching or even surpassing human-level performance. These findings show that abstract reasoning can emerge from structured simplification and minimal data, offering a computational account of concept learning in humans and machines.

INTRODUCTION

Building abstract knowledge and flexibly applying them in diverse real-life situations are the hallmarks of human intelligence (*1*, *2*). Although machine learning models (*3*–*7*) have considerably advanced many fields like computer vision (*8*), natural language processing (*3*), and speech recognition (*9*), it still remains a great challenge for most models to learn abstract concepts. A bulk of research in cognitive science has shown that humans have substantial adaptive learning and reasoning capabilities even without labeled data (*10*). For example, 9-month-old infants are capable of segmenting words from continuous speech streams through self-supervised statistical learning (*11*). In contrast, most existing machine learning models on concept learning focused on concrete concepts rather than abstract relationships as humans typically encounter in our daily life. Despite tremendous progress on self-supervised (*12*), semisupervised (*13*), and representation learning (*14*), most current models for concept learning still fall short as compared to how the human brain learns (*15*). Therefore, self-supervised learning of abstract relation concepts still remains an active research direction in both machine learning and cognitive science (*2*, *16*, *17*).

Humans are not only capable of learning abstract concepts without explicit labels, but they also do so with substantial efficiency. This capacity for efficient self-supervised learning is evident in the flexible generalization of learned knowledge to unseen situations (*18*–*20*). For example, one can readily repurpose a rock as a hammer to secure a tent, drawing on abstract knowledge that rigid objects afford pounding actions—an instance of generalized tool use. In such cases, humans rely on comparisons of abstract functional knowledge (e.g., "a rock and a hammer are both hard") rather than superficial visual features (e.g., "a rock and a hammer have different shapes") (*21*). This flexibility further manifests in the ability to leverage prior knowledge to facilitate the learning of new concepts, thereby circumventing the need to learn from scratch (*22*–*24*). A compelling example is that children can often learn new words or concepts after encountering only a few examples (*25*). While efficient human learning is well established, its implications for advancing machine learning models remain poorly understood.

How can we assess concept learning in humans and machines? This is a challenging question because the human brain is typically equipped with a vast amount of prior knowledge, which is strongly shaped by individual experiences and sociocultural factors (*26*, *27*). Across fields such as psychology, neuroscience, and psychiatry, Raven's Progressive Matrices (RPM) has long been a mainstream tool for measuring intelligence (*28*). Specifically, RPM (Fig. 1, A and B) is designed to evaluate abilities of concept induction, such as recognizing abstract relationships among visual objects and patterns. Similarly, in machine learning, Bongard problems (Fig. 1C) are widely regarded as important benchmarks for assessing concept learning (*29*). Crucially, both RPM and Bongard problems are less likely to be influenced by prior experience or sociocultural background, because they use symbolic visual stimuli rather than natural images or language-based interactions (e.g., Turing tests). As such, they serve as ideal and generic testbeds for evaluating intelligence in both humans and machines (*16*, *30*).

In concept learning, humans often extract more complex concepts by reusing previously learned simpler primitives, a process termed mental bootstrapping (MB) (*31*). Crucially, this occurs largely through self-supervised learning, where humans leverage internal feedback (Fig. 2A) and accumulated experience rather than

[1]School of Systems Science and Engineering, Sun Yat-sen University, Guangzhou 510006, Guangdong, People's Republic of China. [2]Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, People's Republic of China. [3]Brain Health Institute, National Center for Mental Disorders, Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine and School of Psychology, Shanghai 200030, People's Republic of China. [4]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, Guangdong, People's Republic of China. [5]Beijing Key Laboratory of Applied Experimental Psychology, Faculty of Psychology, Beijing Normal University, Beijing 100875, People's Republic of China. [6]School of Psychological and Cognitive Sciences, Peking University, Beijing 100871, People's Republic of China. [7]IDG/McGovern Institute for Brain Research, Peking University, Beijing 100871, People's Republic of China. [8]Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing 100871, People's Republic of China. [9]Beijing Key Laboratory of Brain-Computer Interface and Mental Health Modulation, Peking University, Beijing 100871, People's Republic of China.
*Corresponding author. Email: ruyuanzhang@gmail.com

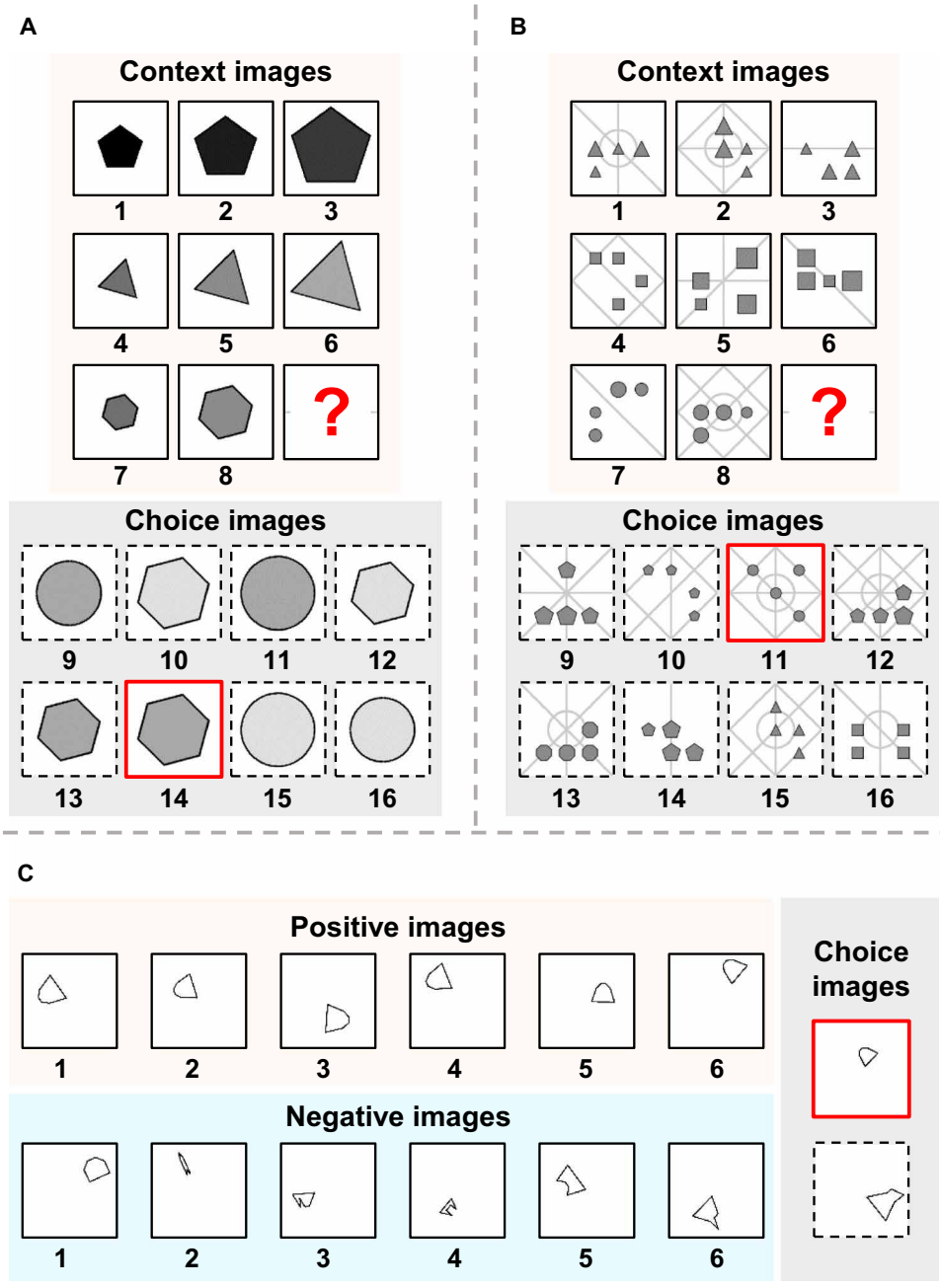


Fig. 1. RPM-like and Bongard problems. (**A** and **B**) Two example RPM-like problems from the RAVEN and PGM-Neutral datasets, respectively. In each RPM-like problem, eight context images and eight choice images are provided. The context images are presented as a 3 × 3 matrix, where each row or column contains abstract concepts to describe their relationships. An observer must identify these concepts and apply them to infer the missing cell in the lower right corner [denoted by a question mark (?)] by selecting the correct answer (highlighted by red boxes) from the eight choice images. The correct answer should make three rows or three columns follow a consistent set of abstract concepts. (**C**) Example of a Bongard problem. A Bongard problem consists of two distinct image sets: a positive set containing six images with similar concepts (i.e., "concave shape" in this example) and a negative set containing six images violating the concepts in the positive images. An observer must induce concepts from only a total of 12 positive and negative images and then classify whether two given choice images are positive (highlighted by red) or negative. Red boxes in all problems denote the correct or the positive answers. Note that, in our SSLvMB, the ground-truth labels, denoted using the red color, are not known.

