BIOMIMETICS

Development of compositionality through interactive learning of language and action of robots

Prasanna Vijayaraghavan, Jeffrey Frederic Queißer, Sergio Verduzco Flores, Jun Tani*

Humans excel at applying learned behavior to unlearned situations. A crucial component of this generalization behavior is our ability to compose/decompose a whole into reusable parts, an attribute known as compositionality. One of the fundamental questions in robotics concerns this characteristic: How can linguistic compositionality be developed concomitantly with sensorimotor skills through associative learning, particularly when individuals only learn partial linguistic compositions and their corresponding sensorimotor patterns? To address this question, we propose a brain-inspired neural network model that integrates vision, proprioception, and language into a framework of predictive coding and active inference on the basis of the free-energy principle. The effectiveness and capabilities of this model were assessed through various simulation experiments conducted with a robot arm. Our results show that generalization in learning to unlearned verb-noun compositions is significantly enhanced when training variations of task composition are increased. We attribute this to self-organized compositional structures in linguistic latent state space being influenced substantially by sensorimotor learning. Ablation studies show that visual attention and working memory are essential to accurately generate visuomotor sequences to achieve linguistically represented goals. These insights advance our understanding of mechanisms underlying development of compositionality through interactions of linguistic and sensorimotor experience.

INTRODUCTION

The problem of generalizing learned behavior to unlearned situations is easy for humans but incredibly challenging for cognitive robots. Compositionality (*1*–*4*) is a major linguistic competency that is essential for generalization of cognitive behavior. Lake and colleagues (*5*) considered compositionality as one of the three fundamental competencies necessary to build machines that can learn to think like humans. Although interpretations of compositionality vary, Hupkes *et al.* (*6*) defined it by identifying its essential components. Among these, systematicity, the ability to recombine known parts and rules for use in a novel context, is a central component of compositionality, on which we focus in the current study. Recent deep learning models seem to trivialize this problem, but in reality, they offer little insight into how language develops in humans. Although it can be argued that using large language models (LLMs) for end-to-end learning shows that they can understand the meanings of words by learning from a large corpus collected in the real world (*7*–*11*), these models cannot access any sensorimotor patterns associated with words and sentences. Our objective is to understand how the aforementioned systematicity aspect of compositionality in language and behavior can codevelop through their interactions, by building an integrative neural network model for conducting robotic simulation experiments.

We used a developmental robotic approach in conjunction with the free-energy principle (FEP) (*12*) to address this problem. Modeling embodied language with developmental robotics is consistent with the constructivist view or usage-based theory of language acquisition (*13*, *14*). Embodiment is considered a necessary precondition for developing higher thoughts (*15*). According to Piaget (*16*), infants develop body-rationality representation through sensorimotor interactions with the environment, accompanied by goal-directed actions. Neuroscience researchers (*17*–*19*) have found that modulation of motor system activity occurs while listening to sentences expressing actions, suggesting that humans infer actions from language and vice versa. Developmental robotics (*20*) also addresses the symbol grounding (*21*) problem, which seeks to understand how symbols commonly used in linguistic expressions are associated with meaning in the real world. Cognitive competencies such as visual attention and visual working memory (VWM) are crucial in development of embodied language (*14*, *22*–*25*). Using developmental robotics, several studies have investigated the association of language and visuo-proprioceptive behavior with hierarchical multimodal recurrent neural networks (*26*–*31*).

In parallel with the developmental robotic approach, recent advances in cognitive neuroscience have underscored the importance of theoretical frameworks, such as the FEP, in modeling cognitive brain mechanisms. According to the FEP, perception and action are modeled in the framework of predictive coding (*32*, *33*) and active inference (AIF) (*34*–*36*), respectively. Predictive coding is a theory of perception that provides a unifying framework for neuronal mechanisms of top-down prediction and perceptual learning of sensory information. AIF is a process for inferring actions that minimize the error between preferred sensation and predicted actional outcomes. Some neural network models (*35*–*39*) have been successfully incorporated into goal-directed planning schemes on the basis of AIF to show that artificial agents can generate adequate goal-directed behaviors on the basis of learning in the habituated range of the world. Some of these models (*35*, *36*) generated action plans by optimizing the policy, and others (*37*–*39*) optimized low-dimensional latent variables by minimizing the future expected free energy. Teleology, a philosophical concept that explains phenomena on the basis of their ultimate goals or purposes rather than merely their causes or origins, aligns closely with the principles of goal-directed behavior observed in these models. In the context of human behavior, teleology offers a framework for interpreting actions as inherently goal directed and purpose driven (*40*, *41*), providing a philosophical underpinning that complements mechanistic insights offered by FEP-based approaches.

Okinawa Institute of Science and Technology, Okinawa, Japan.
*Corresponding author. Email: jun.tani@oist.jp

Inspired by these ideas, we propose a neural network model (Fig. 1A) to study codevelopment of linguistic compositionality paralleling sensorimotor experience (Movie 1). It consists of RNN-based generative networks that handle prediction of vision, proprioception, and language. These modalities are integrated by the associative network. Our model uses an executive layer mechanized by

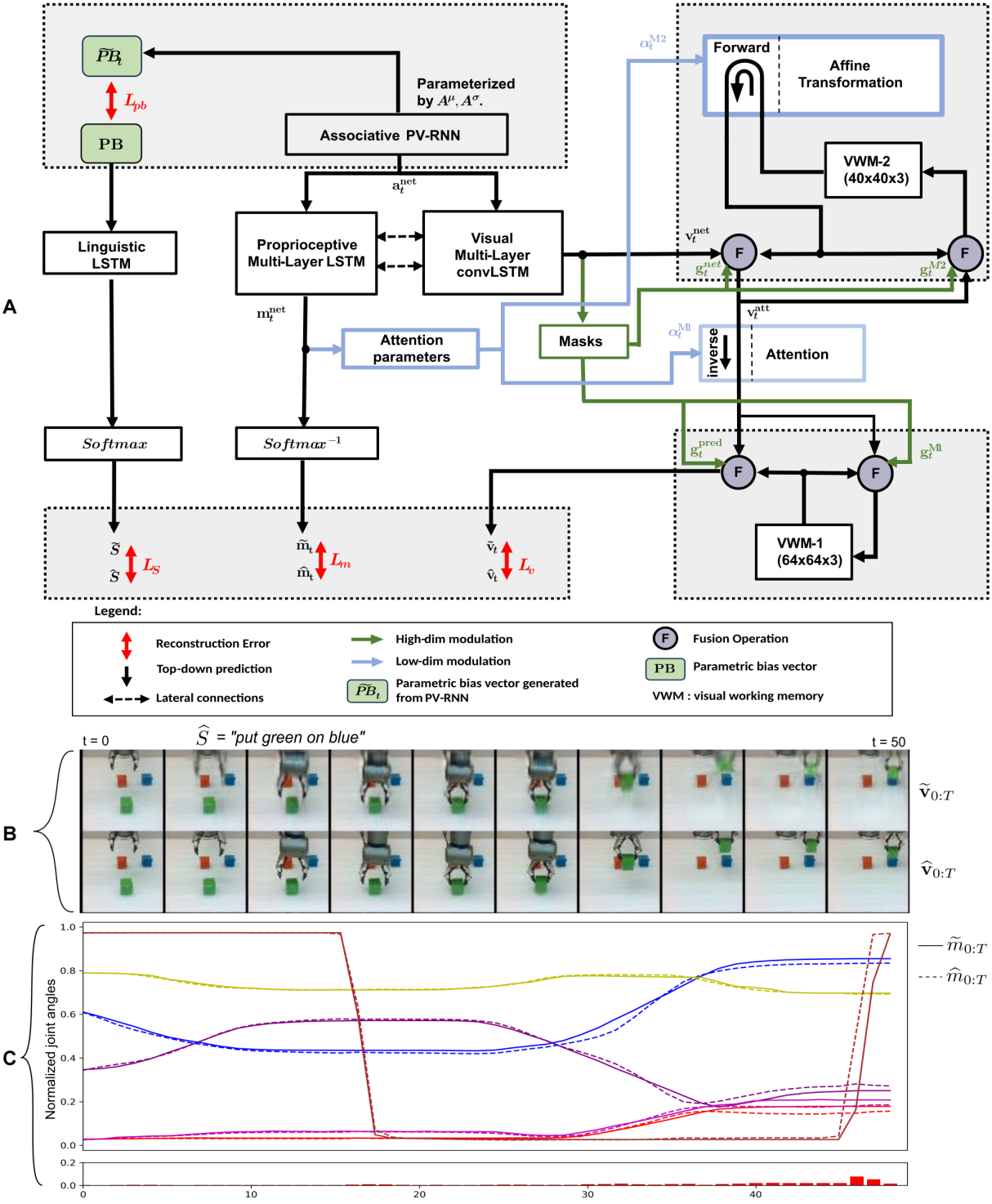

Fig. 1. Neural network model and plan generated by the model to achieve a given linguistic goal. (**A**) Model architecture: Visual (convLSTM) and proprioceptive (LSTM) modalities are integrated by the associative PV-RNN, and the linguistic LSTM is bound to the associative PV-RNN via **PB**. Visual predictions are enhanced by VWM-1, VWM-2, and attention mechanisms. For the given linguistic goal "put green on blue," (**B**) the predicted visual sequence was compared with the observed visual sequence and (**C**) the predicted joint angle trajectory was compared with the observed joint angle trajectory, showing the motor prediction error.

Movie 1. Summary of the model and experimental results.

predictive coding–inspired variational RNN (PV-RNN) (*42*) to integrate language with visuo-proprioceptive sequences. PV-RNN, a neural network consistent with the FEP, contains probabilistic latent variables that allow it to learn probabilistic structure hidden in the data. Some studies (*43*, *44*) have found that abstract representations, such as vector representations, reconcile compositional generalization with distributed neural codes.

