PNAS RESEARCH ARTICLE | NEUROSCIENCE COMPUTER SCIENCES OPEN ACCESS

Linking global top-down views to first-person views in the brain

Jinwei Xing[a], Elizabeth R. Chrastil[a,b], Douglas A. Nitz[c], and Jeffrey L. Krichmar[a,d,1]

Edited by Thomas Albright, Salk Institute for Biological Studies, La Jolla, CA; received February 3, 2022; accepted September 7, 2022

Humans and other animals have a remarkable capacity to translate their position from one spatial frame of reference to another. The ability to seamlessly move between top-down and first-person views is important for navigation, memory formation, and other cognitive tasks. Evidence suggests that the medial temporal lobe and other cortical regions contribute to this function. To understand how a neural system might carry out these computations, we used variational autoencoders (VAEs) to reconstruct the first-person view from the top-down view of a robot simulation, and vice versa. Many latent variables in the VAEs had similar responses to those seen in neuron recordings, including location-specific activity, head direction tuning, and encoding of distance to local objects. Place-specific responses were prominent when reconstructing a first-person view from a top-down view, but head direction–specific responses were prominent when reconstructing a top-down view from a first-person view. In both cases, the model could recover from perturbations without retraining, but rather through remapping. These results could advance our understanding of how brain regions support viewpoint linkages and transformations.

cognitive map | head direction cells | place cells | robotics | variational autoencoders

Significance

The ability to link between a first-person experience and a global map is an important cognitive function. To investigate the underlying computations needed to complete this task, we used variational autoencoders to reconstruct the top-down images from a robot's camera view, and vice versa. We observed that place-specific coding is more prevalent when linking a top-down view to a first-person view, and head direction selectivity is more prevalent in the other direction. In both cases, the system recovers from perturbations by population coding and instantaneous remapping. This modeling brings a fundamentally different approach to understanding transformations between perspectives and suggests testable predictions in the nervous system.

Humans are able to translate their location and navigational goals on an external map into decision-making behaviors in the environment. A glance at a map can help place you in your local surroundings. Conversely, when looking at one's local surroundings, one can place oneself on a global map. The ability to seamlessly move between top-down views (TDVs) and first-person views (FPVs) may be important for navigation and memory formation, as well as many cognitive tasks (e.g., building a cabinet from a plan drawn on paper, or finding an extra screw after the cabinet is constructed, and referring back to the plan to find out where the screw should go). Evidence from other animals suggests that they also have the ability to translate their position from one spatial reference frame to another (1–4). In particular, bats appear to have the ability to translate a TDV while flying above the landscape to an FPV when navigating on the ground or foraging for food (4–7).

Studies suggest that the entorhinal cortex (EC), retrosplenial cortex (RSC), subiculum (SUB), posterior parietal cortex (PPC), and hippocampus (HPC) could play significant roles in linking locations and orientations relative to one view to locations and orientations relative to another (1, 2, 8–12). The computations and neural implementations that manifest this cognitive ability have scarcely been addressed, despite numerous navigation experiments in humans and rodents. Computational modeling suggests that these transformations and linkages could be accomplished through specific encoding of parameters (8, 12, 13), or mixed selectivity that responds to multiple variables (14). However, it is unclear whether mechanisms for linkage and transformation among perspectives operate to form a single mapping of location from both perspectives, or serve to link analogous locations in two different mappings. Mapping of location and orientation is robustly observed in the rodent EC, HPC, and SUB (15–18), which provide input to RSC and other brain regions. The PPC provides egocentric information to the RSC (19). Furthermore, the visual system plays an important role in driving spatial activity (3). Still, the exact role of these brain regions and their neural computations, especially in the context of viewpoint transformations, remain poorly understood.

In the present article, we attempt to answer the following open questions: 1) What architectures might support these transformations and linkages? 2) What are the computations and neural implementations underlying linkages and transformations between TDVs and FPVs? 3) What cues or landmarks are required to make these transformations and linkages? To answer these questions, we take a model-free approach by using variational autoencoders (VAEs) to reconstruct the FPV from the TDV of a robot simulation, and vice versa (20). The latent variables, which make a transformation between the encoding network layers and the decoding network layers, will be compared with brain responses.

Author affiliations: [a]Department of Cognitive Sciences, University of California, Irvine, CA 92697; [b]Department of Neurobiology and Behavior, University of California, Irvine, CA 92697; [c]Department of Cognitive Science, University of California San Diego, La Jolla, CA 92093; and [d]Department of Computer Science, University of California, Irvine, CA 92697

Author contributions: E.R.C., D.A.N., and J.L.K. designed research; J.X. and J.L.K. performed research; J.X., E.R.C., D.A.N., and J.L.K. analyzed data; and J.X., E.R.C., D.A.N., and J.L.K. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: jkrichma@uci.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2202024119/-/DCSupplemental.

Published November 2, 2022.