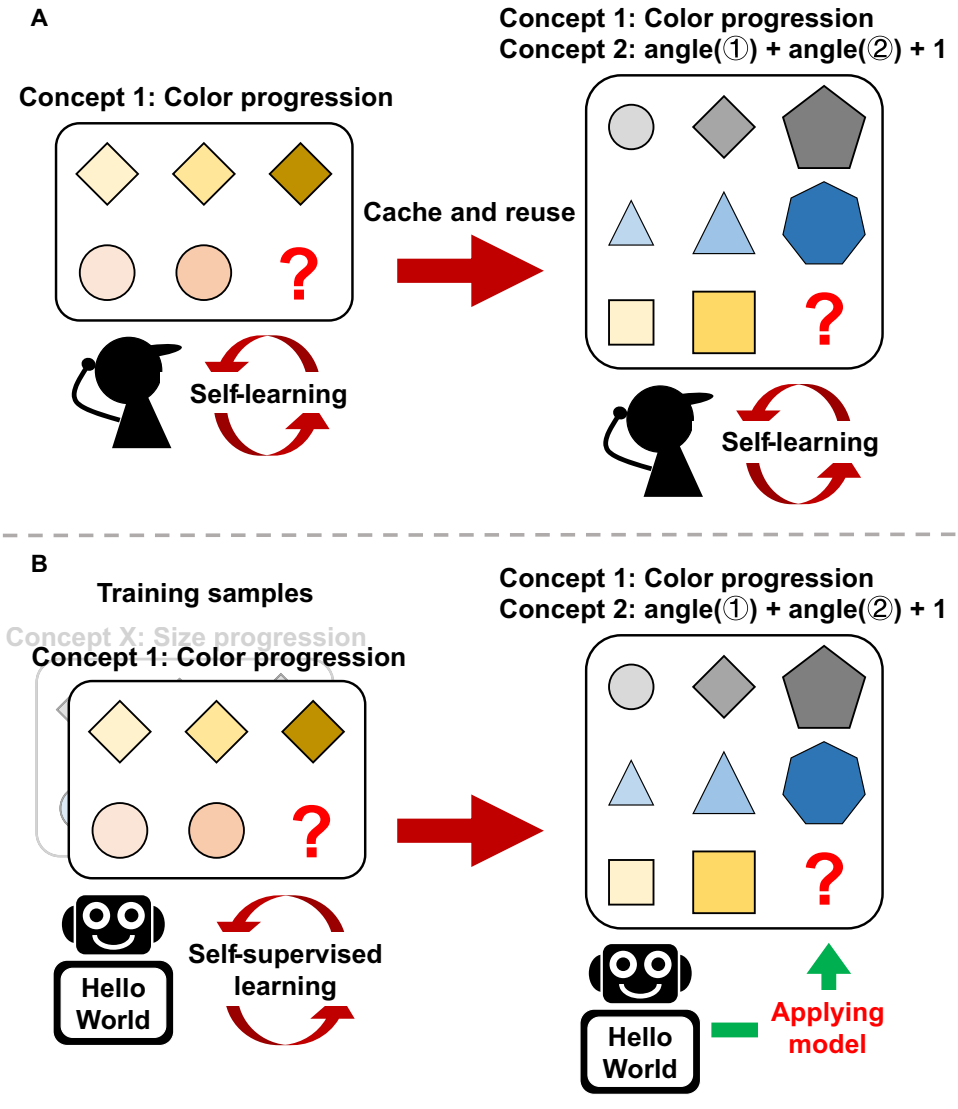


Fig. 2. Illustration of MB. (**A**) Core principles of MB demonstrated through two abstract concepts: color progression and angular arithmetic [defined as angle(①) + angle(②) + 1]. "angle(①)" ["angle(②)"] is the number of angles of the first (second) image in each row. Humans construct complex concepts by bootstrapping from primitive representations through autonomous self-learning cycles and memory caching/reuse without external supervision. (**B**) Main scheme of our proposed machine learning framework inspired by MB principles. Because of the challenges in applying MB to self-supervised machine learning—particularly the absence of explicit complexity indicators for concepts—we designed our MB-inspired algorithm with two distinct phases: a training phase [(B), left] using a reduced problem with fewer contexts of each RPM-like or Bongard problem and a testing phase evaluating on full contexts of problems. Apparently, the training phase contains simple concepts (e.g., size/color progression), while the testing phase has more complex concepts (e.g., composition of multiple abstract concepts such as color progression and angular arithmetic). As Halford's relational complexity theory shows (*37*), the increasing number and integration of context images in the right panels elevate the cognitive demand by requiring the processing of higher-arity relations and more complex feature bindings. While illustrated using RPM-like problems in this picture, our method follows the same design principle for Bongard problems—training on the reduced problem with fewer contexts that impose lower cognitive demands and testing on full contexts requiring greater cognitive load.

explicit external supervision. For example, young children who learn simple numbers (e.g., "one," "two," and "three") and the rule of incremental counting can form a more abstract concept of numbers, enabling them to learn an infinite number of integers (*31*). By reusing and composing previously learned knowledge, MB minimizes the resource demands of learning new concepts and enhances flexible generalization. Behavioral experiments have suggested MB as an essential strategy of human self-supervised concept learning (*31*, *32*) because learning primitive concepts constrain the hypothesis space and guide the subsequent learning of more complex concepts. However, MB requires the information of concept complexity as a priori knowledge, which is usually unavailable in self-supervised machine learning. It is still an open question how to leverage the human MB process to develop concept learning and inference algorithm.

In this study, we propose an effective computational framework—self-supervised learning via MB (SSLvMB)—for concept learning. SSLvMB embodies the human MB procedure as training and testing phases where the training phase includes relatively simpler concepts and the testing phase introduces more complex situations (Fig. 2B).

Notably, our model is able to automatically find abstract visual concepts without any labels. We tested our framework on three key tasks: in-distribution concept induction, out-of-distribution (OOD) concept generalization, and few-shot concept learning. We systematically measured human performance on identical tasks, and the head-to-head comparisons suggest that our model achieves performance comparable to or even surpassing humans. The human-level performance endowed by SSLvMB shed light on the development of more intelligence systems to learn abstract knowledge.

RESULTS

Tasks

To rigorously evaluate the effectiveness of our proposed SSLvMB, we conducted a comprehensive benchmark across 12 diverse datasets, covering two fundamental abstract visual concept learning tasks: RPM-like problems and Bongard problems (representative examples in Fig. 1).

An RPM-like problem (both Fig. 1, A and B) consists of eight context images and eight choice images. The eight context images are arranged into a 3 × 3 matrix, with the last cell missing. Observers should choose the correct image from the eight choice images to complete the matrix and ensure consistent abstract visual concepts (e.g., progression and constant) across rows or columns. Here, an abstract concept indicates the relationship between objects across rows or columns. For example, the abstract concept in Fig. 1A is that the images across rows have "constant color and shape but increasing size." The challenge in RPM-like problems lies in abstract reasoning and inductive inference, as observers must identify hidden concepts from only two complete rows or columns of context images in the matrix. The concepts can be vastly complex, involving multiple visual attributes like shape, color, and orientation and diverse rules such as constant, progression, union, and location. The same abstract concepts can be represented by different visual objects, and the same object can represent different abstract concepts in different problems. It is far more difficult than a visual recognition task. Following this logic, RAVEN, I-RAVEN, and RAVEN-FAIR are the mainstream datasets for testing in-distribution concept induction. Procedurally Generated Matrices (PGM) includes eight subdatasets for testing in-distribution concept induction and OOD concept generalization (see details in data descriptions in Materials and Methods).

The Bongard problem is a type of visual puzzle designed to test few-shot concept learning. Each problem consists of two sets of images: a positive set (six images), where all the images share similar concepts, and a negative set (six images), where the images contain concepts differing from the ones in the positive set. The task is to determine which set (positive or negative) a given image belongs to on the basis of the concepts induced from the two sets of context images. For example, the abstract concept in the positive set in Fig. 1C is "convex," and the negative set violates this concept. The concepts in Bongard problems can involve various attributes such as shape, color, orientation, or spatial relationships, and an observer must induce these concepts from merely 12 images and then classify the candidate images correctly.

SSLvMB framework

In concept learning under the MB theory, humans often build more complex concepts by reusing previously learned simpler primitives, thereby avoiding the need to learn new concepts entirely from scratch. Figure 2A depicts the core principles of human MB, whereby humans bootstrap their own learning—from simple concepts (left: color progression in each row embedded in five context images) to complex compositions (right: color progression combined with angular arithmetic embedded in eight context images). For example, the composite rule "angle(①) + angle(②) + 1" necessitates the integration of multiple visual features within a single row. More broadly, the composition of multiple concepts imposes greater cognitive demands as demonstrated by relational complexity theory (*32*), which posits that task difficulty scales with the number of concepts that must be simultaneously integrated to reach a compound concept solution.

While the concept of human MB is intuitively appealing, its application in self-supervised machine learning of abstract visual concepts presents two fundamental challenges. First, because of the lack of labeled data in the self-supervised fashion, it is not feasible to directly train models on concepts with different levels of complexity. Second, because of the progressive learning process—moving from simpler to more complex concepts (as in curriculum learning)—it would require access to concept-related information as ground truth. For example, RPM datasets provide meta-information associated with each problem, such as shape and abstract rules, which can be used as ground truth to benefit training. However, incorporating such meta-information introduces additional auxiliary signals, which changes the problem setting because such auxiliary data are typically unavailable when a human solves such a problem. As such, the successful construction of MB in self-supervised concept learning remains a vast challenge.

In this study, we operationalize the process of MB to develop a computational method for self-supervised concept learning. As machine models lack the innate self-learning abilities observed in humans, we explicitly structure the MB process into two distinct phases—training and testing—applicable to both RPM-like and Bongard problems. Figure 2B illustrates the core idea of this approach. The training phase (Fig. 2B, left) presents the learning process of machine models using a set of reduced problems, constructed with fewer contexts, while the testing phase (Fig. 2B, right) evaluates generalization across full problem sets. Although Fig. 2 illustrates the method using RPM-like problems, the same principle applies to Bongard problems: Training on reduced problems with fewer contexts, and testing requires full-context inference. Crucially, increasing the number of contexts elevates concept complexity, in line with Halford's framework (*33*). This design forces the model to construct robust feature representations during training, enabling it to solve more complex composite problems during testing.

This two-stage structure embodies core principles of human MB, enabling the compositional reuse of prior knowledge to support increasingly complex inferences. The progression from simplified to composite concepts parallels human cognitive development, where integrating more contextual information imposes greater cognitive demands (*33*).

Figure 3 provides a conceptual overview of the SSLvMB framework. The model first uses a Visual Perceptual Module (VPM) to extract features with different types of each visual reasoning problem. To mimic human MB, we then explicitly construct reduced problems (examples are shown in Fig. 4) that retain only partial context information. As shown in Fig. 4A, for an RPM-like problem, the original setting requires the model to complete a missing cell by inferring concepts across all three rows or columns. In the reduced formulation (Fig. 4B), only the first two rows or columns are used to establish concepts regularity. Note that the observer must infer a

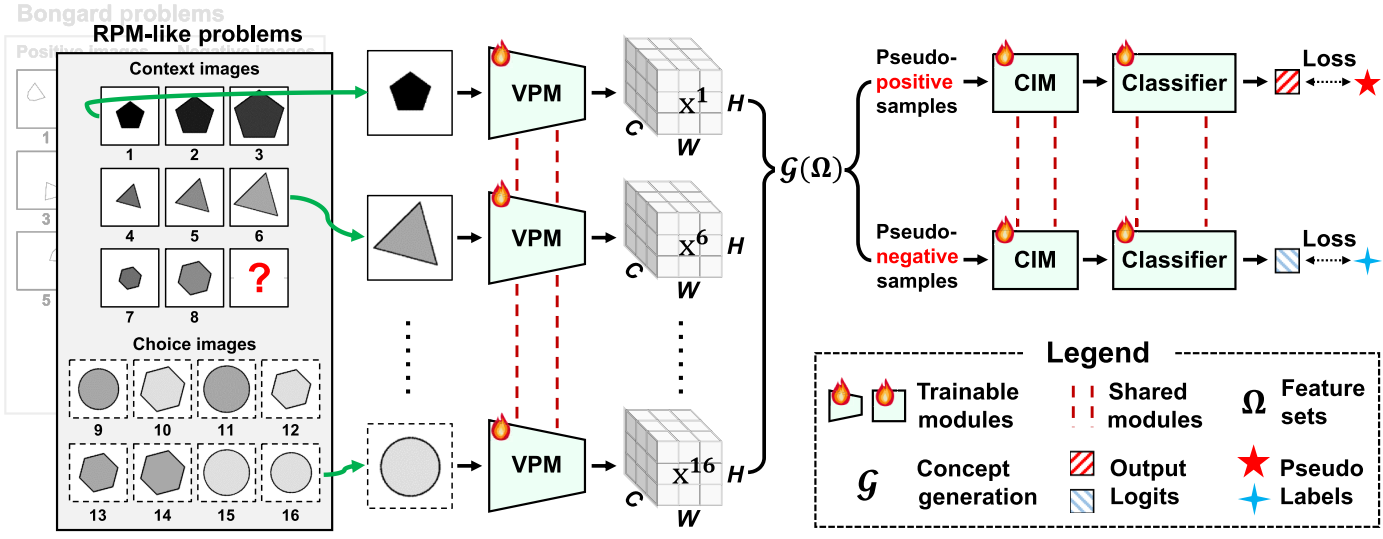


Fig. 3. Framework of our novel self-supervised learning method—SSLvMB. A VPM separately transforms each image to features in both RPM-like and Bongard problems. After that, we explicitly construct pseudopositive and pseudonegative samples [$\mathcal{G}(\Omega)$] for model training, where each sample contains multiple image features with reduced contexts. Furthermore, a CIM is used to extract abstract concepts within each constructed sample. In CIM, two convolutional layers are used to simultaneously extract concepts along the image axis (N, the number of images) and spatial cues along the spatial axis (L = the spatial height $H \times$ width W).

single underlying abstract relationship that is shared across all three rows or columns. If the concept is inferred using only the first two rows or columns, multiple candidate hypotheses typically remain consistent with the observations. Consequently, the third row or column is required to provide sufficient constraints to identify a unique solution. We also conducted an ablation study to remove the third row or column (i.e., context images 7 and 8) to demonstrate its constraint effects on rule abstraction (table S3). A similar approach is adopted for the Bongard problem (Fig. 4, C and D). Here, the original 12-context setting is reduced to 10 by applying dropout, and the omitted samples (e.g., $\mathbf{x}_p^4$ and $\mathbf{x}_n^2$ in this example) are repurposed as active test choices. As (*33*) demonstrated, the reduced problem requires lower cognitive demands as compared with the full problem. This configuration encourages the model to extract generalizable concepts under limited supervision, aligning with the MB principle of learning from fewer contexts.