We introduce a parametric bias (**PB**) (*45*, *46*) vector as the language latent variable. **PB** is a low-dimensional latent state vector used in recurrent neural network models. In training of multiple temporal patterns, the **PB** vector space is self-organized such that each temporal pattern is encoded by a specific point in the **PB** vector space. Sugita and Tani (*26*) showed that word sequences and corresponding behavioral temporal patterns can be bound using the **PB**, facilitating the development of linguistic compositionality. In the current model, the associative PV-RNN is constrained by the **PB** vector, which influences learning of associations among vision, proprioception, and linguistics.

This model learns to generate visuo-proprioceptive sequences with appropriate linguistic predictions by minimizing evidence free energy. The trained model generates appropriate visuo-proprioceptive sequences to achieve linguistically represented goals via goal-directed planning by means of AIF. Figure 1 (B and C) shows the visuomotor sequences predicted by the model compared with observed ground truth. We used a teleology-inspired approach to goal-directed planning proposed by Matsumoto *et al.* (*38*). The underlying concept is that goal expectation is generated at every time step instead of expecting the goal at a distal step.

Through simulation experiments, we aimed to study how variations in training compositions affect generalization and to understand the mechanisms underlying the development of compositionality. Our simulation experiments revealed the following: First, generalization in learning improves significantly as the number of variations in task compositions increases. Second, the compositional structure that emerges in the linguistic latent state representation is strongly influenced by sensorimotor learning. Specifically, we observed that the linguistic latent representation of actional concepts develops by preserving similarity among corresponding sensorimotor patterns. Last, by performing ablation studies, we found that the model's ability to accurately generate visuo-proprioceptive sequences is significantly affected by the presence of visual attention and working memory modules.

RESULTS

Task description

In the current study, we introduced vision-based object manipulation tasks with a robotic arm (fig. S1). The tasks included grasping, moving (in four different directions—left, right, front, and back), and stacking. These tasks were performed on 5-cm cubic blocks of five colors (red, green, blue, purple, and yellow). Tasks were linguistically represented by sentences like "grasp X," "move X left," "move X right," "move X front," "move X back," and "put X on Y"; "X" indicates the color of the object being manipulated (any of the five colors), and "Y" is the color of the object at the base (we used green, blue, or yellow) in the stacking task. Examples of visuomotor sequences of different types of tasks are shown in fig. S2. In total, there are 40 possible combinations (five nouns and eight verbs). The model was trained with data collected from the physical robot. However, all evaluations were performed on the basis of the ability of the model to generate mentally simulated trajectories of visuo-proprioceptive sequences.

We performed two experimental evaluations with the above setup. First, we evaluated model performance for generalization to unlearned object positions and unlearned linguistic compositions. We further emphasized this by comparing model performance among different degrees of sparsity in training data. We also evaluated the model's ability to understand visuo-proprioceptive behavioral sequences by inferring the appropriate linguistic description. Second, we performed an ablation experiment to study the impact of visual attention and working memory on the model's generalization capability.

Evaluation of ability to generalize: Experiment I

In this experiment, we evaluated the ability of the model to generalize to unlearned object positions and unlearned language compositions. The dataset was divided into four groups, each with a different number of combinations. Group A contained 40 combinations (five nouns and eight verbs). Group B comprised 30 combinations (five nouns and six verbs). Group C included 15 combinations (five nouns and three verbs), and Group D contained 9 combinations (three nouns and three verbs). To evaluate model performance in generalization in learning, we further divided each group with different ratios of training. Details of different training ratios in each group are described in Table 1. Note that in group D, because the

Table 1. Training ratios. Ratios of various compositions of behaviors used for training.

Groups	1	2	3
Group A (5 × 8)	32/40 (80%)	24/40 (60%)	16/40(40%)
Group B (5 × 6)	24/30 (80%)	18/30 (60%)	12/30 (40%)
Group C (5 × 3)	12/15 (80%)	9/15 (60%)	6/15 (40%)
Group D (3 × 3)	7/9 (77%)	6/9 (66%)	3/9 (33%)

total number of combinations is nine, we used the ratios 77, 66, and 33% instead of 80, 60, and 40%. Details of individual compositions in the four groups are illustrated in figs. S3 to S5.

Inference of visuo-proprioceptive sequences to achieve linguistically specified goals

The model was trained with visuo-proprioceptive sequences that started at random initial configurations of objects in the work space. As previously mentioned, the trained model performed goal-directed planning using AIF for novel object positions and unlearned compositions (U-Cs) of linguistically represented goals, as well as for novel object positions (U-Ps) and learned linguistically represented goals. The error between the visuo-proprioceptive plan generated by the network and the ground truth of visuo-proprioceptive trajectory was measured to evaluate the model's generalization performance.

An example of successful generation of a goal-directed action plan to achieve a linguistically represented goal, when trained with group A1, is shown in Fig. 2. A visuo-proprioceptive mean squared error of 0.0113 was observed in this successful example. This figure shows mental simulation of the generated motor plan and the expected visual trajectory associated with the linguistically specified goal ("put green on blue"). Figure 2G compares ground truth joint-angle trajectories of the test sequence with inferred trajectories from the planning process. Trajectories 0 to 4 (red, blue, green, yellow, and purple) represent joint angles of all five active rotary joints of the robot arm and joint. Joint number 5 (brown) refers to linear actuators

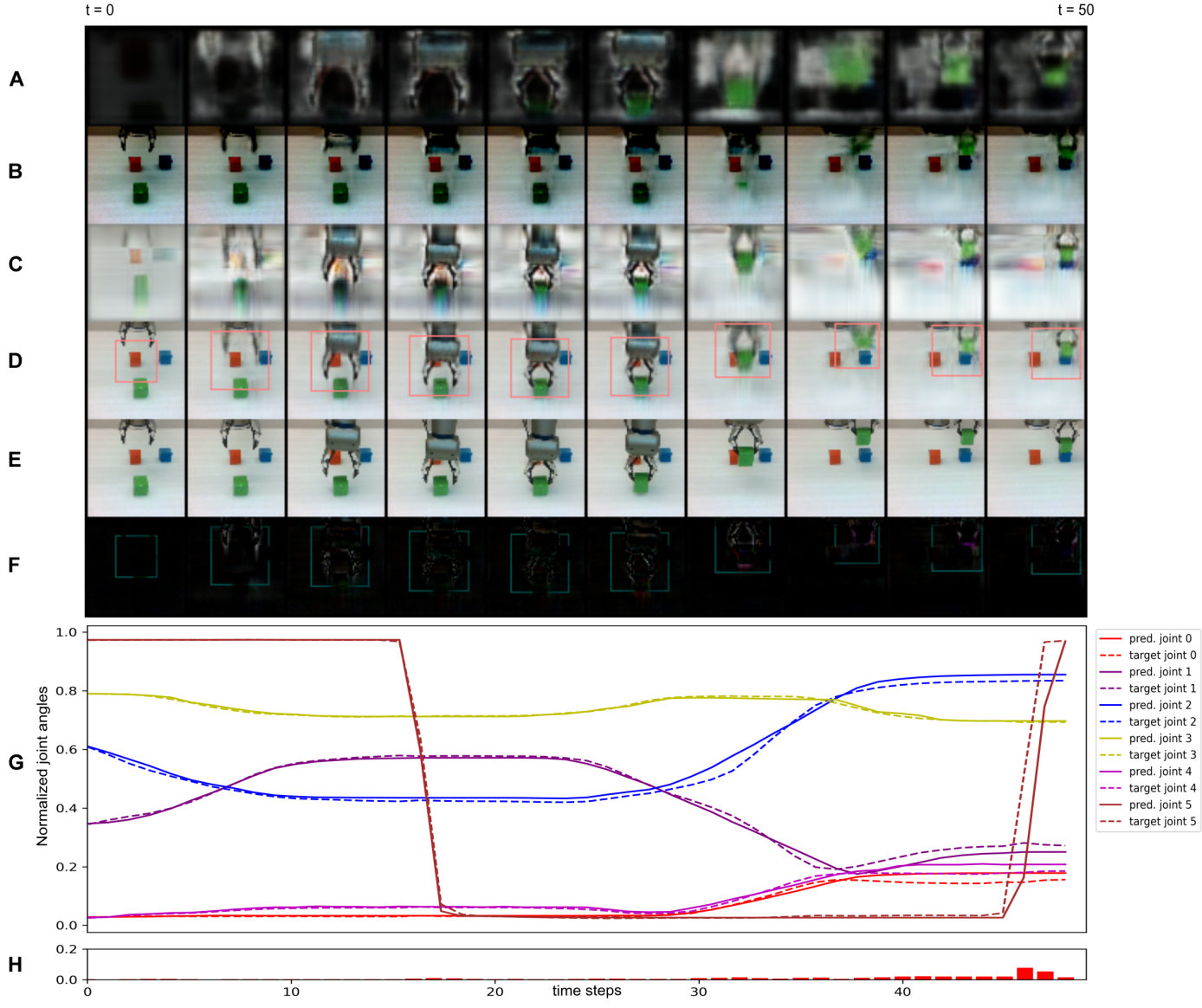

Fig. 2. Goal-directed planning using AIF. The model generated the above visuo-proprioceptive sequence for the linguistically specified goal "put green on blue." (**A**) Masked representation of VWM-2. (**B**) VWM-1. (**C**) Model prediction of attended visual stream. (**D**) Final simulated prediction of the visual stream; the red box indicates coordinates for attention predicted by the proprioceptive LSTM. (**E**) The ground truth target for comparison. (**F**) Difference between the predicted visual stream and the ground truth target. (**G**) Normalized joint angle trajectory predicted by the model compared with the corresponding observed joint angle trajectory. (**H**) Mean difference between the predicted joint angles and the observed joint angles.

of the robot gripper. The visual stream shows every fifth time step of the generated sequence of the model. The current focal area, in terms of size and position of the attention transformation, is indicated by a red square. Parameterization of the attention transformer is generated as an additional output of the multilayer proprioceptive LSTM (long short-term memory), as previously mentioned. Attention (red box in Fig. 2D) is directed toward the object to be manipulated and the gripper when the gripper starts to approach the object. The contents of VWM-2 and VWM-1 are illustrated in Fig. 2 (A and B), respectively. The shape and color of the manipulated object are represented in VWM-2. Note that when the object is being moved, it disappears from VWM-1 and appears in VWM-2. This information flow between VWM-1 and VWM-2 emerged in the visual network through training, which is essential for generalization to novel situations, based on our previous work (*39*). We leveraged this emergent mechanism in the current model to facilitate grounding of language to visuo-proprioceptive behavior. These observations show that a mental image of a continuous visuo-proprioceptive pattern can be generated using goal-directed planning to achieve a linguistically specified goal.