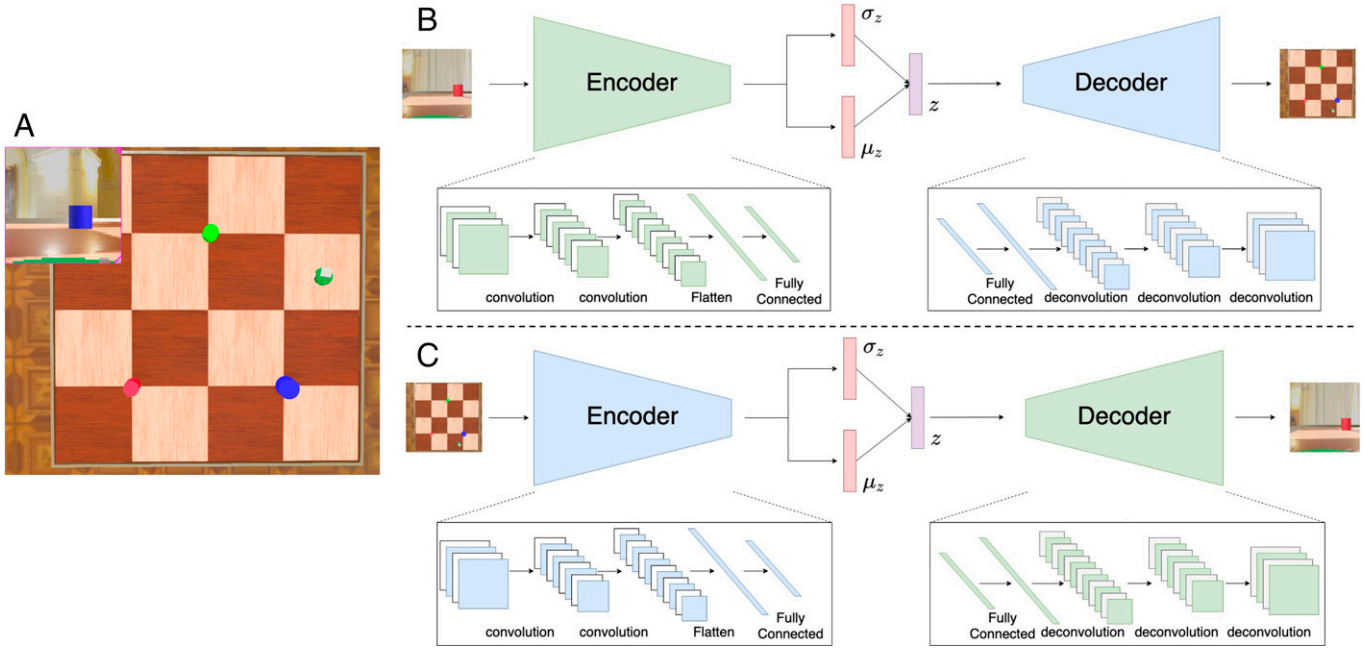

Fig. 1. Simulation setup and model architectures. (*A*) Robot freely explored a square arena, which had three colored cylinders. The robot is located on the middle right facing the blue cylinder. *Inset* shows the robot's camera view. Note that the camera view did not overlay the top-down images during data collection. The robot was simulated using Webots (21, 22). (*B* and *C*) VAEs reconstruct images from robot simulation. The latent variables between the encoder and decoder are analyzed to understand the transformations and linkages between views. (*B*) Takes an FPV as input and reconstructs a TDV. (*C*) Takes a TDV as input and reconstructs an FPV.

In contrast to neurobiologically inspired models of these transformations (8, 13), which suggest that the computations in each direction utilize the same circuit by inverting the transformation operation, our results indicate that each direction of transformation is dictated by different computations carried out by separate circuits. Whereas going from a TDV to an FPV required specific representations of place and objects, going from an FPV to a TDV tended to use mixed representations and strong head direction signaling. In both cases, one view was accurately reconstructed from the other. In addition, both neural codes were flexible and adaptive to perturbations. We suggest that this is a possible neural implementation that could support important navigation functions.

Results

Robot Simulation and Modeling Transformations. To test the ability to link TDV to FPV and vice versa, data were collected with the Webots (21, 22) robot simulation environment (Fig. 1*A*). The simulated robot was a Khepera with a camera, and proximity sensors to detect objects and boundaries. The robot freely explored its space. Approximately every second, the overhead view of simulation (TDV) and the robot's camera image (FPV), position, heading, and distance to the three cylinders were saved. Ten thousand data points were collected: 8,000 for training and 2,000 for testing. In all conditions, even when environmental conditions changed (e.g., removing an object or changing the background), the robot position, heading, and trajectory were identical for the 10,000 data points.

Two VAEs were constructed: one for reconstructing the TDV of the simulation environment from the FPV of the robot (Fig. 1*B*) and another for reconstructing the FPV from the TDV (Fig. 1*C*). The number of latent variables (μ, σ, and z in Fig. 1 *B* and *C*) varied from 30 to 50 to 100. In the results presented below, only the 2,000 testing data points were used for analysis.

The VAE was able to reconstruct a TDV from an FPV, and vice versa. Fig. 2 shows how the reconstructions improved as the loss decreased during training with 100 latent variables. After 20,000 epochs of training, the median reconstruction losses for the FPV to TDV and the TDV to FPV transformations were less than 0.01. The process was repeated five times with different random number generator seeds. All five runs for each number of latent variables were used for analysis. *SI Appendix*, Figs. S1–S3 shows the loss for simulations with 30, 50, and 100 latent variables. Although the medians were roughly similar for both types of transformations (e.g., with 100 latent variables, the median was 0.0078 for FPV to TDV and 0.0084 for TDV to FPV in *SI Appendix*, Fig. S3), the TDV to FPV transformation had more outliers (i.e., images it had difficulty reconstructing). Because of this, the two distributions for all numbers of latent variables were significantly different ($p < 0.000001$; Wilcoxon signed-rank test). *SI Appendix*, Figs. S4 and S5 shows examples of the reconstructions after training.