Building on the reduced problem, we construct training samples using the task-defined sampling function—$\mathcal{G}(\Omega)$—where each sample comprises multiple image features. These samples are then processed by the Concept Induction Module (CIM; Fig. 5), which infers latent concepts by integrating information across the constituent features. A shared classifier subsequently maps each sample to a scalar logit, and training is performed using a contrastive loss that encourages discriminability between distinct training samples (Fig. 3). Details of our SSLvMB are presented in Materials and Methods.

SSLvMB enables high performance on in-distribution concept induction

We first tested our SSLvMB on four datasets for in-distribution concept induction: RAVEN (*34*), I-RAVEN (*35*), RAVEN-FAIR (*36*), and PGM-Neutral (*37*) (see Materials and Methods for descriptions of all datasets). All four datasets are RPM-like problems (Fig. 1, A and B). All four datasets have separate training and test sets. Concepts in the training and test sets in each dataset follow the same distribution (i.e., the same abstract concepts are included in both train and test sets). Therefore, the four datasets constitute a good test for the in-distribution concept induction.

To our best knowledge, Noisy Contrast and Decentralization (NCD) (*38*) and Pairwise Relations Discriminator (PRD) (*39*) are only two self-supervised (unsupervised) models on RPM-like problems. Thus, we compared SSLvMB with these two models (Table 1). Across all four datasets, NCD and PRD yielded similar averaged accuracy (47.2 and 45.6), whereas the average performance of our SSLvMB in all four datasets is 82.3. SSLvMB achieved more than 80% accuracy on RAVEN (83.8 ± 2.9) and I-RAVEN (89.7 ± 2.7) and even more than 90% on RAVEN-FAIR (91.3 ± 2.4). The improvement over NCD and PRD is substantial. On PGM-Neutral, which is more challenging, our SSLvMB achieved the accuracy of 64.4 ± 1.2, but both NCD (47.6) and PRD (34.8) obtain performance below 50%. Moreover, SSLvMB accomplishes this with only 1.27 million (M) parameters, which is an order of magnitude smaller than both NCD (11.24 M) and PRD (11.18 M). To assess the internal representations of abstract concepts in our SSLvMB, we used GradCAM (*40*) to visualize the abstract concepts (fig. S2) and found that SSLvMB can attend to different object features to extract abstract concepts. Furthermore, we trained a linear classifier on the layer after CIM using the 12 rule categories specified in the dataset's meta-information. It yielded 79.4% accuracy of decoding concepts. Together, these results demonstrate SSLvMB's superior capability of visual concept induction, consistently outperforming existing methods across diverse datasets.

We note that NCD (*38*) and PRD (*39*) also focus on constructing positive and negative samples but share several fundamental differences from ours. NCD, for example, selects the first two rows of an RPM-like problem as positive samples and fills the ninth position in the third row with eight choice images as negative samples. This strategy, however, carries a notable risk of mislabeling the correct answer as a negative sample. Similarly, PRD uses contrastive learning to align representations of the first two rows while pushing them away from the first or second rows of other RPM-like problems,

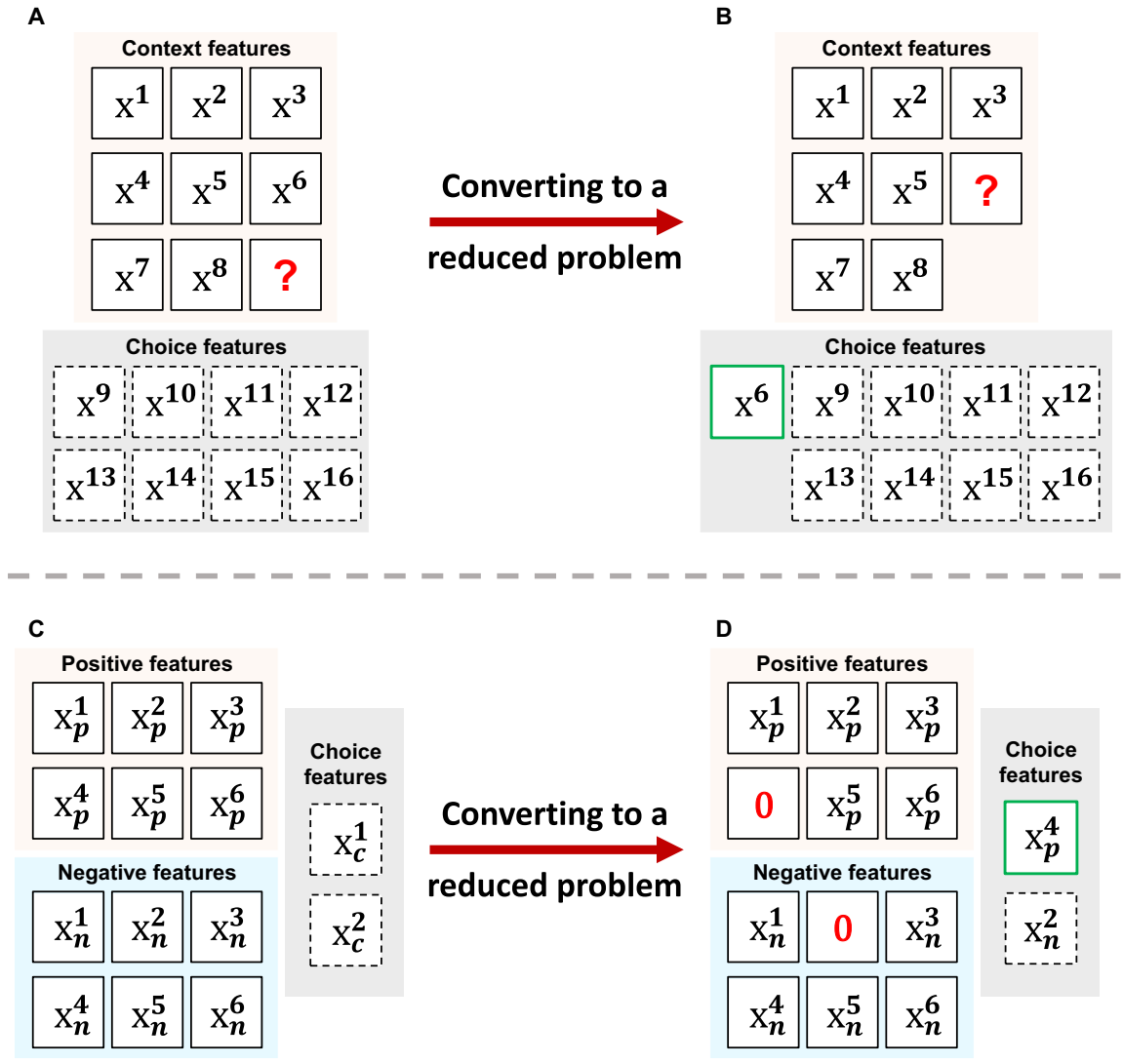


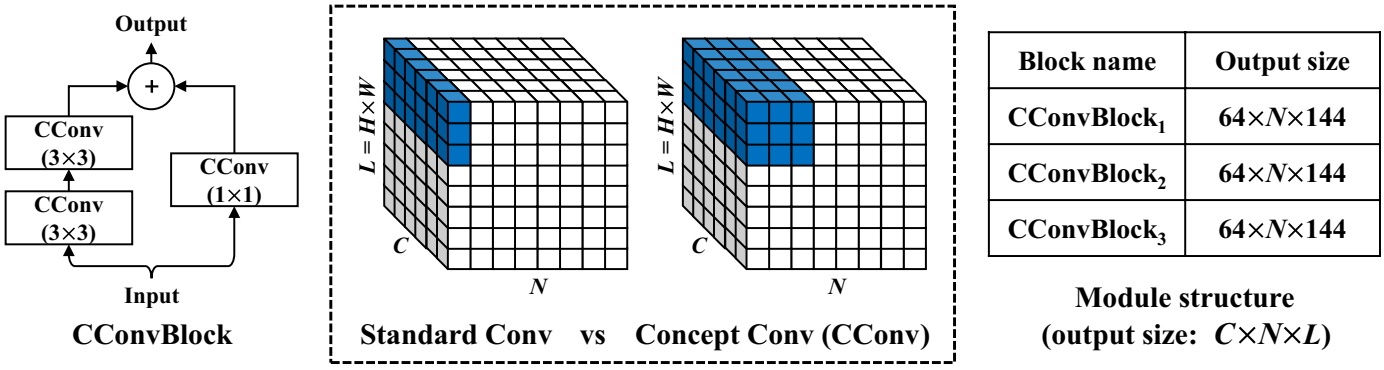
Fig. 4. Problem reduction strategies inspired by human MB. (**A** and **B**) In RPM-like problems, the original full setting (A) presents all eight context images, requiring the model to determine whether concepts are preserved across three rows or columns. To emulate the principle of human MB, the problem is reformulated into a reduced version (B) by randomly masking one of the context positions (e.g., the sixth location, rather than the standard bottom right). The masked feature is treated as a pseudo-correct choice, allowing the model to focus on assessing consistency between two rows or columns. (**C** and **D**) In Bongard problems, the full setting (C) consists of 12 context features divided into positive and negative sets, and the model is tasked with checking whether a choice feature aligns with or violates the underlying concepts in the positive or negative set. In the reduced version (D), one positive feature and one negative feature (e.g., $\mathbf{x}_p^4$ and $\mathbf{x}_n^2$) are randomly removed using dropout, and the omitted features are repurposed as pseudochoices, replacing the original choice features. This design mirrors the mechanism of human MB, wherein concepts are acquired from incomplete evidence, enabling the model to generalize under reduced contexts (10 instead of 12). Green boxes are pseudopositive labels.

assuming that randomly chosen problems do not share the same concept. In practice, this assumption does not always hold, leading to potential inclusion of incorrect negative samples. Moreover, both NCD and PRD rely solely on information from the first two rows (images 1 to 6) to construct pseudopositive samples, ignoring the third row as a constraint. Without this, the first two rows may correspond to multiple plausible concept solutions, increasing ambiguity and making concept learning and generalization across all three rows more difficult.

SSLvMB achieves the best average accuracy on compositional generalization

The aforementioned experiments primarily focus on in-distribution concept induction, wherein the training and test sets adhere to the same concept distribution. However, the key aspect of concept learning is to flexibly generalize learned concepts to novel, unlearned concepts, namely OOD concept generalization. The difficulty of concept learning lies at searching a correct answer from a large pool of candidate concepts (*41*). Moreover, the number of the complete set of concepts is enormous and infeasible to be learned in a single pass. This is particularly important when an agent learns concept primitives and reuses the primitives to compose a solution for a more complex problem. We argue that this reasoning capability reflects more accurately the inductive reasoning abilities of both humans and machine models (*18*, *42*).

We first tested the composition generalization of our SSLvMB on seven PGM subdatasets. The other seven PGM subdatasets besides



Block name	Output size
CConvBlock_1	$64\times N\times 144$
CConvBlock_2	$64\times N\times 144$
CConvBlock_3	$64\times N\times 144$

Module structure (output size: $C\times N\times L$)

Fig. 5. Overall architecture of CIM. Our CIM captures concept-level representations by jointly modeling spatial (along the *L* axis) and cross-image dependencies (along the *N* axis). It comprises stacked Concept Convolution Blocks (CConvBlocks), where each CConvBlock integrates cross-image and cross-spatial information through CConv layers. The module outputs features of size $C \times N \times L$, with residual connections facilitating hierarchical concept abstraction.