The model is adept at associating linguistically represented goals with corresponding visuo-proprioceptive sequences. Details of the model's performance for generalizing to U-Ps with learned linguistically represented goals and to U-Cs of linguistically represented goals are provided in table S1. Despite fewer variations of linguistic composition in training, the model still maintained a low error, ≤0.0405 (group D3), for U-Ps. Figure 3 shows the comparison of generalization performance between U-Ps and U-Cs for different groups with the highest training ratio. Although U-P performance does not change significantly, depending on the composition scale, U-C performance improves as the variation of task composition in the training increases from group D1 to group A1.

Figure 4 illustrates the difference in U-C performance between groups when trained with different training ratios, as mentioned above. In the majority of failed cases, the model confused colors of the object or misinterpreted the action to be performed on the object.

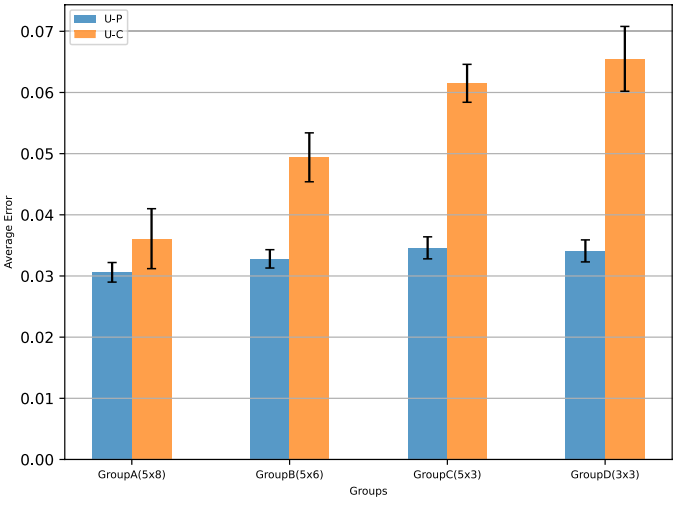

Fig. 3. Comparison between position generalization and compositional generalization. Average error for the inference of visuo-proprioceptive plans compared between U-Ps and U-Cs among groups with different number of compositions with the highest training ratio of 80%. Error bars show SD with sample size $n = 5$.

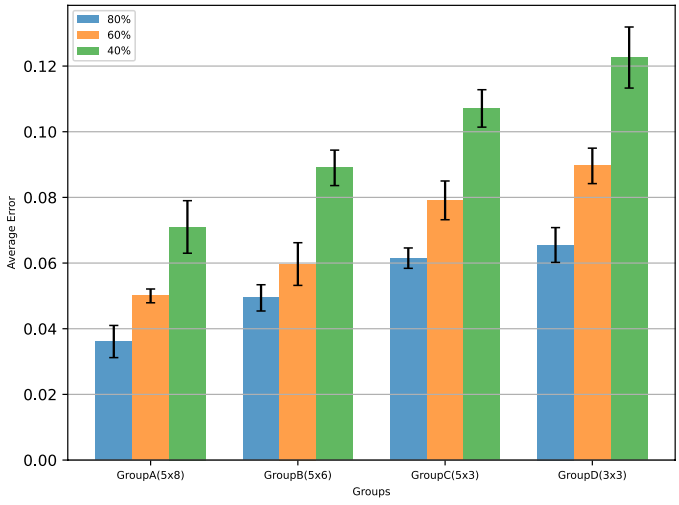

Fig. 4. Effect of variation and sparsity in training on generalization. Comparison of average visuo-proprioceptive error between groups with different training ratios for inference of visuo-proprioceptive plans to achieve U-Cs of linguistically represented goals. Error bars show SD with sample size $n = 5$.

This indicated that failure in generalization occurred at the abstract level, given that low-level predictions still generated sequences corresponding to a different action or performed the specified action on a different object. Examples of visuo-proprioceptive sequence generated by the model in cases of successful and failed generalizations are shown in figs. S6 and S7, respectively. Even when the model fails to generalize, it still generates visuo-proprioceptive sequences that do not result in large errors. This likely explains why, despite poor performance in some instances, the average error remains relatively low.

We qualitatively analyzed a latent space, the **PB** space, of linguistic-LSTM with kernel principal components analysis (KPCA) (*47*) using linear kernels. Figure 5 shows a scatterplot of mean KPCA values of each cluster in the **PB** space of the model for all groups with the highest training ratio. Figure S9 shows the data distribution corresponding to mean values shown here. The topology of hidden states corresponding to verb phrases ("grasp," "move left," etc.) seems to have a common structure when visualized separately for each object noun. For example, in Fig. 5A, representations of actions related to the stacking task are aligned on the left side and follow the order of "put X on green," "put X on yellow," and "put X on blue" (X refers to the object noun that corresponds to the object color in our experiments), from top to bottom. Similarly, in Fig. 5A, representations of actions related to "move" are clustered on the right side and follow the order "move X right," "move X back," "move X front," and "move X left," from top to bottom. Also, the representation of actions related to "grasp" is always between moving and stacking actions. Two things can be said from these observations. First, **PB** vectors corresponding to similar action categories become similar, and second, this structure is largely similar for different colors, which implies compositionality between verbs and nouns.

This structure is more consistent among different object colors when the variety of task compositions in learning is high, as seen in Fig. 5 (A and B). In contrast, in Fig. 5 (C and D), topologies of hidden states are inconsistent among different object colors. We found that structural relationships between learned compositions are extrapolated by

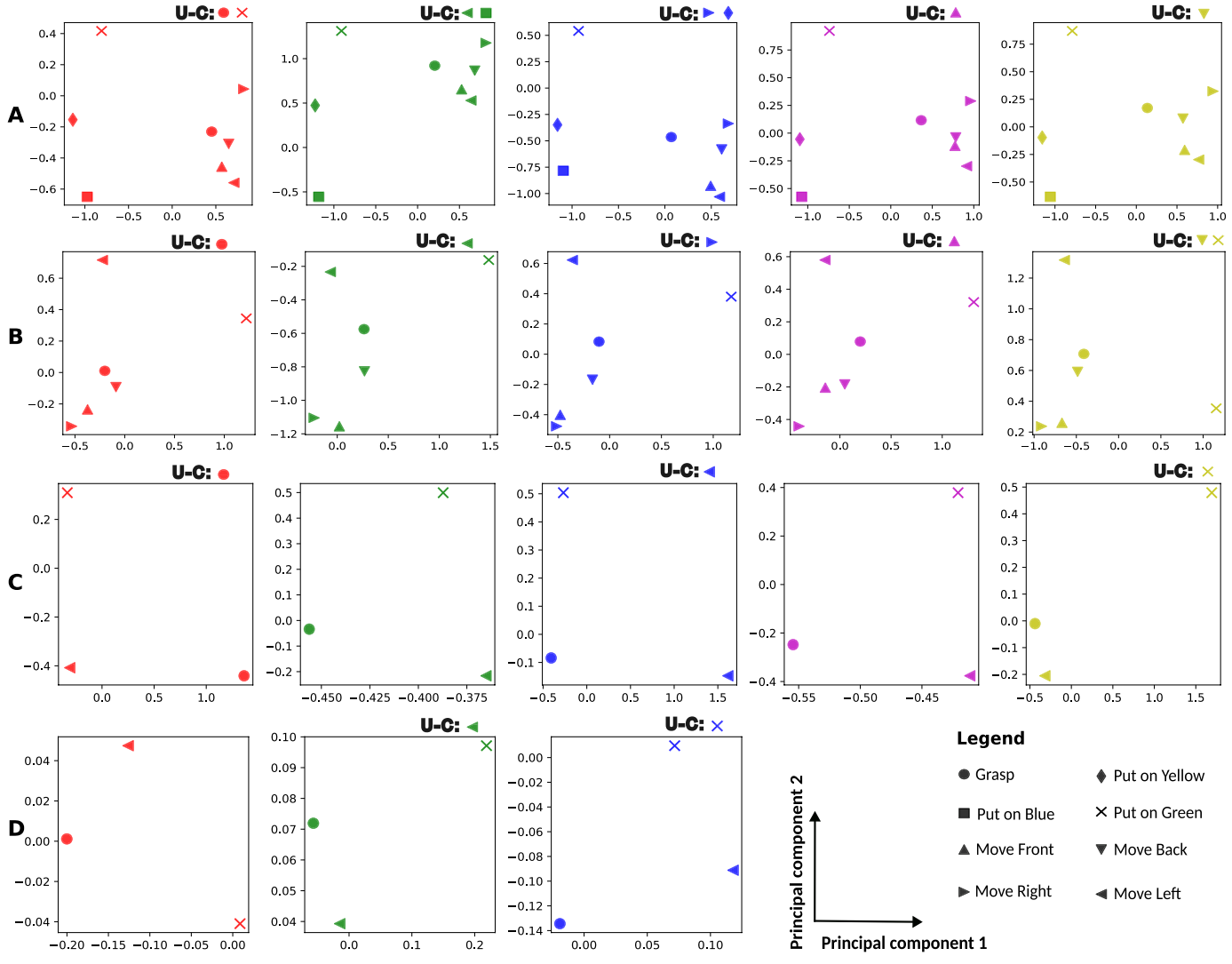

Fig. 5. Latent state PB representation. Scatterplot of mean kernel PCA values of latent state **PB** vectors for all groups with the highest training ratio. (**A**) Group A1 (5 × 8, 80%). (**B**) Group B1 (5 × 6, 80%). (**C**) Group C1 (5 × 3, 80%). (**D**) Group D1 (3 × 3, 77%). Colors of markers indicate the color of the object being manipulated. The variance explained by the two components of KPCA for all groups was greater than 90%.

the model to generalize to U-Cs of nouns and verbs, illustrated by positions of U-Cs (Fig. 5). The model's ability to extrapolate this structure to U-Cs improves as the number of variations in task compositions increases. We also performed Welch's unequal variances *t* test (*48*) to compare the generalization performance (for U-Cs) between groups, the results of which are shown in tables S3 and S4. The model showed robust generalization performance when trained with only 60% of training data for groups with a greater variety of task compositions (groups A and B). Groups trained with 40% sparsity performed poorly, irrespective of the variety of task compositions in training. We observed a significant difference in model performance between subsequent groups, where groups with more variations in task compositions always outperformed groups with fewer elements for compositions.