Spatial Representations in Latent Variables. We wondered whether the latent variables of the VAEs had similar qualities to spatial representations observed in RSC, HPC, and SUB recordings. Indeed, many latent variables were sensitive to place, heading, and objects. We looked at how well a latent variable correlated with the robot's distance to a cylinder, to an idealized head direction cell (cosine tuning curve with one of 16 preferred directions), or to an idealized place cell (two-dimensional [2D] Gaussian with one of 16 preferred locations). Table 1 shows the percentage of significant correlations in each case. The TDV to FPV transformations had significantly more latent variables that were strongly correlated with distance to the cylinder objects, and significantly more latent variables strongly correlated with place fields than the FPV to TDV transformations. In contrast, the FPV to TDV transformations had significantly more latent variables sensitive to head direction than the TDV to FPV transformation. *SI Appendix*, Fig. S6, which plots the correlations

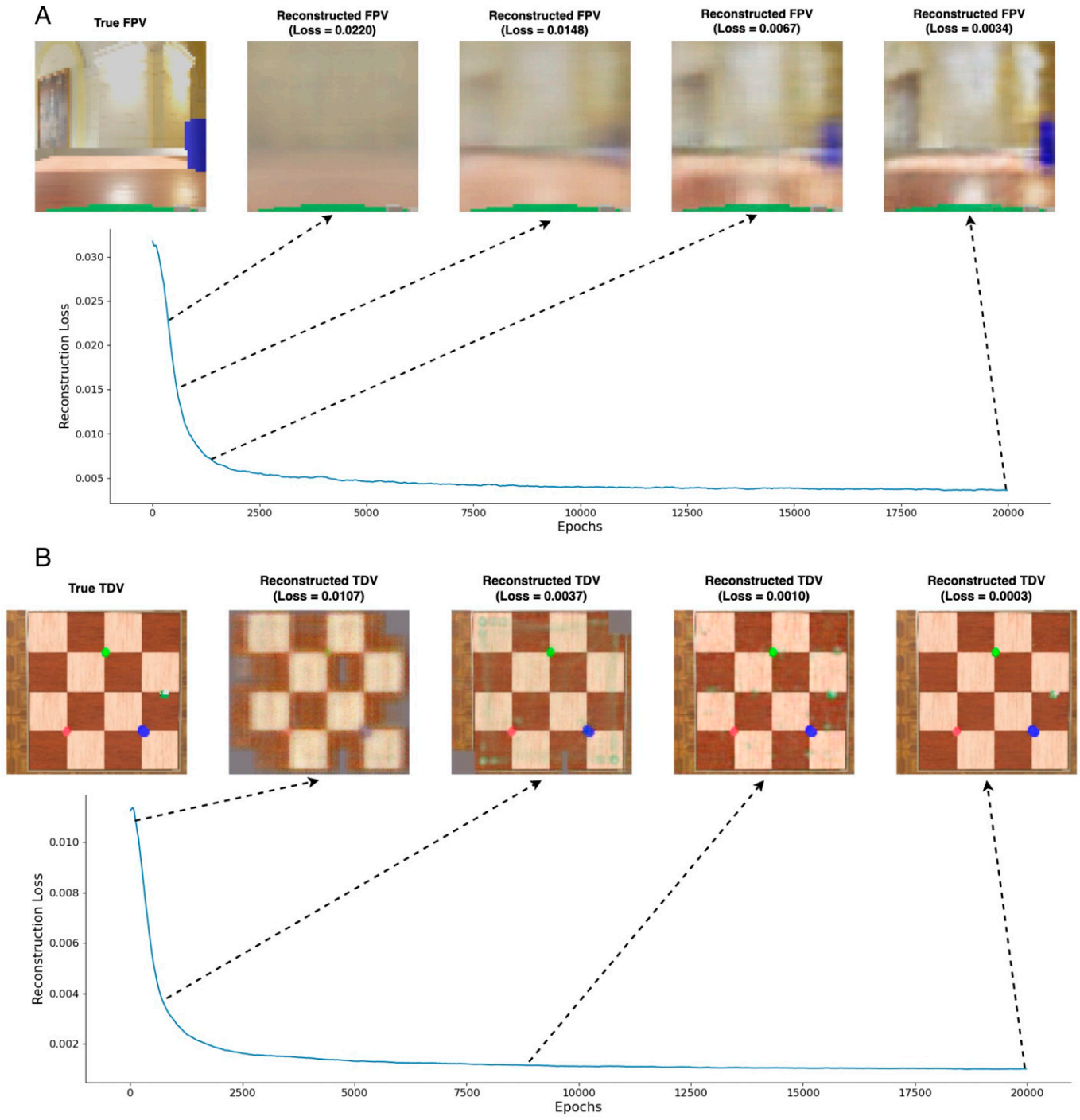

Fig. 2. Reconstruction loss during VAE training with 100 latent variables. The true image is the VAE target, and the other images are reconstructions at different points in the training. (*A*) TDV to FPV transformation. (*B*) FPV to TDV transformation.

for all latent variables, shows these trends, especially in the tails of the distributions where latent variables were strongly positively and negatively correlated.

Fig. 3 shows representative latent variable examples of head direction and location-specific tuning. Interestingly, the place fields for TDV to FPV tended to be sharp, whereas the FPV

Table 1. Percentage of strong correlations ($p < 0.01$)

LV	Hdg (FPV to TDV), %	Hdg (TDV to FPV), %	Obj (FPV to TDV), %	Obj (TDV to FPV), %	Plc (FPV to TDV), %	Plc (TDV to FPV), %
30	49*	33	50	73*	50	73
50	54*	26	49	73*	47	70*
100	55*	30	49	67*	47	69*

*Denotes significantly more strong correlations for that transformation direction ($p < 0.01$; Wilcoxon rank sum test).

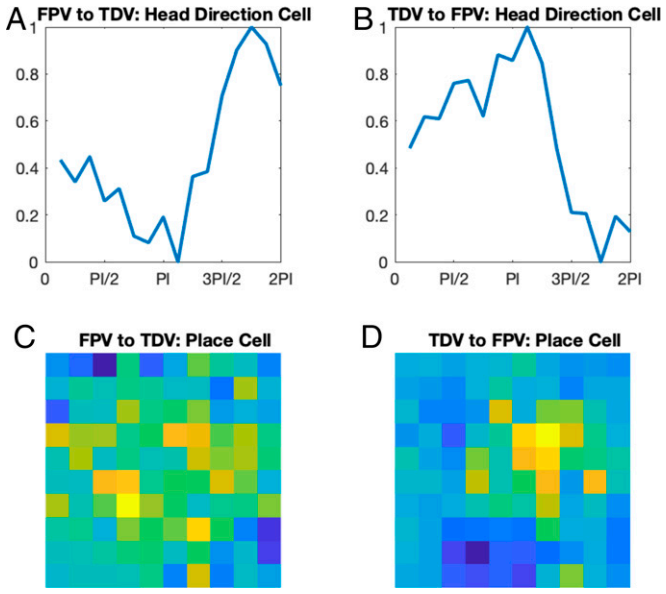

Fig. 3. Representative latent variable responses during simulations with 100 latent variables. (*A* and *B*) Latent variables that responded similarly to head direction cells. (*C* and *D*) Latent variables that responded similarly to place cells. Note that *C* was typical of an FPV to TDV transformation, and *D* was typical of a TDV to FPV transformation.