Table 1. In-distribution concept induction on RAVEN, I-RAVEN, RAVEN-FAIR, and PGM-Neutral. For all three RAVEN datasets, accuracy is calculated by averaging across all seven configurations. Results of other methods are either taken from their original publications or obtained using their publicly available codes. SSLvMB was run five times on all RAVENs and three times on PGM-Neutral because of its much larger size. These repetitions were run with different random seeds. The mean accuracy and standard deviation across multiple repetitions are reported in the table.

Datasets	Model size	RAVEN	I-RAVEN	RAVEN-FAIR	PGM-Neutral	Avg
NCD	11.24 M	37.0	48.2	56.0	47.6	47.2
PRD	11.18 M	37.9	55.9	54.5	34.8	45.6
SSLvMB	1.27 M	83.8 ± 2.9	89.7 ± 2.7	91.3 ± 2.4	64.4 ± 1.2	82.3

PGM-Neutral are designed specifically for testing OOD concept generalization. Their training and test sets include distinct concepts, i.e., different objects' attributes, different rules among objects, etc. These PGM subdatasets provide robust benchmarks for assessing whether a machine model is capable of inducing abstract concepts and generalizing them to novel situations.

We again compare our SSLvMB against NCD and PRD on all seven PGM subdatasets for OOD generalization (Table 2). SSLvMB is the best model in five of seven datasets. Overall, SSLvMB achieves the state-of-the-art average score of 30.9 across all seven subdatasets, surpassing NCD's 25.8 and PRD's 20.4, demonstrating its superiority in handling OOD generalization. In particular, in the case of Interpolation (Int; see methods for dataset descriptions), SSLvMB achieves an accuracy of 62.1 ± 3.1, much higher than NCD's 47.0 and PRD's 31.5. In the case of Extrapolation (Ext), SSLvMB's 15.6 ± 0.7 is not the highest but still competitive. Notably, in the Held-Out (H.O.) sets like H.O.TP, H.O.T, H.O.P, and H.O.SC, SSLvMB consistently performs well, often achieving the best or near-best results. These results provide strong evidence for the superior abilities of SSLvMB in OOD concept generalization. Moreover, we conducted an additional experiment on RAVEN-FAIR to validate the easy-to-hard learning mechanisms (text S2).

SSLvMB obtains high performance with few-shot concept learning

Besides flexible generalization, another hallmark of efficient concept learning is the ability to rapidly acquire and apply new concepts from a limited set of examples. This scenario closely parallels how humans learn from new environments, where they must quickly learn and infer new concepts from only a few unseen instances. It is widely recognized that humans have substantial few-shot learning capabilities (*43*)—the ability to generalize from limited examples—a feature notably absent in current machine learning models that predominantly rely on large datasets. In this section, we extend our analysis to evaluate the model's ability to use this sparse information for few-shot concept learning.

We used the Bongard-LOGO (*44*) dataset (an example Bongard problem is shown in Fig. 1C). The training and test sets in Bongard-LOGO include distinct objects and concepts. The context information of each problem consists of a positive set (i.e., six images) and a negative image set (i.e., six images). In addition, the test set of this dataset consists of four splits, i.e., Free-form (FF) shape, Basic (BA) shape, Combinatorial (CM) abstract shape, and Novel (NV) abstract shape. These four test splits are used to comprehensively evaluate different perspectives of few-shot learning in humans and machines (see Materials and Methods for data descriptions). Given that no self-supervised models have been proposed for this dataset, we compared our SSLvMB model against several state-of-the-art supervised models.

Table 3 shows all results. Astonishingly, our SSLvMB, as a self-supervised learning framework with no labels, outperforms other fully supervised models. In particular, SSLvMB achieves the best performance on the FF ($75.6_{\pm 0.3}$), BA ($93.5_{\pm 0.7}$), and CM ($72.7_{\pm 1.1}$) splits and the second best on the NV ($71.4_{\pm 0.6}$) split. SSLvMB obtains the overall average accuracy of 78.3, which is the highest among all models. These results are noteworthy considering that SSLvMB operates without access to ground-truth labels during training, whereas all competing methods require complete supervised information. The performance advantage highlights the effectiveness of our self-supervised approach in learning meaningful visual representations without manual annotation.

Table 2. OOD concept generalization on seven PGM subdatasets. The evaluation subdatasets are Interpolation (Int), Extrapolation (Ext), and Held-Out (shortened as H.O.) TriplePairs (TP)/Triples (T)/Pairs (P)/ShapeColor (SC)/LineType (LT). All the seven subdatasets have training and test sets. The training and test sets contain different concept distributions, i.e., some concepts occur in the test set but not in the training set. Results of other methods are either taken from their original publications or obtained using their publicly available codes. Our SSLvMB is run three times with different random seeds on all seven subdatasets. The mean accuracy and standard deviation across three repetitions are reported.

Datasets	Model size	Int	Ext	H.O.TP	H.O.T	H.O.P	H.O.LT	H.O.SC	Avg
NCD	11.24 M	47.0	24.9	33.1	13.2	19.0	29.5	13.8	25.8
PRD	11.18 M	31.5	12.7	28.1	18.6	14.1	25.2	12.8	20.4
SSLvMB	1.27 M	62.1 ± 3.1	15.6 ± 0.7	43.0 ± 0.9	28.6 ± 1.7	21.7 ± 2.1	32.4 ± 2.1	12.6 ± 0.1	30.9

Table 3. Few-shot concept learning on Bongard-LOGO. The test set of Bongard-LOGO dataset consists of four splits, i.e., FF shape, BA shape, CM abstract shape, and NV abstract shape. Machine models are trained on the training set and evaluated on these four test splits. As studied in (*44*), some objects and concepts occur in the test set but not in the training set. Moreover, because no self-supervised (unsupervised) models have been developed on Bongard-LOGO, we include several state-of-the-art supervised models for comparison. All models were run three times with different random seeds. The mean accuracy and standard deviation across the three repetitions are reported.

Methods	Model size	Learning type	FF	BA	CM	NV	Avg
Meta-Baseline-SC	8.10 M	Supervised	66.3 ± 0.6	73.3 ± 1.3	63.5 ± 0.3	63.9 ± 0.8	66.8
Meta-Baseline-MoCo	8.10 M	Supervised	65.9 ± 1.4	72.2 ± 0.8	63.9 ± 0.8	64.7 ± 0.3	66.7
ProtoNet	8.10 M	Supervised	64.6 ± 0.9	72.4 ± 0.8	62.4 ± 1.3	65.4 ± 1.2	66.3
PredRNet	1.27 M	Supervised	74.6 ± 0.3	75.2 ± 0.6	71.1 ± 1.5	68.4 ± 0.7	72.3
SVM-Mimic	9.89 M	Supervised	73.3 ± 0.3	84.3 ± 0.8	69.4 ± 0.8	74.2 ± 0.3	75.3
SSLvMB	1.27 M	Self-supervised	75.6 ± 0.3	93.5 ± 0.7	72.7 ± 1.1	71.4 ± 0.6	78.3

SSLvMB surpasses human performance in concept learning

The above results have demonstrated SSLvMB's efficacy in three aspects: in-distribution concept induction, OOD concept generalization, and few-shot concept learning, which are the three key abilities in concept learning that remain hallmarks of human intelligence. It is noteworthy that human concept learning is often influenced by specific knowledge backgrounds and unique experiences, and the underlying mechanisms remain difficult to test. Furthermore, existing research on concept learning has predominantly focused on the linguistic domain (*45*). An intriguing question arises: To what extent can humans learn, induce, and generalize novel visual concepts? This question holds particular significance in the context of RPM-like and Bongard reasoning, because both of them were originally designed for assessing human or machine intelligence and have served as a standard testing tool in social sciences, psychology, and neuroscience for nearly a century (*28*, *29*, *46*). To conduct a comprehensive head-to-head comparison between human and machine performance, we performed behavioral experiments to collect human performance on exactly the same tasks and stimuli (see Materials and Methods for experimental details).

Figure 6 presents comparisons between the performance of SSLvMB and humans across various tasks. The evaluated datasets include RAVEN, I-RAVEN, RAVEN-F (RAVEN-FAIR), and PGM-N (PGM-Neutral) for in-distribution concept induction; PGM-E (PGM-Extrapolation) and PGM-I (PGM-Interpolation) for OOD concept generalization; and the four test splits of Bongard-LOGO (BP-FF, BP-BA, BP-CM, and BP-NV) for few-shot concept learning. Human performance on RAVEN, BP-FF, BP-BA, BP-CM, and BP-NV are directly taken from their original papers (*34*, *44*). Human performance on other datasets was measured in this study.

We found that SSLvMB can achieve or even outperform human performance in all scenarios except the FF split of the Bongard-LOGO (Fig. 6). Notably, the model demonstrates a notable advantage in tasks such as I-RAVEN, RAVEN-F, PGM-N, and PGM-I, where its accuracy surpasses human performance by a substantial margin. SSLvMB's mean accuracy computed across all tasks achieves an average accuracy of 72% compared to the mean accuracy of 54% in humans. These results strongly suggest that SSLvMB not only matches but also exceeds human capabilities in some concept learning tasks, highlighting its potential applications in artificial intelligence and cognitive computing. We also conducted additional experiments to let human subjects learn examples and perform the task to investigate the effect of explicitly learning on AVR problems, although such learning improved human performance a little bit but still fall far short of model performance. We included these learning experiments in fig. S1.

DISCUSSION

Induction, generalization, and few-shot learning of abstract concepts have long been recognized as key strengths of human intelligence as well as substantial challenges for current machine learning models. To address these challenges, we propose a self-supervised framework designed for RPM-like and Bongard visual reasoning tasks. Inspired by humans' MB capability, where simpler concepts are reused to construct more complex concepts, our model is trained