Language inference from observed visuo-proprioceptive sequences

The model's ability to infer linguistically represented goals from visuo-proprioceptive behaviors was evaluated, with success defined by accurate inference of unlearned word sequences. Generalization performance for U-Ps and linguistic U-Cs is shown in table S2, with Welch's *t* test results in tables S5 and S6. Comparisons across different training ratios are illustrated in fig. S8, demonstrating that increased training composition size and ratios improve generalization. This is analogous to results obtained for inference of visuo-proprioceptive plan sequences (Fig. 4). Although the model minimizes the error between generated visuo-proprioceptive sequences and the ground truth, success is measured by accuracy of inferred linguistic goals (fig. S8 and table S2). Generalization performance in inferring appropriate linguistic goals was poorer than expected. This was possibly caused by the relatively noisy visual image sequence that was used for inference. (Note that this goal inference, as well as training of the network, was performed by sampling real visual data operated by a physical robot.)

Ablation study: Experiment II

To assess the impact of visual attention and working memory on the current model's performance, we conducted ablation studies to

evaluate the model's ability to generate mental simulation of visuo-proprioceptive sequences when provided with linguistically represented goals for group A1 (5 × 8, 80%). We ablated the visual attention and VWM-1 and VWM-2 (Table 2) and trained the model from scratch before evaluation. Ablation of either working memory or the attention module notably affected the ability of the model to generate successful visual predictions, which resulted in poor generalization performance (Table 2). The model showed significantly degraded generalization performance for both U-Ps and U-Cs. The proprioceptive prediction capability of the model was also reduced significantly but less than visual prediction capability, because of ablation of the visual network. This is reflected in the proprioceptive accuracy (Table 2).

The model performed better when there was at least one VWM with attention compared with having no attention module. These results highlight the importance of the interaction between visual attention and working memory for generating accurate visuo-proprioceptive predictions. Further implications of these results are discussed below.

DISCUSSION

This study investigated how generalization in compositionality can be achieved through the process of associative learning between action and language, with limited amount of the experience, even with limited training data. We hypothesized that increasing task composition variation used in learning would improve generalization. This hypothesis was evaluated by conducting a set of simulated robotic experiments using the framework of AIF (*35*). More specifically, we studied how robots can learn to generate goal-directed action plans by adequately inferring visuo-proprioceptive sequences to achieve linguistically specified goals. We also examined how such a robot can infer linguistically represented goals from observation of provided visuo-proprioceptive sequences. For this purpose, we built on our previous work (*39*) by adding a language-processing LSTM and a PV-RNN in the associative layer and by using a teleological approach for goal-directed planning (*38*). The proposed model used a complex visual network with submodules, including visual attention and VWMs, to facilitate grounding of language to visuo-proprioceptive behavioral sequences.

Our analysis of simulation experiment results to infer action plans for linguistically specified goals led to the following findings. First, generalization performance in learning U-Cs improves with increased vocabulary, as well as the training ratio. This is supported by an analysis of the **PB** space that showed more consistent relational structures among different concepts combining actions and object nouns. These emerge for cases with more variations of task composition in learning. We found that representations in the **PB** space self-organized on the basis of the similarity of actions. This validates our basic hypothesis. Second, performance for position generalization does not depend on the size of composition used in learning. This result can be understood by considering that position generalization competency should be developed in the lower level of the network model, which does not interact directly with linguistic compositional processing.

Although several studies have investigated grounding of language with visuo-proprioceptive behavior using recurrent neural network models (*26*–*30*, *49*), few have addressed the mechanism underlying compositionality. To the best of our knowledge, there have been no models that examined the compositional nature of language grounded in visuo-proprioceptive behavior using the AIF framework. The current study explored a possible underlying mechanism for linguistic compositionality, codeveloped with sensorimotor skills to manipulate objects, using AIF. Extended studies should investigate how this mechanism can be developed gradually through incremental learning, as human children do.

By conducting ablation studies, we found that generalization performance in learning is significantly reduced when either VWM or visual attention is deleted from the model. This can be explained by considering that in the current model, visual image is perceived on the basis of structures rather than as simple pixel patterns by means of visual attention and working memory mechanisms. Actually, previous work (*39*) using a similar visual network showed that coupled mechanisms of visual attention and working memory enable a manipulated object to be segmented from the background. Therefore, it is highly likely that the marriage of structural visual information processing and compositional linguistic information processing enhances generalization in learning in the current task. An analysis of the model's performance in inferring linguistically represented goals from the observation of visuo-proprioceptive sequences showed results analogous to those obtained for inferring visuo-proprioceptive plan sequences from provided linguistic goals.

According to Hupkes *et al.* (*6*), compositionality of a neural network model should satisfy several characteristics. Systematicity means that the model should systematically combine known parts and rules. Productivity refers to the model's ability to extend its predictions beyond the lengths observed during training. Substitutivity ensures that the model's predictions are robust to synonymous substitutions. Localism concerns the degree to which the meaning of a compositional expression depends on its immediate, local structures rather than global structures. Overgeneralization means that

Table 2. Ablation study prediction error.

	Visual error (μ ± SD) %		Proprioceptive error (μ ± SD) %	
Condition	**U-P**	**U-C**	**U-P**	**U-C**
VWM-1 and 2 with attention	0.0196 ± 0.0016	0.0249 ± 0.0038	0.0110 ± 0.0014	0.0117 ± 0.0020
Only VWM-1 with attention	0.0407 ± 0.0041	0.0543 ± 0.0011	0.0136 ± 0.0008	0.0158 ± 0.0014
Only VWM-2 with attention	0.0391 ± 0.0031	0.0506 ± 0.0074	0.0149 ± 0.0002	0.0155 ± 0.0015
VWM-1 & 2 with no attention	0.0480 ± 0.0050	0.0657 ± 0.0046	0.0154 ± 0.0007	0.0169 ± 0.0075
No VWM-1 & 2 and no attention	0.0734 ± 0.0052	0.0960 ± 0.0068	0.0198 ± 0.0004	0.0238 ± 0.0037

the model should avoid favoring particular rules or exceptions during training and should not overgeneralize. We see evidence that our model demonstrates at least a rudimentary level of systematicity through its ability to generate appropriate visuo-proprioceptive sequence plans for U-Cs of actions and nouns, effectively generalizing to novel scenarios. This capability confirms that the model can synthesize new, meaningful sequences from previously learned parts and rules, meeting a key criterion for compositionality. Although testing for every aspect of compositionality is beyond the scope of the current study, it is a promising avenue for future research.

We showed that generalization performance in learning unseen compositions increases as the size of the vocabulary and variations in task composition increase, evidenced by the best performance seen in the largest group (5 × 8 composition). This result offers a minimal potential solution to the poverty of stimulus problem (*50*). If the dimensions of composition increase to include not only verbs and object nouns but also various modifiers, such as adverbs and adjectives, and each dimension consists of hundreds of elements, as we experience in daily life, covering all possible combinations across all dimensions would result in a combinatorial explosion. Faced with this problem, our expectation is that the required amount of experience for generalization in learning is not proportional to the product of the number of elements across all dimensions but rather is proportional to their summation, provided that the elemental size of each dimension is relatively large. Future studies should evaluate this possibility, which aims beyond generalization shown by typical neural network models, by conducting the same experiments under drastically scaled settings. The current study was necessary to examine basic mechanisms accounting for how generalization can be achieved in language-behavior compositionality with rigorous analysis, before scaling up the system. In addition, it aimed to investigate how the model's capability improves with increments in sizes of individual elements in each dimension.

Previous work with PV-RNN–based models (*51*, *38*) used the online error regression scheme, where prediction error serves as an input feature to retroactively correct the history of predictions (postdiction) while simultaneously predicting future behavior. This approach, although effective, is computationally intensive and relies on low-dimensional input features for feasibility. This limitation is especially notable given that our model evaluations were not conducted on physical robots. Although the proposed model shows competitive performance in generating appropriate visuo-proprioceptive trajectories compared with ground truth trajectories, executing the generated trajectories with a real robot may not yield the desired level of performance. Inferring object positions accurately from 64 pixel–by–64 pixel RGB (red-green-blue) images in the video can result in large errors in the real world. A single-pixel error from noise in the 64 pixel–by–64 pixel RGB image can result in a position error of several centimeters, which will substantially affect the behavioral performance of the robot, especially when the robot attempts to grasp objects. If a 256 pixel–by–256 pixel RGB image can be used, the position error could be reduced to <1 cm. This scheme, however, prohibits real-time computation (it takes several minutes to generate a single visuomotor plan trajectory), because expensive back-propagation through time computation should be conducted through the convLSTM. Future studies should investigate more efficient solutions to speed up this part of the computation, e.g., developing a C++ compiler for the whole system instead of using the current Python-based program to achieve real-time operation of physical robots using the proposed model. Upon solving these issues, the model will have the potential to be scaled up to more complex tasks with rich linguistic descriptions for cognitive robots to interact with the real world. We are actively working on this problem to make future iterations of the model more computationally efficient.