to TDV tended to be more diffuse. Even though there were fewer head direction–sensitive latent variables in the TDV to FPV transformation, the shape of the head direction latent variables was similar for both transformation directions (see *SI Appendix*, Figs. S7 and S8 for more examples). In contrast, FPV to TDV latent variables that were strongly correlated with place tended to be more diffuse than those in the TDV to FPV direction. For example, compare the FPV to TDV place fields in *SI Appendix*, Fig. S11 to the TDV to FPV place fields in *SI Appendix*, Fig. S12.

To understand the difference in spatial coding between transformation directions, we measured the spatial information (23) and spatial coherence (24) of the latent variables. In general, spatial information measures the extent to which activity rates are high across a small subset of locations and low or nonexistent across the remainder of an environment. Coherence measures the extent to which high activity rates cluster in a single location, as in "place fields" of HPC neurons. Together, they provide a good metric for how strongly the latent variables encode locations. Fig. 4 shows the distributions for these spatial metrics in simulations with 100 latent variables. For both transformation directions, the spatial information and spatial coherence were significantly larger than a random distribution containing the same latent variables with their positions shuffled ($p < 0.000001$; Wilcoxon signed-rank test). Furthermore, the spatial metrics were significantly larger for the TDV to FPV transformation than the FPV to TDV transformation ($p < 0.000001$; Wilcoxon signed-rank test). Latent variable place fields were distributed throughout the environment, with some tendency for the centers of place fields to be on the borders and corners (*SI Appendix*, Fig. S36). The sparsity metric (23), which is roughly the fraction of the environment in which the latent variable was active, ranged from very specific to broad (*SI Appendix*, Fig. S37). Together, these measures suggest that both VAEs had strong spatial tuning and that TDV to FPV had stronger spatial tuning than FPV to TDV. Overall, these results suggest that the TPV to FPV transformation relied more on place-specific coding, whereas the FPV to TPV transformation relied more on head direction coding with diffuse place fields.

Spatial representations, such as place cells, grid cells, and head direction cells, appear early in rodent development (25) and almost immediately upon entering a new environment (26). We looked at the spatial metrics of the latent variables in the VAE model at 20, 200, and 2,000 epochs of training (Fig. 5). At each of these training stages, the spatial information and coherence were significantly larger than random in both transformation directions

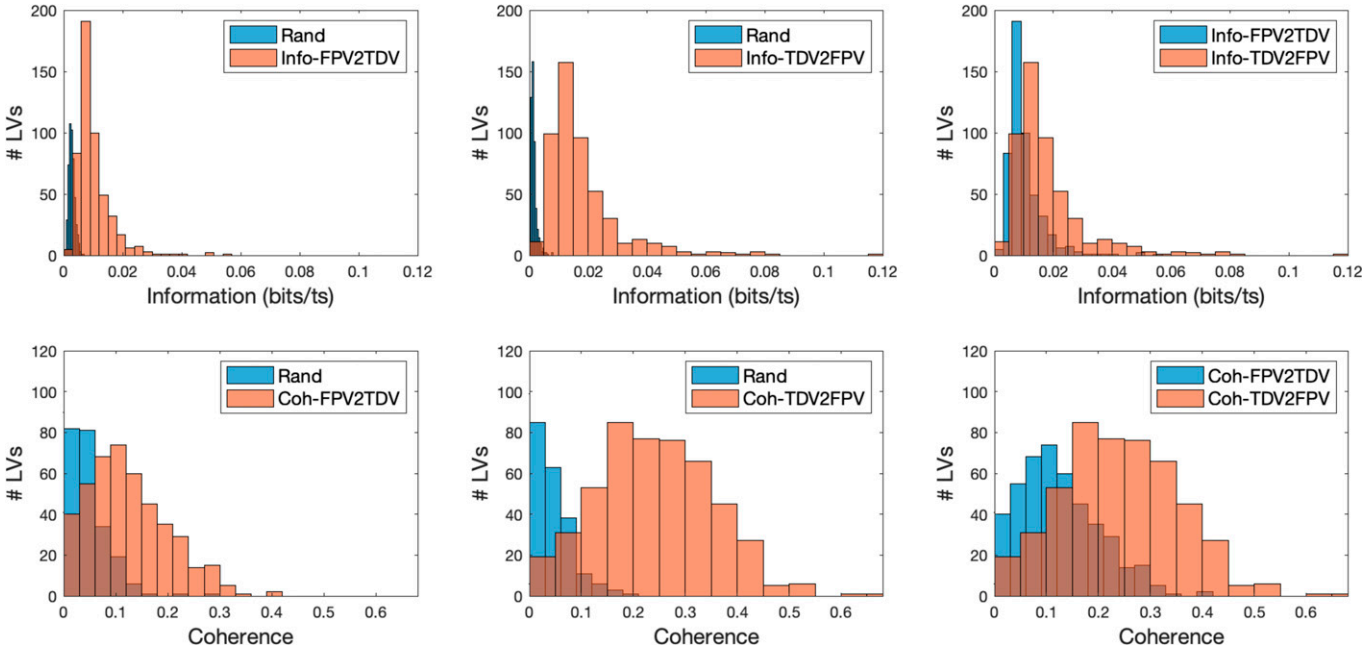

Fig. 4. Spatial metrics for latent variables. *Top* shows the distributions of spatial information, and *bottom* shows the distributions of spatial coherence for simulations with 100 latent variables. *Left* compares the spatial metrics for FPV to TDV transformations (orange) with a random distribution (blue) in which the location activity bins were shuffled. *Middle* compares the spatial metrics for TDV to FPV transformations (orange) with a random distribution (blue). *Right* compares TDV to FPV transformations (orange) with FPV to TDV transformations (blue). The TDV to FPV transformation had significantly stronger spatial metrics then the FPV to TDV transformation for both information and coherence.