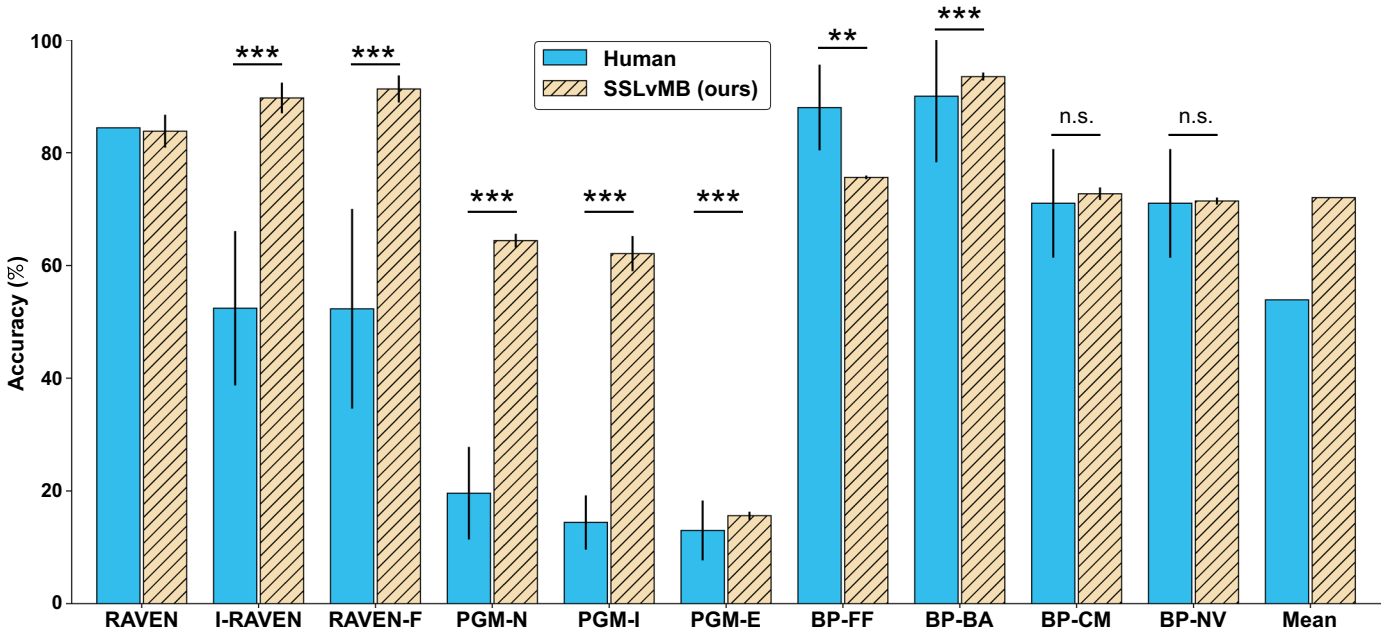


Fig. 6. Comparisons of humans and SSLvMB. Experiments include in-distribution concept induction on RAVEN, I-RAVEN, RAVEN-F, and PGM-Neutral, OOD concept generalization on PGM-Interpolation (PGM-I) and PGM-Extrapolation (PGM-E), and few-shot concept learning on four test splits of Bongard-LOGO (BP-FF, BP-BA, BP-CM, and BP-NV). Human performance on RAVEN and all test splits of Bongard-LOGO is directly obtained from their original papers (*34*, *44*), respectively. Human performance on all other datasets, including I-RAVEN ($N = 25$), RAVEN-FAIR ($N = 25$), PGM-N ($N = 30$), PGM-I ($N = 30$), and PGM-E ($N = 30$) were measured in this study. The bar chart illustrates the accuracy for both human observers (blue bars) and our SSLvMB model (striped bars). The error bars of human results represent the standard deviation across subjects. The error bars of model performance represent the standard deviation across three repetitions with random seeds. The significance conventions are as follows: $^{**}P < 0.01$; $^{***}P < 0.001$; n.s., nonsignificant. See statistical details in Materials and Methods.

on simpler knowledge and then uses learned knowledge to solve more difficult problems during inference. All learning is self-supervised without any label. Our approach has demonstrated strong performance on in-distribution concept induction, OOD concept generalization, and few-shot learning. Our model not only surpasses supervised models in some scenarios but also achieves or surpasses human performance on the identical tasks and stimuli. These results challenge the long-standing view of human superiority in these cognitive areas.

Human concept learning via MB

A bulk of studies in cognitive science has suggested that MB is an essential concept learning strategy of humans (*31*, *32*). For instance, after human observers first acquire simpler, basic primitive concepts, subsequent learning of more complex concepts can be much faster if the complex concepts consist of previous primitives. Conversely, if human observers are first exposed to more complex concepts, they must search a much larger hypothesis space to induce the complex concepts, making the solution computationally intractable given the limited cognitive resources. Thus, learning primitive concepts first helps guide and constrain the hypothesis space for learning more complex combinations. This simple-to-complex learning process has also been adopted in machine learning such as supervised curriculum learning (*47*). In supervised curriculum learning, the model is trained by gradually using samples from easy to hard. In contrast, in our SSLvMB, the model is trained on simple primitives but evaluated on more complex situations.

We emphasize that MB in our model is a training paradigm rather than a computational module. To further highlight the efficacy of our MB-based training methods, we used the exactly same model architecture (i.e., VPM + CIM) but trained it using competitor methods (i.e., NCD and PRD). We found that using different training methods substantially degrades model performance (table S4), providing strong evidence for our MB-based training approach. Moreover, MB here is a concept from cognitive science and highlights that humans can solve complex problems by reusing (i.e., bootstrapping) existing knowledge. This is different from statistical bootstrapping implemented by resampling data. MB is unlikely to be constrained by the number of features in a model.

Unique advantages of our SSLvMB

Our self-supervised framework for abstract visual concept learning is of particular value in this line of research. Most of existing models on the abstract concept learning tasks like RPM and Bongard problems are supervised models. However, we emphasize that the key aspect of these tasks is to test whether a machine model can achieve human-level abilities in concept learning, especially in self-supervised learning.

Our self-supervised learning method is inspired by human MB, which involves training on simpler cases during the learning phase and then reasoning about more complex cases during inferences. In particular, in RPM-like problems, the learning process focuses on the consistency of concepts in partial context images, eventually extending this consistency to the inferences between two different variants of the same problem. For Bongard problems, training considers the consistency of six images (five context images and one choice) that form a concept, while testing involves inferences with seven images to assess the consistency of the concept. This easy-to-hard concept learning and inference are aligned with the human MB

process (*31*, *32*). In addition, in the construction of positive and negative samples, we only use one RPM or Bongard problem, excluding the potential confounding of introducing wrong concepts from other problems as in NCD and PRD. Moreover, our SSLvMB framework has been tested on three key aspects of concept learning: in-distribution concept induction, OOD concept generalization, and few-shot concept learning, covering a broad range of datasets and testing scenarios.

We emphasize that several benchmark datasets, including the seven OOD subsets of PGM and the four splits of Bongard-LOGO, are specifically designed to probe compositional generalization—the capacity to extrapolate from primitives to more complex or novel concept compositions. For instance, the BA shape test set in Bongard-LOGO directly targets this ability: It contains 480 problems where each concept corresponds to a new combination of two basic shapes unseen during training, providing a controlled evaluation of generalization from simpler to more complex concepts.

Differences between humans and machines in concept induction and reasoning

Our behavioral experiments also provide evidence for a head-to-head comparison between human and machine performance in abstract visual concept learning. A key gap in existing studies is that human performance on these tasks is largely unknown. A common assumption is that humans excel at these flexible reasoning tasks, yet this assumption has rarely been systematically tested. For instance, Nie *et al.* (*44*) tested three novices on Bongard-LOGO and reported 83% accuracy. No human performance has been reported on I-RAVEN, RAVEN-FAIR, and all PGM subdatasets. Unexpectedly, we found that human performance was suboptimal across five datasets: I-RAVEN, RAVEN-FAIR, PGM-Neutral, PGM-Interpolation, and PGM-Extrapolation. Notably, we found that human performance in the OOD generalization scenario on the PGM subdatasets was unexpectedly low. This suggests that the OOD problems in the PGM may be inherently more difficult than those in-distribution tests.

Our method leverages the human-like mechanism of MB: learning easy primitives and scaling to complex problems. Our intention is not to replicate every aspect of human reasoning, and the difficulty control process is an engineering realization. By masking out a context and reinserting it as a candidate (or, in Bongard problems, by dropping one example from the positive and one from the negative set), we generate pseudopositive and pseudonegative pairs that convert the original full reasoning task into simpler contrastive comparison subtasks. This mechanism is supported by cognitive studies showing that humans rely on comparisons between positive and negative exemplars to solve the problem (*48*). However, it remains an open question whether humans can solve complex problems by dividing into simpler subquestions. Future studies may design more rigorous experiments to test this hypothesis.

Induction of abstract concepts is in general difficult in machine learning. Here, we used CIM to extract abstract concepts. The convolution-based CIM is particularly useful to compute local contingency between context images (e.g., for an RPM problem, images 1/2/3 for the first row, and images 4/5/6 for the second row). We conducted a control experiment where the convolution operation was replaced by a transformer-like architecture (text S1 and table S1). The results show suboptimal performance, supporting the efficacy of simple convolution in concept induction. We speculate that given the self-supervised setting and limited context information (e.g., only eight context images in an RPM problem), convolution is a more powerful inductive bias. We also acknowledge that the transformer-like operation may be better suited for other forms of tasks, such as sequence learning or temporal reasoning. We emphasize that our main focus is MB-based training, not the architecture of CIM. The format of CIM can be flexibly adjusted according to the task.

The results from human participants challenge the traditional view that humans have extraordinary abilities of concept learning. At the very least, the performance of our model reaches the level of human performance on these particular tasks. There are three potential reasons for why human performance is suboptimal. First, humans excel at perceiving and recognizing visual attributes but may suffer limited resources to search the correct concept from a large number of candidate rules. For example, it has been shown that humans can only track or construct relationships among a limited set of targets (*33*, *49*). This reflects the inherent limitations of human cognitive resources. In contrast, SSLvMB does not impose computational constraints on searching candidate abstract rules when solving a specific abstract visual reasoning (AVR) problem. We speculate that this is the key part that is not human-like. We also emphasize that SSLvMB may not have constraints when solving a AVR problem, but it does face computational constraints when learning new concepts. Otherwise, SSLvMB can learn any new concepts from scratch, which is in practice infeasible. Second, our approach extracts concepts between visual features through the CIM. Given the large number of learnable parameters, the CIM may capture more complex intertarget relationships, thus leading to better performance in these concept learning tasks. Third, humans may not strictly follow the simple-to-complex progressive learning in this scenario. As a result, humans may exhibit greater uncertainty when facing complex conjunction of concepts, leading to poorer outcomes. Future studies may ask human observers to learn simpler concepts and then advance to more complex concept conjunctions, and the learning effects might be more pronounced. We also emphasize that our human experiments here are just a proof of concept that humans and machine models can be evaluated on the same testbed. However, we noted that the exact computational mechanisms through which humans solve the problems still remain unclear. A process model is needed in cognitive science to quantify human behavior. However, such a model may require an enormous amount of human behavioral data. Only with large-scale datasets can we characterize the general behavioral patterns of human problem-solving.

MB as a general approach for machine learning

More generally, MB offers a domain-agnostic principle for designing more effective machine learning algorithms: Learning can be made more efficient and generalizable by organizing experience around progressively structured, lower-complexity abstractions that support the reuse of prior knowledge. Beyond abstract visual reasoning, MB-style learning naturally extends to domains such as language, reinforcement learning, and multimodal learning, where agents must infer a latent structure from sparse feedback. For example, in natural language processing, models may benefit from first acquiring partial syntactic or semantic regularities before being exposed to full linguistic complexity, mirroring human language acquisition (*2*, *47*). In reinforcement learning, MB aligns with approaches that decompose long-horizon problems into simpler subgoals or skills, enabling agents to bootstrap complex policies from learned

primitives rather than learning end to end from sparse rewards (*50*, *51*). Similarly, in multimodal and representation learning, MB suggests training models on simplified cross-modal correspondences to form stable abstract representations that can later support transfer and few-shot learning (*52*–*54*). Across these domains, MB emphasizes the importance of structuring self-supervised experience by conceptual complexity, providing a unifying framework for improving data efficiency, OOD generalization, and lifelong learning in artificial systems (*17*, *55*).

Limitations and future directions

While SSLvMB demonstrates promising results in abstract visual concept learning, several limitations should be acknowledged. First, SSLvMB trains the model using simpler concepts during the training phase but makes inferences on more difficult scenarios during the testing phase. A better approach might be to directly train a model from simpler to difficult concepts during the training phase. However, this requires the complexity information of each problem as a priori. Second, SSLvMB's performance, although superior to existing unsupervised methods, still falls short of many state-of-the-art supervised (*30*, *56*–*58*) methods on concept learning tasks. This suggests that the current architecture may not fully capture the compositional and hierarchical nature of human concept learning. Third, SSLvMB is evaluated on abstract visual concept datasets. It is unclear how such a model could extend to real-world scenarios, such as visual question answering (*59*) or interactive reasoning (*60*). Future models could incorporate the principle of MB to solve real-word problems, potentially improving its robustness in diverse concept induction and generalization scenarios.

MATERIALS AND METHODS

Ethics and human participants

All experimental protocols were approved by the institutional review board of Shanghai Jiao Tong University (H20240606C) and Beijing Normal University (ICBIR_A_0091_012). All research was conducted in accordance with relevant guidelines and regulations. Informed written consent was obtained from all participants (see demographical information in table S2).

All datasets

Experiments were conducted on several widely used datasets for abstract visual concept learning: three RAVEN datasets (*34*–*36*), the PGM dataset (*37*), and the Bongard-LOGO dataset (*44*).