Models like CLIP (contrastive language–image pretraining) (*52*) and CLIP-guided generative latent space search (CLIP-GLaSS) (*53*) learn to associate images and textual descriptions in a joint embedding space. They use a contrastive learning objective to align embeddings of images and their corresponding textual descriptions. Although our model may share some features with these approaches, the **PB** vector is not a shared embedding for behavior and language. Instead, it acts as a bottleneck to constrain both visuo-proprioceptive sequences and word sequences such that they share similar structures. Our rationale for not using a shared embedding space for visuo-proprioceptive behavior and language is to maintain flexibility in learning behavioral patterns that can achieve the same linguistic goals. For example, the robot may learn multiple trajectories to achieve the linguistically represented goal of "put red on green" depending on object positions.

A major limitation of the current study is the absence of communication in a societal context. Tasks were executed by a single robotic arm, lacking active engagement with other agents, which is essential for language development in humans (*22*). The model operates in a small workspace with a limited vocabulary. One possible way to scale the model is to increase the number of compositional elements (adverbs, adjectives, conjunctions, etc.) to form longer sentences, as discussed previously. Moreover, extending the model to incorporate multiple agents, such as in RT-2 (*7*), each with their own models, could facilitate examination of communication dynamics within a societal context.

Recent advances in LLMs have shown incredible performance with robots working in real-world environments (*54*–*57*). These models, equipped with sophisticated language-processing capabilities, have enabled robots to comprehend and generate human-like language, facilitating seamless interaction with users and enhancing their overall functionality. However, it is important to note a substantial difference in the way language is acquired by these models compared with humans. Although humans develop language skills through interaction with their environment, as we have tried to emulate in our model by intermingling linguistic cues with physical experiences and sensory inputs, the language capabilities of LLMs are predominantly acquired through passive exposure to vast linguistic datasets. The extent to which LLMs can truly understand language in a human-like manner (*58*–*61*) is an intriguing question. As robotics continues to advance, bridging the gap between language understanding in machines and humans remains a key research challenge. Efforts to incorporate embodied interaction and sensorimotor experiences into language learning processes for robots hold promise to enhance the naturalness and robustness of their linguistic abilities in real-world scenarios.

MATERIALS AND METHODS

We propose a hierarchically organized generative model to handle multiple modalities, including vision, proprioception, and language. The model was trained end to end through supervised learning, using training examples containing visuo-proprioceptive sequences paired with corresponding linguistic expressions. The model learned to generate visuo-proprioceptive sequences corresponding to associated linguistic

expressions by minimizing evidence free energy. The trained model was used to generate goal-directed plans in which the goal was represented by linguistic expression. Goal-directed visuo-proprioceptive sequences were generated by minimizing expected free energy by means of AIF. This model inherited the vision module, with multiple VWM modules and visual attention (*37*). To integrate language into this model, we used the **PB** scheme proposed by Sugita and Tani (*24*). For all experiments, the model was trained and evaluated with five random seeds, as described in the "Implementation details" section of the Supplementary Materials.

Model architecture

The overall architecture of the current model is illustrated in Fig. 1A. The model consists of RNN-based generative networks that handle prediction of vision, proprioception, and language. These modalities are integrated by the associative network. There are three layers of stacked LSTM and convLSTM for the proprioception and vision networks, respectively. The language network is implemented as a single-layer LSTM with a **PB** vector (Fig. 6). The associative network is a single-layer PV-RNN that connects to the top layer of the visual and proprioceptive networks. Language is bound to the associative network through a binding loss between the linguistic **PB** vectors and $\widetilde{\mathrm{PB}}_t$ that is generated by the associative network (Fig. 6).

Each layer in the visuo-proprioceptive pathway receives contextual information from neighboring layers. Top-down connectivity provides a signal from the subsequent higher-level layer or from the associative layer of the model. Those layers propagate the prediction or belief of the network down to the sensorimotor level. A deconvolution operation is applied in the visual pathway, with dimensionality increasing from top to bottom. Visual and proprioceptive LSTM cells on the same layer of the model are connected via lateral connections. As in top-down processing, a deconvolution operation is applied to expand the low-dimensional space of proprioceptive representations to match the dimensions of the feature space of convLSTMs. Bottom-up connectivity sends neural activation from the lower layer of the model or the current sensory input, i.e., vision or proprioception, into the subsequent higher-level layer. Visual input is processed by an attention module, and a convolution operation is applied to reduce the dimension of projections to the next higher layer. Visual processing incorporates visual attention and saving/

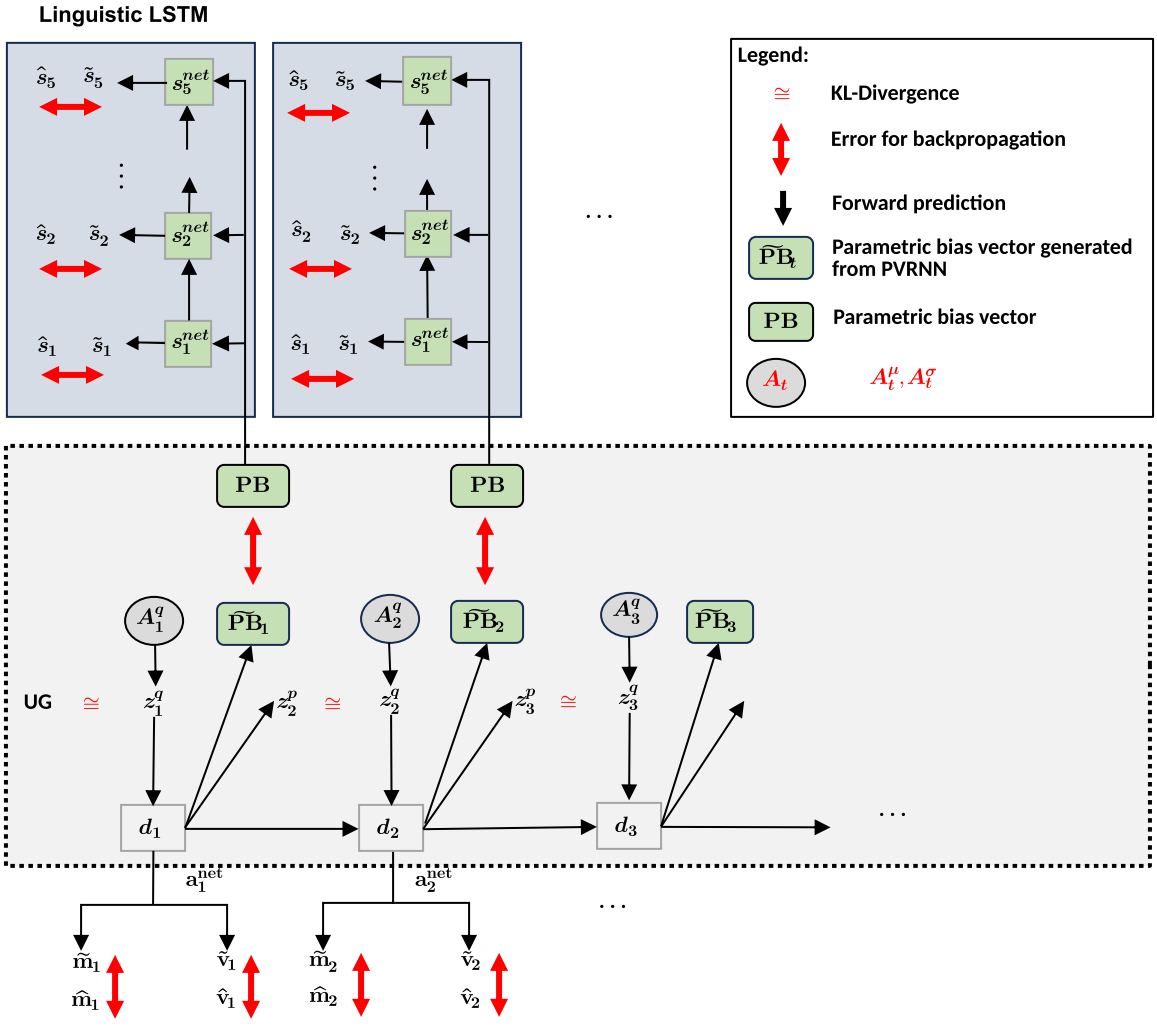

Fig. 6. Graphical representation of the associative PV-RNN and linguistic LSTM.

reading of visual images, using VWM. These are performed autonomously as part of the inference process.

The proprioceptive network predicts sequences of joint angles ($\tilde{\mathbf{m}}_t$) and multiple low-dimensional control signals (α_t^{att} and α_t^{M2}). These control signals act as parameters of visual attention and visual image transformation (*39*, *62*) to modulate information flow in the visual system. Visual attention performs dynamic adjustment of the pixel density in different regions of the image generated by the visual network. This allows the model to focus on and predict the visual appearance of manipulated objects in greater detail, whereas static parts of the generated images can be retrieved from the VWM-1. Furthermore, an additional parametric control of pixel-wise transformation (*62*) of images stored in VWM-2 allows the model to imitate dynamic changes of object images during its manipulation. This parametric control is performed by the output of proprioceptive LSTM (α_t^{M2}). This transformation is limited to affine image transformations considering the nature of the task used in our experiments.

The visual network predicts pixel images of the currently attended region ($\mathbf{v}_t^{att}$) and a set of masks that are used to mix the predicted image with the contents of each VWM (see Fig. 1A). The final visual image is predicted through the interaction of this convLSTM prediction with parameterized visual image operations, including attention, inverse attention, fusion, and transformation. Attention is performed by application of the current attention filter, parameters of which are predicted by the proprioceptive LSTM on the plain visual image. Inverse attention is simply an inverse transformation of the attended image. The fusion operations (denoted by "F" in a gray circle) fuse two sources of visual streams with a pixel-wise mixing ratio using outputs and corresponding masks generated from the multilayer convLSTM. Fusion operations were used to compose the final prediction, as well as to update the VWMs.