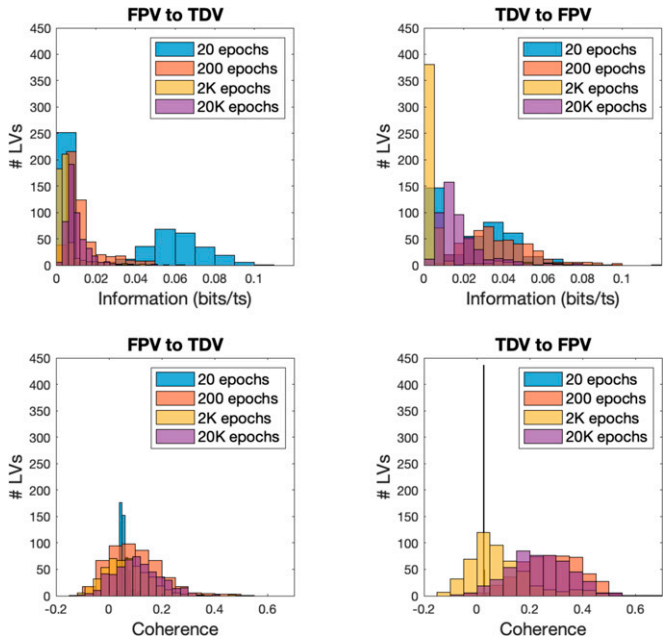

Fig. 5. Spatial metrics for latent variables during early training. *Top* shows the distributions of spatial information, and *Bottom* shows the distributions of spatial coherence, for simulations with 100 latent variables after 20, 200, 2,000, and 20,000 epochs of training.

(Wilcoxon signed-rank test; $p < 0.001$). At 20 epochs, the spatial information was bimodal, with many latent variables close to zero. By 200 epochs and continuing to the end of training (20,000 epochs), the spatial information had a high degree of overlap with the end of training (Fig. 5, *Top*). After 20 epochs of training, coherence was extremely low. Similar to spatial information, from 200 epochs until the end of training, the coherence had a high degree of overlap with the end of training (Fig. 5, *Bottom*). *SI Appendix*, Figs. S43–S54 shows example latent variables that respond to place and head direction after 20, 200, and 2,000 training epochs. Given the spatial metric values and low reconstruction loss by 200 training epochs, it appears the model can support spatial navigation after limited environmental exposure.

Effect of Latent Variable Ablations. We next tested how sensitive reconstruction was to particular latent variables. We ablated (i.e., set to zero) the most sensitive (top 25%) latent variables to environmental features. Fig. 6*A* shows the relative FPV to TDV reconstruction losses, and Fig. 6*B* shows the relative TDV to FPV reconstruction losses. Relative loss was calculated by dividing the ablation loss by intact loss for each image. For the sensitivity to cylinders, head direction, and location, the loss was significantly greater for the TDV to FPV transformation when the top 25% of latent variables were ablated. This may be due to different representation schemes; whereas TDV to FPV relies more on specific selectivity, which could be sensitive to ablations, FPV to TDV may rely on a more distributed population code. Example reconstructions are shown in *SI Appendix*, Figs. S13–S24.

Effect of Environmental Perturbations. We wondered how the VAEs would respond to perturbation of local and distal cues in the environment. Therefore, we ran simulations where the robot took the same trajectory for 10,000 data points, but some aspect of the environment was changed. For example, the original background, which was an entrance hall (Fig. 1*A*), was changed to mountains, while leaving everything else the same. This was an example of perturbing distal cues. In the other cases, we perturbed local cues by rendering the green cylinder invisible, or by rendering both the green and blue cylinders invisible. We then examined how the VAE, which was trained on the original environment, responded to the 2,000 test data points in the perturbed environment (Fig. 7).

Perturbing the local and distal cues had three effects. First, the relative loss was significantly greater for the TDV to FPV than the FPV to TDV transformation for both distal and local cues ($p < 0.0000001$; Wilcoxon signed-rank test). In fact, the median loss was zero for FPV to TDV transformations when the green cylinder was missing, meaning that there was no loss in many cases. This makes sense, since images in the FPV do not always contain a cylinder. Second, in both transformation directions, the loss due to the background change was significantly larger than removing cylinders, and the loss due to removing the green and blue cylinder was significantly greater than removing just the green cylinder. Large loss due to background change makes sense, since more pixels in the images are affected. Third, in most cases, the transformation losses were relatively small, suggesting that, with the exception of outliers, the VAEs were able to recover many features in a view image (*SI Appendix*, Figs. S25–S32). Furthermore, at the population level, these perturbations appeared to