RAVEN/I-RAVEN/RAVEN-FAIR

The three RAVEN datasets are designed to emulate the structure of RPM (*46*), a well-known tool for assessing human Intelligence Quotient. Introduced in 2019 (*34*), the RAVEN dataset includes four concepts—"Progression," "Constant," "Union," and "Arithmetic Calculations"—to describe the relations between objects. RAVEN also contains seven spatial configurations—"Center,""2x2Grid," "3x3Grid," "Left-Right," "Up-Down," "Out-InCenter," and "Out-InGrid". All models are trained on all seven spatial configurations of all datasets. Each configuration contains 10,000 problems, and each problem includes eight context images and eight choice images. There are a total of 70,000 problems and 1,120,000 images. Some studies (*35*, *36*) suggested that machine models trained only on the eight choice images (without the eight context images) on the RAVEN dataset can achieve unexpectedly high performance, revealing the potential bias of the RAVEN dataset. I-RAVEN (*35*) and RAVEN-FAIR (*36*) were constructed to eliminate the bias. Both datasets keep all context images from RAVEN but introduce different strategies to generate negative choice images. Experiments demonstrate that models trained exclusively on choice images show chance-level performance on I-RAVEN and RAVEN-FAIR. I-RAVEN and RAVEN-FAIR are thus more robust benchmarks for evaluating concept learning.

Note that these three RAVEN datasets provide not only the correct label for each problem but also several types of ground-truth meta-information (e.g., shape attributes and abstract rules) associated with each item. Our study did not use any meta-information.

PGM

PGM is also an RPM-like dataset. PGM consists of eight subdatasets, each containing 1,222,000 problems. Similar to RAVEN, each problem in PGM includes eight context images and eight choice images. Each problem is governed by abstract visual concepts involving relationships such as "XOR," "OR," "Progression," and "AND," among objects. In the PGM-Neutral subdataset, the concepts in training and test sets share the same distribution. In contrast, the concepts in the training and tests in other seven subdatasets belong to different distributions. The other seven subdatasets are thus aimed for OOD concept generalization. OOD concept generalization is the hallmark of human inductive inferences. Similar to the RAVEN datasets, PGM also provides meta-information of each problem, which was not used in our study.

Bongard-LOGO

The Bongard-LOGO dataset (*44*) is designed to evaluate few-shot concept learning. It is inspired by the classical Bongard problems (*29*), where observers are tasked with inferring visual concepts from just a few context images. The Bongard-LOGO dataset consists of 12,000 binary classification problems. Each problem includes six positive images that share a common concept and six negative images that do not. Observers must first infer the underlying concept by comparing the positive and negative context images. Learning is limited to these 12 context images. After this learning phase, the observer is asked to determine whether two test images share the same concept as the positive images. Some objects and concepts in the test set do not appear in the training set, making generalization more challenging. Concept learning in Bongard-LOGO is considered more difficult than in RPM-like tasks, as models must handle novel objects and abstract relationships using only limited visual information.

Model architecture

Visual Perceptual Module (VPM)

We modified the image encoder in (*56*) as our VPM for feature extraction. Specifically, VPM comprises four residual blocks. Each block consists of two branches: a residual branch and a shortcut branch. The residual branch consists of three convolutional layers with kernel sizes of 3×3. The shortcut branch first uses an average pooling layer to process input features and then uses a 1×1 convolutional layer to match the output dimension of the residual branch. The output of the shortcut branch is then added to the residual branch for the next block. In addition, in the first three blocks, we downsample the input features with a stride of 2 to enlarge the receptive fields. This structure facilitates neurons to extract higher-level information. After four blocks, we use another 1×1 convolutional layer to reduce the feature dimension for further processing. In summary, VPM can be formulated as

$$\mathbf{x}_l^i = \mathrm{ResBlock}_{s=2}^{l}\left(\mathbf{x}_{l-1}^i\right),\ \ l \in \{1,2,3\} \tag{1}$$

$$\mathbf{x}_4^i = \mathrm{ResBlock}_{s=1}^{4}\left(\mathbf{x}_3^i\right) \tag{2}$$

$$\mathbf{x}^i = \mathrm{Conv}_5\left(\mathbf{x}_4^i\right) \tag{3}$$

where $\mathbf{x}_0^i \in R^{1\times 96\times 96}$ is the image i in a problem. $\mathbf{x}_l^i \in R^{C\times H\times W}$ is the lth block's features of the image, where C, H, and W indicate feature channels, spatial height, and width, respectively (Fig. 3). s is the stride used in each ResBlock. Given an input image with a $1 \times 96 \times 96$ size, VPM transforms it into an embedding with a size of $\mathbf{x}^i \in R^{64\times 12\times 12}$. VPM separately extracts features from images (i.e., eight context images and eight choice images in an RPM-like problem) without considering any concepts between images.

Constructing pseudo training samples [$\mathcal{G}(\Omega)$] on the basis of the reduced problems

To mimic human MB, we propose an approach in SSLvMB by transforming RPM-like and Bongard problems into their reduced versions. These reduced formulations present fewer visual contexts for concept discovery while retaining the compositional structure necessary for abstraction (see Fig. 4, A to D). On the basis of these reduced problems, we define a pseudo sample construction function $\boldsymbol{\mathcal{G}}(\Omega)$. Let $\boldsymbol{\mathcal{X}}$ denote the set of features for all images in a single problem obtained by our VPM (for example, for an RPM-like problem, $\boldsymbol{\mathcal{X}}$ comprises eight context and eight choice features, while for a Bongard problem, $\boldsymbol{\mathcal{X}}$ comprises 14 images' features).

We then let $\Omega \subseteq \boldsymbol{\mathcal{X}}$ denote a subset of these features. Define $\boldsymbol{\mathcal{G}}$: $\Omega \to \boldsymbol{\mathcal{S}} \times \boldsymbol{\mathcal{Y}}$ be a function that maps input features to a set of pseudo training samples $\mathbf{s} \in \boldsymbol{\mathcal{S}}$ with binary labels $y \in \boldsymbol{\mathcal{Y}} = \{\mathbf{0}, \mathbf{1}\}$. After that, each sample $\mathbf{s}$ comprises multiple image features, serving as either pseudopositive or pseudonegative training signals. Given the structural differences between RPM-like and Bongard problems, we specialize $\mathcal{G}(\Omega)$ accordingly and present its details below.

RPM-like problems

Let the full context features be $\Omega_c = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^8\}$, $\mathbf{x}^i \in R^{C\times H\times W}$, and let the choice features be:$\Omega_a = \{\mathbf{x}^9, \mathbf{x}^{10}, \ldots, \mathbf{x}^{16}\}$. We define a reduced problem by masking out a randomly selected context feature $\mathbf{x}_{\mathrm{mask}}^k = \mathbf{x}^k$, where k is a random integer sampled from the range [1, 8]. The remaining context features can be represented as $\Omega_c^k = \Omega_c \setminus \{\mathbf{x}^k\}$. After that, the pseudopositive sample is defined by filling the missing cell [denoted using a question mark (?) in Fig. 4B] with a current masked feature

$$\mathbf{s}^+ = \Omega_c^k \cup_? \{\mathbf{x}^k\}, \quad y = 1 \tag{4}$$

where $\cup_?$ denotes placing the right-hand element into the placeholder "?." The pseudonegative samples are constructed by

$$\mathbf{s}^j = \Omega_c^k \cup_? \{\mathbf{x}^j\}, \quad y^j = 0, \mathbf{x}^j \in \Omega_a \tag{5}$$

To further increase the variability of pseudonegative samples, we apply a sampling strategy directly to the original context set Ω_c and define the resulting pseudonegative sample as

$$\mathbf{s}^{sp} = \mathrm{Sample}\left(\Omega_c, 8\right), \quad y^{sp} = 0 \tag{6}$$

where $\mathrm{Sample}\left(\Omega_c, 8\right)$ denotes a random sampling procedure without replacement, producing a reordered or partially duplicated set of eight context features. With the above definition, each type of $\mathbf{s}$ contains multiple image features with a size of $8 \times C \times H \times W$. Moreover, the pseudopositive sample $\mathbf{s}^+$ contains consistency concepts in the first two rows (or columns), and its final row (or column) serves as a constraining context that helps disambiguate competing hypotheses, particularly when multiple concept candidates remain consistent with the reduced problem. The pseudonegative samples $\mathbf{s}_j$ and $\mathbf{s}^{sp}$ disrupt concepts embedded in first two rows (or columns) and therefore are treated as negative samples. As a result, the final pseudo sample construction function $\boldsymbol{\mathcal{G}}(\Omega)$ for RPM-like problems is defined as

$$\mathcal{G}_{\mathrm{RPM}}\left(\Omega_c, \Omega_a\right) = \{\left(\mathbf{s}^+, y\right)\} \cup \{\left(\mathbf{s}^j, y^j\right), j = 9, 10, \ldots, 16\} \cup \{\left(\mathbf{s}^{sp}, y^{sp}\right)\} \tag{7}$$

Bongard problems

Let the positive and negative features be $\Omega_p = \{\mathbf{x}_p^1, \mathbf{x}_p^2, \ldots, \mathbf{x}_p^6\}$ and $\Omega_n = \{\mathbf{x}_n^1, \mathbf{x}_n^2, \ldots, \mathbf{x}_n^6\}$, respectively. As shown in Fig. 4 (C and D), we apply dropout-like operation to each feature set, in which one feature is randomly selected and dropped from each set. Specifically, the dropped features are denoted as $\mathbf{x}_p^k$ and $\mathbf{x}_n^z$, where k and z are integers randomly sampled from {1, 2, …, 6}. This procedure effectively converts the original 12-context problems to 10-context problems. In accordance with the standard dropout formulation, the remaining features are rescaled by a factor of 6/5 to preserve the expected activation magnitude. The positive and negative feature sets in the resulting reduced problem are defined as follows

$$\tilde{\Omega}_p^k = \begin{cases} \frac{6}{5}\mathbf{x}_p^i, & i \neq k \\ 0, & i = k \end{cases}, \tilde{\Omega}_n^z = \begin{cases} \frac{6}{5}\mathbf{x}_n^i, & j \neq z \\ 0, & j = z \end{cases} \tag{8}$$

The pseudopositive and pseudonegative training samples in the reduced Bongard setting are constructed as follows

$$\begin{aligned} \mathbf{s}^1 &= \tilde{\Omega}_p^k \cup \mathbf{x}_p^k, \quad y^1 = 1 \\ \mathbf{s}^2 &= \tilde{\Omega}_n^z \cup \mathbf{x}_n^z, \quad y^2 = 1 \\ \mathbf{s}^3 &= \tilde{\Omega}_p^k \cup \mathbf{x}_n^z, \quad y^3 = 0 \\ \mathbf{s}^4 &= \tilde{\Omega}_n^z \cup \mathbf{x}_p^k, \quad y^4 = 0 \end{aligned} \tag{9}$$

where $\mathbf{s}^i \in R^{6\times C\times H\times W}$ contains five context features and a query feature. The relationship can be extracted within each $\mathbf{s}^i$. As a result, the final pseudo sample construction function $\boldsymbol{\mathcal{G}}(\Omega)$ for Bongard problems is defined as

$$\mathcal{G}_{\mathrm{Bongard}}\left(\Omega_p, \Omega_n\right) = \{\left(\mathbf{s}^i, y^i\right)\}_{i=1}^4 \tag{10}$$

Concept Induction Module (CIM)

In both RPM-like and Bongard problems, solving the task requires relational abstraction—an observer must infer latent concepts (e.g., object identity, attribute regularities, and interobject relations) from the set of context images and then select the most consistent answer from the choice images. This inductive reasoning process inherently requires a comparison across images and compositions, which VPM alone cannot provide.