Generation of top-down signals from the associative PV-RNN is based on the adaptive parameter **A** (representing the approximate posterior probability distribution in terms of the mean and SD at each time step), which is optimized together with network parameters during training, to minimize the evidence free energy. The language network predicts a short sequence of words ($\tilde{\mathbf{S}}$) at each time step of the visuo-proprioceptive sequence. Each word ($\tilde{\mathbf{s}}_i$) is represented by a one hot vector, and the corpus is limited to 20 possible words. Language prediction depends on parameters of the linguistic-LSTM and the **PB** vector, which is constrained by the predicted $\widetilde{\mathrm{PB}}_t$ from the associative PV-RNN through a binding loss (equation 30 in the Supplementary Materials) (Fig. 6).

During learning, updating adaptive variables and connectivity weights of the network is performed to minimize the reconstruction error between prediction and observation. To this end, back-propagation of the error is performed inversely through the aforementioned top-down and bottom-up pathways to update values of the adaptive latent variables **A** and **PB**. In summary, the model represents a variational Bayes generative model in which learned multimodal spatiotemporal sequence patterns, including proprioception, vision, and language, can be reconstructed by adequately inferring the corresponding probabilistic latent variables **A** and **PB***s*. This is possible because all fabricated functions, such as visual attention, mask operation, and affine transformation, are designed as differentiable functions.

Visuo-proprioceptive and linguistic streams

The lowest layers of vision, proprioception, and language modules receive corresponding sensory inputs, pixel-based images $\mathbf{v}_t$, the softmax representation of the current joint angle configuration $\mathbf{m}_t$, and a series of one-hot vectors as linguistic input $\mathbf{s}_i$. The computation in the input layers of the network can be defined as

$$\mathbf{v}_{l=0,t}^{net} = \mathrm{ATT}\left(\mathbf{v}_t, \alpha_t^{att}\right) \tag{1}$$

$$\mathbf{m}_{l=0,t}^{net} = \mathrm{SoftMax}\left(\mathbf{m}_t\right) \tag{2}$$

$$\mathbf{s}_{l=0,i}^{net} = \mathbf{s}_i \tag{3}$$

with visual attention transformation $\mathrm{ATT}\left(\mathbf{v}_t, \alpha_t^{att}\right)$ parameterized by α_t^{att} applied to the visual input. We use the suffix i to denote individual word steps of the language. It is important to note that although vision and proprioception are synchronized and share the same number of time steps, language is expressed as a sentence in which each word is represented by a one-hot vector. Therefore, linguistic prediction, limited to five word steps, is predicted by the model at each step of the visuo-proprioceptive sequence (Fig. 6).

The proprioceptive prediction $\mathbf{m}_t^{net}$ as well as the low-dimensional parameterizations α_t^{att} and α_t^{M2}, which modulate attention and the affine transformation of VWM-2, are generated from the hidden states of the proprioceptive pathway as

$$\mathbf{m}_t^{net} = \mathrm{FFN}\left(\mathbf{m}_{l=1,t}^{net}\right) \tag{4}$$

$$\alpha_t^{att} = \mathrm{FFN}\left(\mathbf{m}_{l=1,t}^{net}\right) \tag{5}$$

$$\alpha_t^{M2} = \mathrm{FFN}\left(\mathbf{m}_{l=1,t}^{net}\right) \tag{6}$$

with FFN denoting a fully connected feed-forward network. Note that the FFNs do not share the same connectivity weights.

Neural activation in the visual pathway (stacked convLSTM) for layer $l = 1$ to $l = L$ at time step t is defined as

$$\mathbf{v}_{l,t}^{net} = \begin{cases} \mathrm{ConvLSTM}\left(\mathbf{v}_{l-1,t}^{net}, \mathbf{m}_{l,t-1}^{net}, \mathbf{a}_{t-1}^{net}\right), & \text{if } l = L \\ \mathrm{ConvLSTM}\left(\mathbf{v}_{l-1,t}^{net}, \mathbf{m}_{l,t-1}^{net}, \mathbf{v}_{l+1,t-1}^{net}\right), & \text{otherwise} \end{cases} \tag{7}$$

Neural activation in the proprioceptive pathway (stacked LSTM) is defined as

$$\mathbf{m}_{l,t}^{net} = \begin{cases} \mathrm{LSTM}\left(\mathbf{m}_{l-1,t}^{net}, \mathbf{v}_{l,t-1}^{net}, \mathbf{a}_{t-1}^{net}\right), & \text{if } l = L \\ \mathrm{LSTM}\left(\mathbf{m}_{l-1,t}^{net}, \mathbf{v}_{l,t-1}^{net}, \mathbf{m}_{l+1,t-1}^{net}\right), & \text{otherwise} \end{cases} \tag{8}$$

The model uses only one LSTM layer with **PB** for the language network. Its neural activation is defined as

$$\mathbf{s}_i^{net} = \mathrm{LSTM}\left(\mathbf{s}_{i-1}^{net}, \mathrm{PB}\right) \tag{9}$$

Associative network

As previously mentioned, the visual and proprioceptive pathways are connected by lateral connections in each layer of the convLSTM and LSTM blocks, respectively. In addition, the model includes an associative PV-RNN for a combined representation of both pathways in the highest layer. PV-RNN is composed of deterministic **d** and stochastic **z** latent variables. The PV-RNN generates predictions from a prior distribution **p** and infers an approximate posterior distribution **q** by means of prediction error minimization on the generated sensory output **x**. The prior generative model $\mathbf{p}_\theta$ is factorized as shown in following equation.

$$\begin{aligned}&\mathbf{p}_\theta\left(\mathbf{x}_{1:T}, \mathbf{d}_{1:T}, \mathbf{z}_{1:T} \mid \mathbf{d}_0\right)\\&=\prod_{t=1}^{T}\left(\mathbf{p}_{\theta_\mathbf{x}}\left(\mathbf{x}_t \mid \mathbf{d}_t\right)\mathbf{p}_{\theta_\mathbf{d}}\left(\mathbf{d}_t \mid \mathbf{d}_{t-1}, \mathbf{z}_t\right)\mathbf{p}_{\theta_z}\left(\mathbf{z}_t \mid \mathbf{d}_{t-1}\right)\right)\end{aligned} \tag{10}$$

The prior distribution $\mathbf{p}_{\theta_z}\left(\mathbf{z_t} \mid \mathbf{d}_{t-1}\right)$ is a Gaussian, and it depends on $\mathbf{d}_{t-1}$, except at the initial time step $t = 1$, which is fixed as a unit Gaussian with zero mean. Each sample of the prior distribution $\mathbf{z}_t^p$ is computed as shown in the following equation.

$$\begin{aligned}\mu_t^p &= \begin{cases}0, & \text{if } t=1\\ \tanh\left(\mathbf{W}_{\mathbf{d},\mathbf{z}^{\mu^p}}\mathbf{d}_{t-1}\right), & \text{otherwise}\end{cases}\\ \sigma_t^p &= \begin{cases}1, & \text{if } t=1\\ \exp\left(\mathbf{W}_{\mathbf{d},\mathbf{z}^{\sigma^p}}\mathbf{d}_{t-1}\right), & \text{otherwise}\end{cases}\\ \mathbf{z}_t^p &= \mu_t^p + \sigma_t^p * \epsilon\end{aligned} \tag{11}$$

where ϵ is a random noise sample such that $\epsilon \sim \mathcal{N}(0, I)$. **W** is the connectivity weight matrix. We omit the bias term in all equations for the sake of brevity.

Given that computing the true posterior distribution is intractable, the model infers an approximate posterior ($\mathbf{q}_\varphi$) at time step t, $\mathbf{z}_t^q$ computed as shown in Eq. 12. $\mathbf{A}_{1:T}^{\mu}, \mathbf{A}_{1:T}^{\sigma}$ are adaptive variables, inferred through back-propagation by minimizing the prediction error and the complexity term, as detailed below. These are used to compute the mean and SD for the approximate posterior at each step in a sequence.

$$\begin{aligned}\mu_t^q &= \tanh\left(\mathbf{A}_t^{\mu}\right),\\ \sigma_t^q &= \exp\left(\mathbf{A}_t^{\sigma}\right),\\ \mathbf{z}_t^q &= \mu_t^q + \sigma_t^q * \epsilon\end{aligned} \tag{12}$$

We use a multiple timescale recurrent neural network (*63*), adapted for a single layer, as the RNN for the associative PV-RNN layer. The deterministic latent variable of the associative layer $\mathbf{a}_t^{\text{net}}$ is computed as follows

$$\begin{aligned}\mathbf{d}_t &= \left(1-\frac{1}{\tau}\right)\mathbf{d}_{t-1}\\ &+\frac{1}{\tau}\left(\mathbf{W}_{a,a}\mathbf{a}_{t-1}^{\text{net}} + \mathbf{W}_{z,a}\mathbf{z}_t^q + \mathbf{W}_{v,a}\mathbf{v}_{l=L,t-1}^{\text{net}} + \mathbf{W}_{m,a}\mathbf{m}_{l=L,t-1}^{\text{net}}\right)\end{aligned} \tag{13}$$

$$\mathbf{a}_t^{\text{net}} = \tanh\left(\mathbf{d}_t\right) \tag{14}$$

where τ is the time constant that determines the rate at which the network integrates information over time. The associative layer also predicts a parametric bias ($\widetilde{\text{PB}}_\mathbf{t}$) vector at each time step, which is bound to the **PB** vector of the language module through a binding loss (equation 30). The $\widetilde{\text{PB}}_\mathbf{t}$ vector is computed as

$$\widetilde{\text{PB}}_\mathbf{t} = \tanh\left(\mathbf{W}_{d,pb}\mathbf{d}_t\right) \tag{15}$$

$\mathbf{W}_{d,pb}$ is the connectivity weight matrix between **d** and **PB**.