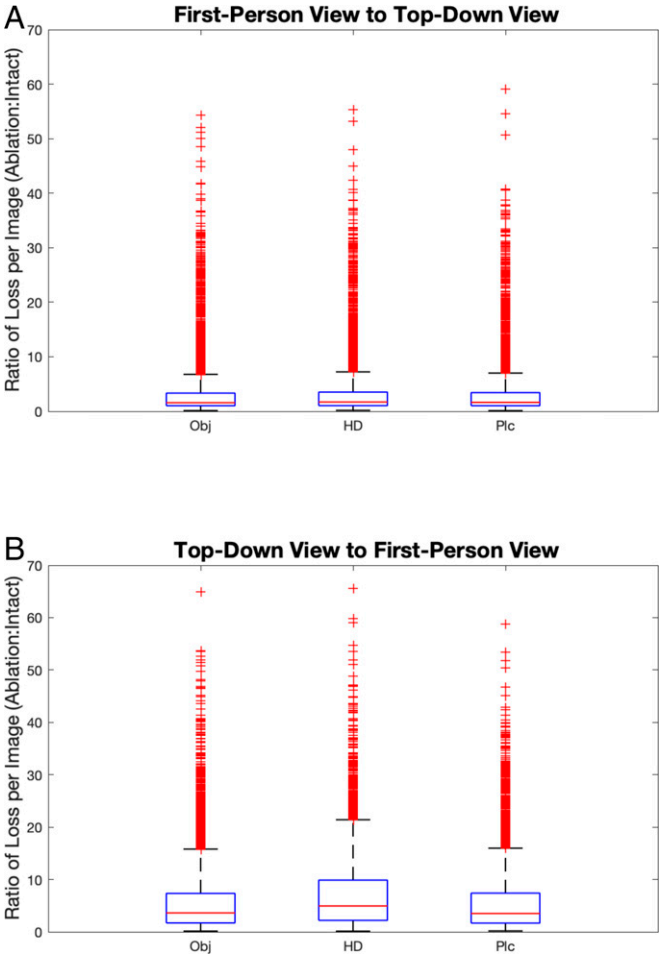

Fig. 6. Relative loss during ablation studies of the top 25% of latent variables that were correlated with objects (Obj), heading (HD), or place (Plc). The figures show the ratio of the ablation loss to the intact model loss for each image. In each box, the central mark is the median (red), the edges of the box are the 25th and 75th percentiles (blue), the whiskers extend to the most extreme data points that are not considered outliers, and the outliers are plotted individually with a red plus sign. (*A*) FPV to TDV transformation. (*B*) TDV to FPV transformation. All reconstruction losses for ablations were significantly larger for TDV to FPV than for FPV to TDV transforms ($p < 0.0000001$, Wilcoxon signed-rank test).

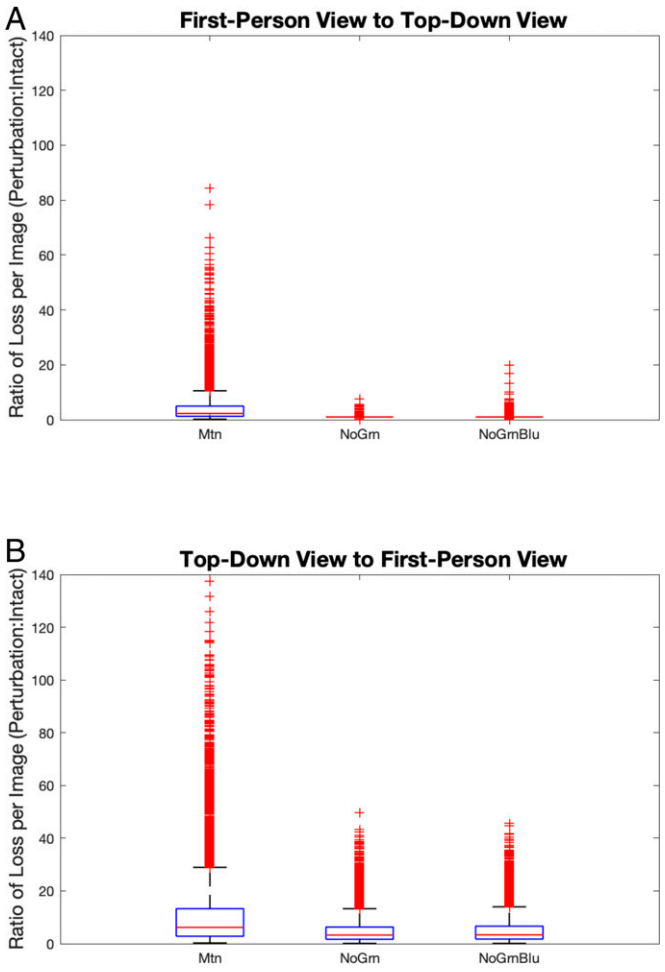

Fig. 7. Relative loss during perturbation experiments. The figures show the ratio of the loss due to a perturbation loss to the intact model loss for each image. Relative losses are shown for changing the background to mountains (Mtn), removing the green cylinder (NoGrn), and removing both the green and blue cylinders (NoGrnBlu). (*A*) FPV to TDV transformation. (*B*) TDV to FPV transformation. All reconstruction losses for ablations were significantly larger for TDV to FPV than for FPV to TDV transforms.

have a minimal effect on latent variable sensitivity to the distance to an object, heading, or place (*SI Appendix*, Figs. S33–S35). The range of correlation values (i.e., many latent variables had strong correlations) and the trends observed in the intact model were preserved.

Perturbations of local and distal cues led to substantial remapping. Table 2 shows the percentage of latent variables that were not correlated before the perturbation but remapped to be significantly correlated with spatial features after the perturbation, and those latent variables that were previously correlated with spatial features prior to the perturbation and remapped to not

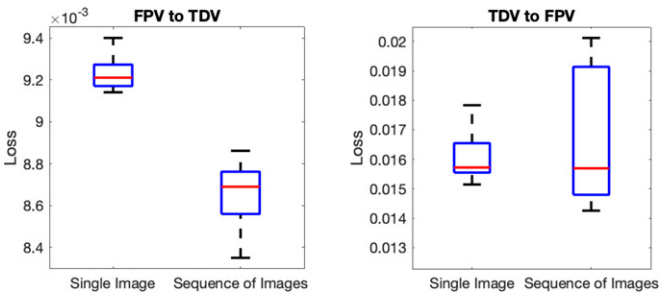

Fig. 8. Loss comparison between a sequence of images and a single image used for reconstruction.

be correlated with spatial features after the perturbation. There was significantly more remapping when the distal cues changed during FPV to TDV transformations (see top row in Table 2), and there tended to be more remapping during FPV to TDV transformations when local cues changed (see bottom two rows in Table 2). Such adaptation and remapping has been observed in the RSC (1, 14).