On the basis of the reduced problem illustrated in Fig. 4 (B and D), a set of pseudo training samples $\mathbf{s}^{\mathbf{i}} \in R^{N\times C\times H\times W}$ and their corresponding pseudo label y^i are constructed, where $N = 8$ (or 7) corresponds to RPM-like (or Bongard) problems. To model cross-image dependencies explicitly, we introduce the CIM, which is built upon a specialized convolutional block, as shown in Fig. 5. Each block comprises a residual branch with two stacked 3×3 convolutional layers and a shortcut branch with a 1×1 convolution. While structurally analogous to a ResBlock, this architecture is tailored for abstract reasoning by learning over sets of image features. To facilitate the extraction of abstract concepts, CIM first reformulates the tensor of $\mathbf{s}^i$ to make relational patterns more explicit

$$\begin{aligned} \tilde{\mathbf{s}}^i &= \text{Reformulate}\left(\mathbf{s}^i\right), \quad \tilde{\mathbf{s}}^i \in R^{C\times N\times L}, \quad L = H\times W \\ c^i &= \text{CConvBlock}_l\left(\tilde{\mathbf{s}}^i_l\right), \quad l \in \{1,2,3\} \end{aligned} \tag{11}$$

where Reformulate transforms the input shape of $N \times C \times H \times W$ to a shape of $C \times N \times L$, where $L = W \times H$ is obtained by flattening along spatial dimensions. After that, the ConvBlock extracts the concept along the second dimension (i.e., N, across different image features) and the third dimension (i.e., L, across feature spatials).

Optimization

Following concept extraction c^i from the i-th pseudo sample, two fully connected layers are used as a classifier to generate a single logit (*36*, *56*, *61*). The prediction is supervised via a binary cross-entropy loss computed against the corresponding pseudo label, formally defined as

$$\mathcal{L}(y,t) = \frac{1}{B\times A}\sum_{i=1}^{B}\sum_{a=1}^{\text{A}} y_i^a \log t_i^a + \left(1-y_i^a\right)\log\left(1-t_i^a\right) \tag{12}$$

where t and y are the output logit and the pseudo label, respectively. A is the number of pseudo samples per RPM-like or Bongard problem. B is the minibatch size. Optimization proceeds via backpropagation, enabling end-to-end training of all learnable parameters within the VPM and CIM modules. Stochastic gradient descent is used for this purpose.

Testing the trained model on full problems

As the model is trained on reduced versions of the problems with only partial context information, a direct application to the full problem setting is a challenge. To address this, we design an inference protocol that preserves the structural alignment with the reduced training configuration while incorporating the complete context required for full problem reasoning.

For an input RPM-like problem comprising eight context and eight candidate choices, each candidate choice $\mathbf{x}^i, i = \{9, 10, \ldots, 16\}$, is reformulated into two subreasoning problems to enable compatibility with the model trained on reduced problems. The first set is defined as $\mathbf{s}^{i1}_{\text{test}} = \{\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3, \mathbf{x}^7, \mathbf{x}^8, \mathbf{x}^i, \mathbf{x}^4, \mathbf{x}^5\}$, which includes the first and third rows of the context with i-th choice features. Here, $\{\mathbf{x}^4, \mathbf{x}^5\}$ serves as the constraint on the answer space. The second set is defined as $\mathbf{s}^{i2}_{\text{test}} = \{\mathbf{x}^4, \mathbf{x}^5, \mathbf{x}^6, \mathbf{x}^7, \mathbf{x}^8, \mathbf{x}^i, \mathbf{x}^1, \mathbf{x}^2\}$ using the second and third rows. $\{\mathbf{x}^1, \mathbf{x}^2\}$ is the constraint. Each of these sets is independently processed by the CIM to extract cross-image concept representations. The resulting concepts are passed through the classifier to yield logits t^{i1}_{test} and t^{i2}_{test}. The final prediction is made by selecting the candidate choice with the maximum combined logit across all options: $\underset{i\in\{9,10,\ldots,16\}}{\text{argmax}}\left(t^{i1}_{\text{test}} + t^{i2}_{\text{test}}\right)$.

For an input Bongard problem, while the training stage uses dropout to transform the original setting to its reduced version, the testing stage evaluates concept consistency using the full set of images and the real choice images without dropout. Specially, after the VPM process, for each choice feature $\mathbf{x}^i_c, i = \{1, 2\}$, as shown in Fig. 4C, we construct two sets: one composed of six positive features and the ith choice feature, $\mathbf{s}^{i1}_{\text{test}} = \{\mathbf{x}^1_p, \mathbf{x}^2_p, \mathbf{x}^3_p, \mathbf{x}^4_p, \mathbf{x}^5_p, \mathbf{x}^6_p, \mathbf{x}^i_c\}$, and one composed of five negative features and the same choice: $\mathbf{s}^{i2}_{\text{test}} = \{\mathbf{x}^1_n, \mathbf{x}^2_n, \mathbf{x}^3_n, \mathbf{x}^4_n, \mathbf{x}^5_n, \mathbf{x}^6_n, \mathbf{x}^i_c\}$. Each of these two sets is independently processed by our CIM to extract concepts embedded across image features and produce logits t^{i1}_{test} and t^{i2}_{test} by the classifier. After that, the final label for the ith choice is determined by the index that achieves the maximum logit: $\underset{j\in\{1,2\}}{\text{argmax}}\left(t^{ij}_{\text{test}}\right)$, where $j = 1$ denotes positive and $j = 2$ denotes negative.

Our SSLvMB's design enables inference over full problems while maintaining architectural compatibility with the reduced training regime. It reflects a simple-to-complex paradigm, wherein abstract concepts are initially acquired from simplified contexts and subsequently generalized to more complex configurations. This progression is consistent with cognitive theories, which suggest that humans learn to abstract relational patterns from limited information before integrating additional contextual elements as cognitive load permits. Such a strategy aligns with the principle of human MB, allowing the model to extend partially learned rules to novel, fully specified reasoning tasks.

Human behavioral experiments

To systematically evaluate human performance on all tasks used on our models, we designed five online experiments to collect human behavioral data. These experiments were conducted using the I-RAVEN (Exp-1), RAVEN-FAIR (Exp-2), PGM-Neutral (Exp-3), PGM-Interpolation (Exp-4), and PGM-Extrapolation (Exp-5). Human performance for other datasets, including RAVEN and the four test splits of Bongard-LOGO, was obtained directly from their original published studies. The original authors of Bongard-LOGO (*44*) reported experimental results for two groups of human subjects: One is provided task instructions and strictly adhered to the reasoning logic in the instructions, and the other did not undergo such rigorous training. In this study, our experimental comparisons include only the results from behavioral experiments conducted without specialized training.

To ensure data quality, we incorporated two validation questions to each tested dataset to assess participants' engagement levels. Only participants who correctly answered both validation questions were included in the final analysis. In addition, we excluded participants with unusually short response times or abnormal IP addresses. This validation process effectively filtered out individuals who were likely to provide random or unreliable answers.

All experiments were conducted through the NAODAO platform, where participants completed a set of questions. Each participant was required to answer 50 questions randomly selected from the respective dataset. The stimulus images were standardized to 7.6% of the screen length and 12.26% of the screen width to maintain consistency across participants. Participants were instructed to complete the task within 50 min (with a maximum allowance of 90 min), with flexibility in time allocation for individual questions. A mandatory 10-min break was provided after the first 25 questions, during which participants could rest or continue at their discretion. This break period was not counted toward the total response time.

The experiment followed a fixed sequence of test questions, with each question presented on a single page. Participants began by reviewing a two-page instruction manual to familiarize themselves with the question types and interface operations. During the task, participants selected their answers by clicking on a choice image, which was automatically placed in the position of the question mark, allowing them to visualize the complete matrix and identify concepts. If participants detected inconsistencies in the patterns across the three rows, they could modify their selection before finalizing their answer. Confirmed answers were submitted using the "Next" button to proceed to the subsequent question. In the human learning experiments, 100 human observers (50 for I-RAVEN and 50 for RAVEN-FAIR) were recruited and instructed to complete the exact same task as the main experiment. The only difference is that the participants can click a link to temporally step out the current question and enter a pool of learning examples. All learning examples are drawn from the respective datasets and accompanied with the correct answer. The participants thus had exposure to a few learning examples to familiarize themselves with RPM concepts.

Through our validation process, we obtained valid data from 25 participants for Exp-1 (I-RAVEN), 25 for Exp-2 (RAVEN-FAIR), 30 for Exp-3 (PGM-Neutral), 30 for Exp-4 (PGM-Interpolation), and 30 for Exp-5 (PGM-Extrapolation). A total of 140 subjects participated in our experiments.

Statistical analysis

We performed several statistical analyses to compare human and model performance. In Fig. 6, We used Welch's *t* test to compare human participants' performance against model performance on all datasets (i.e., I-RAVEN, RAVEN-F, PGM-N, PGM-I, PGM-E, BP-FF, BP-BA, BP-CM, and BP-NV), except RAVEN. The sample size numbers of BP datasets were directly derived from the literature. The human performance on RAVEN is directly derived from the original reference. We also performed Welch's *t* test to compare the human performance with and without learning (fig. S1).

Supplementary Materials

This PDF file includes:

Supplementary Text
Figs. S1 and S2
Tables S1 to S4

REFERENCES

1. K. M. Collins, I. Sucholutsky, U. Bhatt, K. Chandra, L. Wong, M. Lee, C. E. Zhang, T. Zhi-Xuan, M. Ho, V. Mansinghka, A. Weller, J. B. Tenenbaum, T. L. Griffiths, Building machines that learn and think with people. *Nat. Hum. Behav.* **8**, 1851–1863 (2024).
2. B. M. Lake, T. D. Ullman, J. B. Tenenbaum, S. J. Gershman, Building machines that learn and think like people. *Behav. Brain Sci.* **40**, e253 (2017).
3. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Association for Computational Linguistics, 2019).
4. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations* (ICLR, 2021).
5. K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2016), pp. 770–778.
6. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 [cs.CV] (2014).
7. A. Krizhevsky, I. Sutskever, G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25 (NIPS)*, F. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger, Eds. (Curran Associates, 2012), pp. 1097–1105.
8. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2009), pp. 248–255.
9. S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, WaveNet: A generative model for raw audio. arXiv:1609.03499 [cs.SD] (2016).
10. A. R. Romberg, J. R. Saffran, Statistical learning and language acquisition. *Wiley Interdiscip. Rev. Cogn. Sci.* **1**, 906–914 (2010).
11. J. R. Saffran, E. D. Thiessen, Pattern induction by infant language learners. *Dev. Psychol.* **39**, 484–494 (2003).
12. M. Ding, Y. Xu, Z. Chen, D. D. Cox, P. Luo, J. B. Tenenbaum, C. Gan, "Embodied concept learner: Self-supervised learning of concepts and mapping through instruction following," in *Proceedings of the 6th Conference on Robot Learning (CoRL)*, K. Liu, D. Kulic, J. Ichnowski, Eds., vol. 205 of Proceedings of Machine Learning Research (PMLR, 2023), pp. 1743–1754.
13. Y. Mi, W. Liu, Y. Shi, J. Li, Semi-supervised concept learning by concept-cognitive learning and concept space. *IEEE Trans. Knowl. Data Eng.* **34**, 2429–2442 (2020).
14. G. Rajendran, S. Buchholz, B. Aragam, B. Schölkopf, P. Ravikumar, "From causal to concept-based representation learning," in *Advances in Neural Information Processing Systems 37 (NeurIPS)*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, C. Zhang, Eds. (Curran Associates, 2024), pp. 101250–101296.
15. L. M. Schulze Buschoff, E. Akata, M. Bethge, E. Schulz, Visual cognition in multimodal large language models. *Nat. Mach. Intell.* **7**, 96–106 (2025).
16. C. Zhang, B. Jia, Y. Zhu, S. C. Zhu, Human-level few-shot concept induction through minimax entropy learning. *Sci. Adv.* **10**, eadg2488 (2024).
17. B. M. Lake, R. Salakhutdinov, J. B. Tenenbaum, Human-level concept learning through probabilistic program induction. *Science* **350**, 1332–1338 (2015).
18. R. B. Dekker, F. Otto, C. Summerfield, Curriculum learning for human compositional generalization. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2205582119 (2022).
19. P. Schwartenbeck, A. Baram, Y. Liu, S. Mark, T. Muller, R. Dolan, M. Botvinick, Z. Kurth-Nelson, T. Behrens, Generative replay underlies compositional inference in the hippocampal-prefrontal circuit. *Cell* **186**, 4885–4897.e14 (2023).
20. T. M. Houser, A. Resnick, D. Zeithamova, Successful generalization of conceptual knowledge after training to remember specific events. *Front. Cognit.* **3**, 1324678 (2024).
21. D. Gentner, Structure-mapping: A theoretical framework for analogy. *Cognit. Sci.* **7**, 155–170 (1983).
22. V. R. Bejjanki, R. Zhang, R. Li, A. Pouget, C. S. Green, Z. L. Lu, D. Bavelier, Action video game play facilitates the development of better perceptual templates. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 16961–16966 (2014).
23. R. Y. Zhang, A. Chopin, K. Shibata, Z. L. Lu, S. M. Jaeggi, M. Buschkuehl, C. S. Green, D. Bavelier, Action video game play facilitates "learning to learn". *Commun. Biol.* **4**, 1154 (2021).
24. Y. Y. Gao, Z. Fang, Q. Zhou, R. Y. Zhang, Enhanced "learning to learn" through a hierarchical dual-learning system: The case of action video game players. *BMC Psychol.* **12**, 460 (2024).
25. J. B. Tenenbaum, "Bayesian modeling of human concept learning," in *Advances in Neural Information Processing Systems 11 (NIPS)*, M. S. Kearns, S. A. Solla, D. A. Cohn, Eds. (MIT Press, 1999), pp. 59–65.
26. T. L. Griffiths, M. L. Kalish, S. Lewandowsky, Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **363**, 3503–3514 (2008).
27. J. Wang, C. Zhang, J. Li, Y. Ma, L. Niu, J. Han, Y. Peng, Y. Zhu, L. Fan, Evaluating and modeling social intelligence: A comparative study of human and AI capabilities. arXiv:2405.11841 [cs.AI] (2024).
28. J. Raven, The Raven's progressive matrices: change and stability over culture and time. *Cogn. Psychol.* **41**, 1–48 (2000).
29. M. M. Bongard, *The Recognition Problem* (Western Washington University, 1968).
30. M. Hersche, M. Zeqiri, L. Benini, A. Sebastian, A. Rahimi, A neuro-vector-symbolic architecture for solving Raven's progressive matrices. *Nat. Mach. Intell.* **5**, 363–375 (2023).
31. S. T. Piantadosi, J. B. Tenenbaum, N. D. Goodman, Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition* **123**, 199–217 (2012).
32. B. Zhao, C. G. Lucas, N. R. Bramley, A model of conceptual bootstrapping in human cognition. *Nat. Hum. Behav.* **8**, 125–136 (2024).
33. G. S. Halford, W. H. Wilson, S. Phillips, Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behav. Brain Sci.* **21**, 803–831 (1998).
34. C. Zhang, F. Gao, B. Jia, Y. Zhu, S.-C. Zhu, "RAVEN: A dataset for Relational and Analogical Visual REasoNing," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2019), pp. 5317–5327.
35. S. Hu, Y. Ma, X. Liu, Y. Wei, S. Bai, "Stratified rule-aware network for abstract visual reasoning," in *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)* (AAAI, 2021), pp. 1567–1574.

36. Y. Benny, N. Pekar, L. Wolf, "Scale-localized abstract reasoning," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2021), pp. 12557–12565.
37. D. Barrett, F. Hill, A. Santoro, A. Morcos, T. Lillicrap, "Measuring abstract reasoning in neural networks," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, J. Dy, A. Krause, Eds., vol. 80 of Proceedings of Machine Learning Research (PMLR, 2018), pp. 511–520.
38. T. Zhuo, Q. Huang, M. Kankanhalli, Unsupervised abstract reasoning for Raven's Problem Matrices. *IEEE Trans. Image Process.* **30**, 8332–8341 (2021).
39. N. Q. W. Kiat, D. Wang, M. Jamnik, Pairwise relations discriminator for unsupervised Raven's Progressive Matrices. arXiv:2011.01306 [cs.AI] (2020).
40. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)* (IEEE, 2017), pp. 618–626.
41. N. Arcot, P. Srivastava, S. Jaarsveld, "Do constraints in APM solving affect APM-like puzzle creation?," in *Proceedings of the 45th Annual Conference of the Cognitive Science Society* (Cognitive Science Society, 2023).
42. T. Ito, T. Klinger, D. Schultz, J. Murray, M. Cole, M. Rigotti, "Compositional generalization through abstract representations in human and artificial neural networks," in *Advances in Neural Information Processing Systems 35 (NeurIPS)*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh, Eds. (Curran Associates, 2022), pp. 32225–32239.
43. B. M. Lake, T. Linzen, M. Baroni, Human few-shot learning of compositional instructions. arXiv:1901.04587 [cs.CL] (2019).
44. W. Nie, Z. Yu, L. Mao, A. B. Patel, Y. Zhu, A. Anandkumar, "Bongard-LOGO: A new benchmark for human-level concept learning and reasoning," in *Advances in Neural Information Processing Systems 33 (NeurIPS)*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin, Eds. (Curran Associates, 2020), pp. 16468–16480.
45. T. Webb, K. J. Holyoak, H. Lu, Emergent analogical reasoning in large language models. *Nat. Hum. Behav.* **7**, 1526–1541 (2023).
46. J. Raven, *Handbook of Nonverbal Assessment* (Springer, 2003), pp. 223–237.
47. Y. Bengio, J. Louradour, R. Collobert, J. Weston, "Curriculum learning," in *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, A. P. Danyluk, L. Bottou, M. L. Littman, Eds., vol. 382 of ACM International Conference Proceeding Series (Association for Computing Machinery, 2009), pp. 41–48.
48. G. H. Bower, A contrast effect in differential conditioning. *J. Exp. Psychol.* **62**, 196–199 (1961).
49. H. S. Meyerhoff, F. Papenmeier, M. Huff, Studying visual attention using the multiple object tracking paradigm: A tutorial review. *Atten. Percept. Psychophys.* **79**, 1255–1274 (2017).
50. M. P. Kumar, B. Packer, D. Koller, "Self-paced learning for latent variable models," in *Advances in Neural Information Processing Systems 23 (NIPS)*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, A. Culotta, Eds. (Curran Associates, 2010), pp. 1189–1197.
51. R. S. Sutton, D. Precup, S. Singh, Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artif. Intell.* **112**, 181–211 (1999).
52. Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798–1828 (2013).
53. G. E. Hinton, S. Osindero, Y. W. Teh, A fast learning algorithm for deep belief nets. *Neural Comput.* **18**, 1527–1554 (2006).
54. T. Chen, S. Kornblith, M. Norouzi, G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, H. Daumé III, A. Singh, Eds., vol. 119 of Proceedings of Machine Learning Research (PMLR, 2020), pp. 1597–1607.
55. G. Konidaris, L. P. Kaelbling, T. Lozano-Perez, From skills to symbols: Learning symbolic representations for abstract high-level planning. *J. Artif. Intell. Res.* **61**, 215–289 (2018).
56. L. Yang, H. You, Z. Zhen, D. Wang, X. Wan, X. Xie, R.-Y. Zhang, "Neural prediction errors enable analogical visual reasoning in human standard intelligence tests," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett, Eds., vol. 202 of Proceedings of Machine Learning Research (PMLR, 2023), pp. 39572–39583.
57. K. Zhao, C. Xu, B. Si, "Learning visual abstract reasoning through dual-stream networks," in *Proceedings of the AAAI Conference on Artificial Intelligence* (Association for the Advancement of Artificial Intelligence, 2024), pp. 16979–16988.
58. S. S. Mondal, J. D. Cohen, T. W. Webb, "Slot abstractors: Toward scalable abstract visual reasoning," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, F. Berkenkamp, Eds., vol. 235 of Proceedings of Machine Learning Research (PMLR, 2024), pp. 36088–36105.
59. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh, "VQA: Visual Question Answering," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)* (IEEE, 2015), pp. 2425–2433.
60. M. Xu, G. Jiang, W. Liang, C. Zhang, Y. Zhu, "Active reasoning in an open-world environment," in *Advances in Neural Information Processing Systems 36 (NeurIPS)*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine, Eds. (Curran Associates, 2023), pp. 11716–11736.
61. C. Zhang, B. Jia, F. Gao, Y. Zhu, H. Lu, S.-C. Zhu, "Learning perceptual inference by contrasting," in *Advances in Neural Information Processing Systems 32 (NeurIPS)*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett, Eds. (Curran Associates, 2019), pp. 1073–1085.

Acknowledgments: We thank B. Yuan and C. Li for human data collection. **Funding:** This work was supported by the National Natural Science Foundation of China (32441102 to R.-Y.Z. and 62206316 to L.Y.), Shanghai Municipal Education Commission (2024AIZD014 to R.-Y.Z), the State Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology (SKLBI-K2025005 to R.-Y.Z), the Fundamental Research Funds for the Central Universities, Peking University (to R.-Y.Z), the Project of Guangdong Provincial Key Laboratory of Information Security Technology (2023B1212060026 to L.Y., X.X., and J.L.), and Beijing Natural Science Foundation (L247010 to Z.Z.). **Author contributions:** Conceptualization: R.-Y.Z. and L.Y. Data curation: M.L., R.-Y.Z., and L.Y. Formal analysis: Y.-J.W., M.L., R.-Y.Z., X.X., and L.Y. Funding acquisition: R.-Y.Z., Z.Z., X.X., and L.Y. Investigation: M.L., R.-Y.Z., Z.Z., X.X., L.Y., and J.L. Methodology: R.-Y.Z., Z.Z., X.X., and L.Y. Project administration: R.-Y.Z. and X.X. Resources: R.-Y.Z., Z.Z., L.Y., and J.L. Software: Y.-J.W., M.L., R.-Y.Z., Z.Z., and L.Y. Supervision: R.-Y.Z. and X.X. Validation: M.L., R.-Y.Z., Z.Z., L.Y., and J.L. Visualization: Y.-J.W., M.L., R.-Y.Z., and L.Y. Writing—original draft: Y.-J.W., R.-Y.Z., X.X., and L.Y. Writing—review and editing: M.L., R.-Y.Z., X.X., L.Y., and J.L. **Competing interests:** The authors declare that they have no competing interests. **Data, code, and materials availability:** All model code, experiment code, and data and materials for the human experiments needed to evaluate and reproduce the results in the paper are present in the paper and the Supplementary Materials and are publicly deposited in Zenodo: https://doi.org/10.5281/zenodo.17846529. The model code can be used to reproduce Figs. 1 to 6 and fig. S2. The human behavioral data and corresponding plotting code can be used to reproduce Fig. 6 and fig. S1. Details for how to synthesize materials are also included in Materials and Methods. This study did not generate new materials.

Submitted 18 July 2025
Accepted 15 July 2026
Published 28 August 2026
10.1126/sciadv.aea7202

ScienceAdvances

Mental bootstrapping enables human-level concept learning in self-supervised deep models

Lingxiao Yang, Muyang Lyu, Ying-Jie Wang, Xiaohua Xie, Zonglei Zhen, Jian-Huang Lai, and Ru-Yuan Zhang

Sci. Adv. **12** (35), eaea7202. DOI: 10.1126/sciadv.aea7202

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.