Visual attention and working memory

Our previous study (*39*) showed that visual image transformations by attention and inverse attention are among the most important elements for successful development of VWM function during end-to-end learning. As mentioned previously, visual attention is performed by an attention transformation parameterized by scaling and coordinates of a focal position. These parameters are generated by the proprioceptive multilayer LSTM, which receives top-down signals from the associative PV-RNN in the higher level. This means that optimal parameters for visual attention during training and goal-directed planning are determined by the inference of optimal latent-state values $\mathbf{A}_{1:T}$.

The visual attention and VWM systems are applied to the output of the convLSTM block, which includes prediction of the attended visual image $\mathbf{v}_t^{\text{net}}$ and a set of masks, computed as

$$\mathbf{v}_t^{\text{net}} = \tanh\left(\text{Deconv}\left(\mathbf{v}_{l=1,t}^{\text{net}}\right)\right) \tag{16}$$

$$\begin{bmatrix}\mathbf{g}_t^{\text{M1}}\\ \mathbf{g}_t^{\text{pred}}\end{bmatrix} = \text{ATT}^{-1}\left(Sig\left(\text{Deconv}\left(\mathbf{v}_{l=1,t}^{\text{net}}\right)\right), \alpha_t^{\text{att}}\right) \tag{17}$$

$$\begin{bmatrix}\mathbf{g}_t^{\text{M2}}\\ \mathbf{g}_t^{\text{net}}\end{bmatrix} = Sig\left(\text{Deconv}\left(\mathbf{v}_{l=1,t}^{\text{net}}\right)\right) \tag{18}$$

with a sigmoidal activation function *Sig*. Note that the Deconv operations do not share the same connectivity weights. The masks $\mathbf{g}_t^{\text{M1}}$ and $\mathbf{g}_t^{\text{M2}}$ modulate the pixel-wise update of the VWM-1 and VWM-2, respectively. Furthermore, the masks $\mathbf{g}_t^{\text{pred}}$ and $\mathbf{g}_t^{\text{net}}$ decide how much the final visual prediction depends on the VWMs or $\mathbf{v}_t^{\text{net}}$ (see Fig. 1A).

Details of network-wise operations for VWM-1 and VWM-2 are described by the following equations

$$\begin{aligned}\mathbf{vwm}_{t+1}^{\text{M1}} &= \left(1-\mathbf{g}_t^{\text{M1}}\right)\odot\mathbf{vwm}_t^{\text{M1}}\\ &+\mathbf{g}_t^{\text{M1}}\odot\text{ATT}^{-1}\left(\mathbf{v}_t^{\text{att}}, \alpha_t^{\text{att}}\right)\end{aligned} \tag{19}$$

Equation 19 describes how contents of VWM-1 ($\text{vwm}_{t+1}^{\text{M1}}$) can be updated, where $\mathbf{g}_t^{\text{M1}}$ denotes a pixel-wise mask and ATT^{-1} performs inverse attention transformation, parameterized by α_t^{att} (Eq. 5), on the predicted attended visual image $\mathbf{v}_t^{\text{att}}$. The element-wise multiplication operator denoted by the symbol $\odot$ fuses the visual stream and mask.

$$\begin{aligned}\mathbf{vwm}_{t+1}^{\text{M2}} &= \mathbf{g}_t^{\text{M2}}\odot TRAN\left(\mathbf{vwm}_t^{\text{M2}}, \alpha_t^{\text{M2}}\right)\\ &+\left(1-\mathbf{g}_t^{\text{M2}}\right)\odot\mathbf{v}_t^{\text{att}}\end{aligned} \tag{20}$$

Equation 20 describes how VWM-2, vwm^{M2}, can be updated. The variable $\mathbf{g}_t^{\text{M2}}$ denotes a pixel-wise mask that defines the fusion of

transformed contents $TRAN(\mathbf{vwm}_t^{M2}, \alpha_t^{M2})$ of VWM-2 with $\mathbf{v}_t^{att}$, the predicted image in the attended feature space from the previous step. This ensures that the contents of $\mathbf{vwm}^{M2}$ are influenced only by the visual prediction in the attended region and by the transformed vwm^{M2}. Prediction $\mathbf{v}_t^{att}$ of attended visual images is performed by a fusion of the predictions made by the convLSTM, $\mathbf{g}_t^{net}$, and the contents of VWM-2, defined by

$$\mathbf{v}_t^{att} = \mathbf{g}_t^{net} \odot \mathbf{v}_t^{net} + \left(1 - \mathbf{g}_t^{net}\right) \odot TRAN\left(\mathbf{vwm}_t^{M2}, \alpha_t^{M2}\right) \quad (21)$$

Model output

The visual output $\tilde{\mathbf{v}}_t$ is computed by fusion of the contents of VWM-1, $\mathbf{vwm}_t^{M1}$, with the predicted attended image $\mathbf{v}_t^{att}$. Inverse attention transformation ATT^{-1} is applied to $\mathbf{v}_t^{att}$ to make it possible to fuse with the contents VWM-1 defined as

$$\tilde{\mathbf{v}}_t = \mathbf{g}_t^{pred} \odot ATT^{-1}\left(\mathbf{v}_t^{att}, \alpha_t^{att}\right) + \left(1 - \mathbf{g}_t^{pred}\right) \odot \mathbf{vwm}_t^{M1} \quad (22)$$

The final proprioceptive prediction is generated by a decoding of the softmax encoded predictions of the LSTM to get the trajectory, $\tilde{\mathbf{m}}_t$, of joint angle configurations

$$\tilde{\mathbf{m}}_t = \text{SoftMax}^{-1}\left(\mathbf{m}_t^{net}\right) \quad (23)$$

The final linguistic prediction, $\tilde{\mathbf{s}}_i$, is computed through a fully connected output layer followed by a softmax activation function to get a one-hot vector representation for each word

$$\tilde{\mathbf{s}}_i = \text{SoftMax}\left(\text{FFN}\left(\mathbf{s}_i^{net}\right)\right) \quad (24)$$

The sentence predicted at every time step of the behavior is the same and is defined as $\tilde{\mathbf{S}} = \left(\tilde{\mathbf{s}}_1, \tilde{\mathbf{s}}_2, \tilde{\mathbf{s}}_3, \tilde{\mathbf{s}}_4, \tilde{\mathbf{s}}_5\right)$. Examples of sentences describing actions performed by the robot are "put red on green," "grasp red," "move red left," etc., where each word is represented by one hot vector. Note that the sentences are of different lengths; therefore, for all sentences to be the same length, the remaining steps are masked with zero vectors to get a maximum of five word steps. The final character, $\tilde{\mathbf{s}}_5$, of every sentence is always the vector corresponding to ".", indicating the end of the sentence. Details on how training and inference are done by free energy minimization are provided in the Supplementary Materials.

Supplementary Materials

This PDF file includes:

Supplementary Methods
Supplementary Results
Figs. S1 to S11
Tables S1 to S6
Algorithms 1 to 3
References (*64*–*66*)