Alternative Models.

Benefit of sequences for linking views. Unlike the nervous system, the model presented here does not have a temporal component. Rather, a single image from one view is linked to another. We wondered whether a sequence of image views would benefit the ability to link different view perspectives. We created VAEs that took, as input, a sequence of five images from one view and output a reconstruction of the last image in the other view (*SI Appendix*, Figs. S38 and S39). Interestingly, the loss was roughly the same for a TDV to an FPV transformation, but the loss was reduced when a sequence of FPVs was used to reconstruct the TDV (Fig. 8). This makes intuitive sense because a sequence of FPVs would provide more varied information than a sequence of TDV images. Despite these differences, the reconstruction loss was small in both cases, and the spatial metrics in both cases were similar (see *SI Appendix*, Fig. S41 for spatial metrics, and see *SI Appendix*, Figs. S55–S58 for example place- and head direction–sensitive latent variables). Although this may be important for a living organism, the computational overhead may outweigh the benefit of using sequences if this model were deployed on a system like a navigating robot.

Combined VAE that reconstructs FPV to TDV and TDV to FPV simultaneously. An open issue is whether a single VAE could perform the linkage of views in both directions. We constructed a VAE to test this by interleaving FPVs and TDVs as inputs, while conducting the training for reconstructing the transformed view (*SI Appendix*, Fig. S40). This combined model had spatial information and coherence nearly identical to the two separate VAEs (*SI Appendix*, Fig. S42). As before, the spatial information and the

Table 2. Remapping due to environmental perturbations (LV = 100)

Perturbation	Obj (FPV to TDV), %	Obj (TDV to FPV), %	Hdg (FPV to TDV), %	Hdg (TDV to FPV), %	Plc (FPV to TDV), %	Plc (TDV to FPV), %
Mountains	30* (18*)	14 (12)	28* (19*)	8 (10)	28* (19*)	13 (11)
No green cylinder	2 (2)	13* (16*)	1 (1)	8 (11*)	2 (1)	12* (15*)
No green, no blue cylinder	6 (4)	13* (17*)	2 (3)	8 (11*)	6 (4)	13* (15*)

Table entries show the percent of latent variables that became significant ($p < 0.01$), and table entries in parentheses show the percent of latent variables that became insignificant ($p \geq 0.01$) after the perturbation.
*Denotes significantly more remapping for that transformation direction ($p < 0.01$; Wilcoxon rank sum test).

coherence of latent variables were higher going from a TDV to an FPV than from an FPV to a TDV. Furthermore, many latent variables responded to head direction. Example latent variables that are sensitive to place and heading for the combined model can be seen in *SI Appendix*, Figs. S59–S62. It should be noted that, although the reconstruction loss going from TPV to FPV was similar, the loss going from FPV to TDV was significantly higher in the combined model (Wilcoxon signed-rank test; $p < 0.001$). These results suggest that one system could perform this function, but that there may be advantages to having a separate system for each transformation direction.

Discussion

Using tools from machine learning and artificial intelligence (AI) (i.e., VAEs), we investigated a fundamental cognitive computation, which is the ability to change one's perspective from a global mental map to an egocentric sensory experience, and vice versa. This is an important, but oftentimes overlooked, aspect of the cognitive map (27). Our results suggest that two types of neural activity can support this transformation: 1) Specific representations of locations and objects were observed when reconstructing an FPV from a TDV, and 2) specific representations of heading with more diffuse representations of location were observed when reconstructing a TDV from an FPV. The response of the latent variables in the present VAEs has similarities to those observed in the brain, with latent variables resembling head direction cells, place cells, and cells that encode the distance to objects. Critically, we did not explicitly create such cell types in our model; rather, these responses, which resemble place and head direction cells, emerge from the way our model solves this problem.

Neurobiological Evidence for Transformations between Views. Studies with humans and nonhuman primates have revealed neural correlates for transformations between views or perspectives. In humans, evidence suggests that RSC activity is related to route learning from an egocentric viewpoint (28) and to navigating from a first-person perspective after looking at a top-down map perspective (29, 30). Although RSC is more active in first-person navigation compared with top-down navigation (29), the evidence for a transformation is indirect, and there are multiple factors to which the RSC could be responding. Within the first-person perspective, RSC is involved in changing viewpoints to different locations. For example, mentally rotating one's viewpoint to the position of an avatar or an arrow yields activation in RSC and parietal–occipital sulcus (31). In addition, RSC activity is related to the amount of viewpoint change relative to the environmental frame (32), and RSC activity is modulated by the magnitude of a viewpoint shift (33). Furthermore, RSC responds to perspective changes when the magnitude of the shift is unknown ahead of time, indicating that it is helpful in making online perspective changes.

Human intracranial recordings in the medial temporal lobe revealed boundary-anchored neural representations that were modulated by one's own as well as another individual's spatial location (34), and recordings of the EC in the monkey revealed neurons that represent gaze position in multiple spatial reference frames (2). These findings more broadly indicate that multiple brain regions in the primate play a role in orienting and processing view-based information from different perspectives (35, 36).

In the rat, neurons have been observed that respond to specific spatial frames of reference (e.g., allocentric, egocentric or route-centric), as well as multiple spatial reference frames (1, 37, 38). Some RSC neurons have place-specific responses, and the activity of a population of RSC neurons is sufficient to predict the location of a rat in a maze (1). RSC neurons are sensitive to distance and orientation relative to boundaries (39). RSC in both humans (40, 41) and rodents (42) has been implicated in mapping distance to other locations in the environment. RSC head direction neurons encode allocentric orientation relative to environmental boundaries (43). RSC activity is sensitive to distance and orientation relative to boundaries and to left versus right turning actions (42, 44). PPC neurons have been observed to simultaneously map the position in multiple external frames of reference (38). Still, none of these studies have put the rodent in situations where it had multiple viewpoints, which would be difficult to undertake. One study, which is a step in this direction, recorded from the rodent HPC and showed place cell responses to itself and to another rat it was observing (15).