REFERENCES AND NOTES

1. N. Chomsky, *Syntactic Structures* (Mouton and Co., 1957).
2. G. Evans, *The Varieties of Reference* (Oxford Univ. Press, 1982).
3. G. Frege, *Collected Papers on Mathematics, Logic, and Philosophy* (Wiley-Blackwell, 1991).
4. T. Janssen, Frege, contextuality and compositionality. *J. Logic. Lang. Info.* **10**, 115–136 (2001).
5. 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).
6. D. Hupkes, V. Dankers, M. Mul, E. Bruni, Compositionality decomposed: How do neural networks generalise? *J. Artif. Intel. Res.* **67**, 757–795 (2020).
7. C. Lynch, A. Wahid, J. Thompson, T. Ding, J. Betker, R. Baruch, T. Armstrong, P. Florence, Interactive language: Talking to robots in real time. arXiv:2210.06407 [cs.RO] (2022).
8. S. Nolfi, On the unexpected abilities of large language models. arXiv:2308.09720 [cs.AI] (2023).
9. M. Abdou, A. Kulmizev, D. Hershcovich, S. Frank, E. Pavlick, A. Sogaard. Can language models encode perceptual structure without grounding? A case study in color. arXiv:2109.06129 [cs.CV] (2021).
10. S. Yousefi, L. Betthauser, H. Hasanbeig, R. Millière, I. Momennejad, Decoding in-context learning: Neuroscience-inspired analysis of representations in large language models. arXiv:2310.00313 [cs.CL] (2024).
11. E. Pavelick, Symbols and grounding in large language models. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **381**, 20220041 (2023).
12. K. J. Friston, The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* **11**, 127–138 (2010).
13. T. Cameron-Faulkner, E. Lieven, M. Tomasello, A construction based analysis of child directed speech. *Cogn. Sci.* **27**, 843–873 (2003).
14. M. Tomasello, "The usage-based theory of language acquistion" in *The Cambridge Handbook of Child Language*, E. L. Bavin, Ed. (Cambridge Univ. Press, 2009), pp. 69-87.
15. L. Smith, M. Gasser, The development of embodied cognition: Six lessons from babies. *Artif. Life* **11**, 13–29 (2005).
16. J. Piaget, *The Language and Thought of the Child* (Meridian, 1955).
17. G. Buccino, L. Riggio, G. Melli, F. Binkofski, V. Gallese, G. Rizzolatti, Listening to action-related sentences modulates the activity of the motor system: A combined tms and behavioral study. *Cogn. Brain Res.* **24**, 355–363 (2005).
18. F. R. Dreyer, F. Pulvermüller, Abstract semantics in the motor system?—An event-related fMRI study on passive reading of semantic word categories carrying abstract emotional and mental meaning. *Cortex* **100**, 52–70 (2018).
19. F. Pulvermüller, L. Fadiga, Active perception: Sensorimotor circuits as a cortical basis for language. *Nat. Rev. Neurosci.* **11**, 351–360 (2010).
20. P. Oudeyer, G. Kachergis, W. Schueller, Computational and robotic models of early language development: A review. arXiv:1903.10246 [cs.CL] (2019).
21. S. Harnad, The symbol grounding problem. *Physica D.* **42**, 335–346 (1990).
22. M. Tomasello, The social bases of language acquisition. *Soc. Dev.* **1**, 67–87 (2006).
23. M. Tomasello, *First Verbs: A Case Study of Early Grammatical Development* (Cambridge Univ. Press, 2009).
24. L. B. Smith, S. Jayaraman, E. Clerkin, C. Yu, The developing infant creates a curriculum for statistical learning. *Trends Cogn. Sci.* **22**, 325–336 (2018).
25. L. Raggioli, A. Cangelosi, Embodied attention in word-object mapping: A developmental cognitive robotics model, *in 2022 IEEE International Conference on Development and Learning (ICDL)* (IEEE, 2022), pp. 156–163.
26. Y. Sugita, J. Tani, Learning semantic combinatoriality from the interaction between linguistic and behavioral processes. *Adapt. Behav.* **13**, 33–52 (2005).
27. A. Cangelosi, G. Metta, G. Sagerer, S. Nolfi, C. Nehaniv, K. Fischer, J. Tani, T. Belpaeme, G. Sandini, F. Nori, L. Fadiga, B. Wrede, K. Rohlfing, E. Tuci, K. Dautenhahn, J. Saunders, A. Zeschel, Integration of action and language knowledge: A roadmap for developmental robotics. *IEEE Trans. Auton. Mental Dev.* **2**, 167–195 (2010).
28. A. Cangelosi, F. Stramandinoli, A review of abstract concept learning in embodied agents and robots. *Philos. Trans. R. Soc. B* **373**, 20170131 (2018).
29. S. Heinrich, S. Wermter, Interactive natural language acquisition in a multi-modal recurrent neural architecture. *Connect. Sci.* **30**, 99–133 (2018).
30. A. Akakzia, C. Colas, P. Y. Oudeyer, M. Chetouani, O. Sigaud, Grounding language to autonomously-acquired skills via goal generation, poster presented at the Ninth International Conference on Learning Representations (ICLR), 3 to 7 May 2021; https://iclr.cc/virtual/2021/poster/3190.
31. T. Yamada, H. Matsunaga, T. Ogata, Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions. *IEEE Robot. Autom. Lett.* **3**, 3441–3448 (2018).
32. R. P. Rao, D. H. Ballard, Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).
33. K. J. Friston, S. Kiebel, Predictive coding under the free-energy principle. *Philos. Trans. R. Soc. B Biol. Sci.* **364**, 1211–1221 (2009).
34. K. J. Friston, J. Daunizeau, S. J. Kiebel, Reinforcement learning or active inference? *PLOS ONE* **4**, e6421 (2009).
35. K. J. Friston, J. Daunizeau, J. Kilner, S. J. Kiebel, Action and behavior: A free-energy formulation. *Biol. Cybern.* **102**, 227–260 (2010).
36. H. Brown, K. J. Friston, S. Bestmann, Active inference, attention, and motor preparation. *Front. Psychol.* **2**, 218 (2011).
37. T. Matsumoto, J. Tani, Goal-directed planning for habituated agents by active inference using a variational recurrent neural network. *Entropy* **22**, 564 (2020).
38. T. Matsumoto, W. Ohata, F. Benureau, J. Tani, Goal-directed planning and goal understanding by extended active inference: Evaluation through simulated and physical robot experiments. *Entropy* **24**, 469 (2022).
39. J. Queißer, M. Jung, T. Matsumoto, J. Tani, Emergence of content-agnostic information processing by a robot using active inference, visual attention, working memory, and planning. *Neural Comput.* **33**, 2353–2407 (2021).

40. S. R. Sehon, Goal-directed action and teleological explanation, in *Causation and Explanation*, Topics in Contemporary Philosophy, J. Keim Campbell, M. O'Rourke, H. S. Silverstein, Eds. (MIT Press, 2007), pp. 155–170.
41. G. Csibra, S. Bíró, O. Koós, G. Gergely, One-year-old infants use teleological representations of actions productively. *Cogn. Sci.* **27**, 111–133 (2003).
42. A. Ahmadi, J. Tani, A novel predictive-coding-inspired variational RNN model for online prediction and recognition. *Neural Comput.* **31**, 2025–2074 (2019).
43. S. Bernardi, M. K. Benna, M. Rigotti, J. Munuera, S. Fusi, C. D. Salzman, The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell* **183**, 954–967.e21 (2020).
44. T. Ito, T. Klinger, D. Schultz, J. Murray, M. Cole, M. Rigotti, Compositional generalization through abstract representations in human and artificial neural networks, in vol. 35 of *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh, Eds. (Curran Associates, 2022), pp. 32225–32239.
45. J. Tani, M. Ito, Self-organization of behavioral primitives as multiple attractor dynamics: A robot experiment. *IEEE Trans. Syst. Man Cybern. Part A: Syst. Humans* **33**, 481–488 (2003).
46. J. Tani, M. Ito, Y. Sugita, Self-organization of distributedly represented multiple behavior schemata in a mirror system: Reviews of robot experiments using RNNPB. *Neural Networks* **17**, 1273–1289 (2004).
47. B. Schölkopf, A. Smola, K.-R. Möller, Nonlinear Component analysis as a kernel eigenvalue problem. *Neural Comput.* **10**, 1299–1319 (1998).
48. B. L. Welch, The generalization of 'student's' problem when several different population variances are involved. *Biometrika* **34**, 28–35 (1947).
49. T. Taniguchi, Collective predictive coding hypothesis: Symbol emergence as decentralized Bayesian inference. *Front. Robot. AI.* **11**, 1353870 (2024).
50. N. Chomsky, Poverty of stimulus: Unfinished business. *Studies Chin. Linguistics* **33**, 3–16 (2012).
51. W. Ohata, J. Tani, Investigation of the sense of agency in social cognition, based on frameworks of predictive coding and active inference: A simulation study on multimodal imitative interaction. *Front. Neurorobot.* **14**, 61 (2020).
52. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision. arXiv:2103.00020 [cs.CV] (2021).
53. M. G. C. A. Cimino, F. A. Galatolo, G. Vaglini, Generating images from caption and vice versa via clip-guided generative latent space search, in *IMPROVE 2021: Proceedings of the International Conference on Image Processing and Vision Engineering* (ACM, 2021), pp. 166–174.
54. J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. L. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Bińkowski, R. Barreira, O. Vinyals, A. Zisserman, K. Simonyan, Flamingo: A visual language model for few-shot learning, in vol. 35 of *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh, Eds. (Curran Associates, 2022), pp. 23716–23736.
55. M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, K.-H. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao, J. Rettinghouse, D. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, M. Yan, A. Zeng, Do as I can, not as I say: Grounding language in robotic affordances. arXiv:2204.01691 [cs.RO] (2022).
56. D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, P. Florence, PaLM-E: An embodied multimodal language model. arXiv:2303.03378 [cs.LG] (2023).
57. A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, Tsang-Wei Edward Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, B. Zitkovich, Rt-2: Vision-language-action models transfer web knowledge to robotic control. arXiv:2307.15818 [cs.RO] (2023).
58. D. J. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (Oxford Paperbacks, 1997).
59. G. Marcus, E. Davis, Insights for AI from the human mind. *Commun. ACM* **64**, 38–41 (2021).
60. G. Pezzulo, T. Parr, P. Cisek, A. Clark, K. J. Friston, Generating meaning: Active inference and the scope and limits of passive AI. *Trends Cogn. Sci.* **28**, 97–112 (2024).
61. T. Yoshida, A. Masumori, T. Ikegami, From text to motion: Grounding gpt-4 in a humanoid robot "alter3." arXiv:2312.06571 [cs.RO] (2023).
62. M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, Spatial transformer networks, in vol. 28 of *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett, Eds. (Curran Associates, 2015), pp. 2017–2025.
63. F. Shibata Alnajjar, Y. Yamashita, J. Tani, The hierarchical and functional connectivity of higher-order cognitive mechanisms: Neurorobotic model to investigate the stability and flexibility of working memory. *Front. Neurorobot.* **7**, 335–346 (2013).
64. C. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics* (Springer-Verlag, 2006).
65. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization. arXiv:1412.6980 [cs.LG] (2017).
66. R. Pascanu, T. Mikolov, Y. Bengio, On the difficulty of training recurrent neural networks, in *Proceedings of the 30th International Conference on Machine Learning*, S. Dasgupta, D. McAllester, Eds. (MLResearchPress, 2013), pp. 1310–1318.

Acknowledgments: We are grateful for the help and support provided by the lab members in the Cognitive Neurorobotics Research Unit and the Scientific Computing section of the Research Support Division at OIST. We thank T. Matsumoto for providing the base code for data collection and S. Aird for editing the language of the manuscript. **Funding:** We thank Okinawa Institute of Science and Technology (OIST) Graduate University for supporting this work. J.T. was partially supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI, Transformative Research Area (A): unified theory of prediction and action (24H02175). **Author contributions:** P.V., J.F.Q., and J.T. designed the model. P.V. and J.T. designed the experiments. P.V. performed all experiments and data analysis. J.F.Q., S.V.F., and J.T. contributed to the interpretation of the results. P.V. wrote the paper, and J.F.Q., S.V.F., and J.T. edited it. J.T. supervised the study. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** The code used for experiments is available at Zenodo and accessible via https://doi.org/10.5281/zenodo.14359361. The dataset used for training and evaluation of the model are available at Dryad and accessible via https://doi.org/10.5061/dryad.xsj3tx9qc.

Submitted 25 March 2024
Accepted 17 December 2024
Published 22 January 2025
10.1126/scirobotics.adp0751

Science Robotics

Development of compositionality through interactive learning of language and action of robots

Prasanna Vijayaraghavan, Jeffrey Frederic Queißer, Sergio Verduzco Flores, and Jun Tani

Sci. Robot. **10** (98), eadp0751. DOI: 10.1126/scirobotics.adp0751

Science Robotics (ISSN 2470-9476) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Robotics* is a registered trademark of AAAS.