However, these studies only involve changes between different first-person viewpoints. Extending this, our simulations suggest a neural solution that uses strong heading signals plus a mixture of place responses to link FPVs to TDVs and more-specific place responses with heading to link TDVs to FPVs.

An interesting parallel to the task carried out in our simulations are studies with freely behaving bats. Place cells, head direction cells, and grid cells have been observed in the bat both on the ground when crawling and in the air when flying (4, 5, 7). Similar to ref. 15, social place cells have been found when the bats are viewing other bats (6). GPS tracking of foraging bats over long time periods has demonstrated the ability to use landmarks and take novel routes from a TDV (45, 46). Taken together, there is evidence suggesting that encoding and utilizing different spatial perspectives during navigation and memory is a common cognitive function across multiple organisms and multiple brain regions.

Modeling Transformations between Views. Computational neuroscience models have attempted to simulate transformations between allocentric position and orientation in the real world and the egocentric, retina-framed view at that location and orientation. In one influential model, head direction or gaze direction cells modulated activity in RSC by rotating environmental variables (8, 13). This modulation converted allocentric border or object vector cells into an egocentric bearing to boundaries and objects, and vice versa. Such gain-modulated fields have been postulated and observed in PPC (47, 48). In another model, RSC acted as an arbitrator, which, depending on the model's confidence in the current task, would activate an allocentric reference frame in the HPC or an egocentric frame in the PPC (12). While these models and others have been useful in suggesting the pathways and neural activity that might produce these transformations, they make assumptions on the underlying computations. For example, Byrne et al. (8) and, later, Bicanski and Burgess (13) suggested that the same circuit computed the transformation for both directions. In addition, they did not specifically examine the linkages between FPVs and TDVs.

The present model attempts to be agnostic on how these computations are implemented. Rather than creating a neural network model based on the known responses or connectivity in specific brain regions, we used VAEs to solve the transformation task (20), and then tested their feasibility by comparing their responses (i.e., hidden layers and latent variables) to empirical experiments. The latent variables in these VAEs indicate different responses and computations, depending on the transformation direction. Furthermore, the separate models for each transformation had less reconstruction loss than a combined model. Whereas strong spatial coding by individual latent variables was observed in TDV to FPV transformations, head direction coding and diffuse place coding were more prevalent in FPV to TDV transformations.

Applying Artificial Neural Networks to Neuroscience. Neuroscientists are turning to AI methods to explain their data (49). For instance, using deep convolutional neural networks (CNNs) as models of hierarchical feature representation in the ventral visual stream can show different cortical responses in the hidden layers (50–52). Moreover, CNNs have shown cortical responses in the dorsal visual stream (53). Others proposed similar models to synthesize control images to maximally activate specific neuron sites in the monkey V4 (54). In a somewhat related robotics study, a deep learning network used the robot's local views and geographic hints, such as satellite images or road maps, to plan paths over a variety of environments (55). In the present work, these TDVs were used to predict FPV, and vice versa. This might be an alternative method to localization and mapping in robotics.

The present work compared the sensitivity of latent variables with neural responses. Similarly, latent representations have been used to model the human visual system during working memory tasks (56). In another modeling study, a latent factor analysis using dynamical systems was applied to monkey and human motor cortex data to accurately predict behavioral variables and neural dynamics (57). Deep VAEs have been used to interpret fMRI data where there is a lack of labeled data (58). Our work is another example of how VAEs can be used to model the nervous system and make valuable predictions about the computations and implementations underlying cognitive function.

The present modeling work suggests a means to link an FPV to a TDV, and vice versa. Although the modeling work is not based in neurobiology, the different encodings depending on the transformation direction may be compatible with the RSC anatomy (59). Whereas the dysgranular RSC has greater connectivity with cortical regions, such as the visual cortex, which provide first-person information (60), granular RSC interacts more with the hippocampal formation and SUB, which is more sensitive to the allocentric coordinates (39, 42, 61). In our simulations, we showed that the model can recover from perturbations, without retraining, much like place cells in the HPC. Moreover, the system did not collapse when large proportions of latent variables were ablated. These perturbation and ablation simulations suggest that the model can flexibly and rapidly adapt to change, which is a hallmark of neural systems.

In summary, we present a computational model for linking perceptual views, which suggests a potential neural implementation for this cognitive function. Furthermore, it makes predictions regarding the functional anatomy suggesting separate encodings depending on the direction of the view transformation, and the ability to adapt without retraining when challenged with perturbations. Although this model provides a possible implementation, we do not yet know exactly how the mammalian brain carries out such a task. Therefore, it will be of interest to follow up this modeling study with similar experiments tailored for humans and other animals. Furthermore, linking different views, as in ref. 55, may be applicable to robot navigation.

Materials and Methods

Robot Simulations. The Webots robot environment (21, 22) was used to simulate an animal freely exploring its environment (Fig. 1). The Khepera robot, which is a two-wheeled robot produced by K-Team, was used for the simulations. During exploration, the robot had a 50% chance of moving straight, 25% chance of veering (i.e., an arcing turn) toward the left, and 25% chance of veering to the right. The robot has eight distance sensors that were used to detect the arena walls and the cylinders. If detected, the robot rotated, with a 50% chance, either clockwise or counterclockwise until the front-facing distance detectors were clear. A camera was mounted on top of the robot for the FPV. Every update cycle, the camera frame was converted into a 64 × 64 RGB image (FPV), and a simulated overhead camera took a JPEG image (TDV) of the robot in its environment. During the exploration, the TDV from the simulator, the FPV from the robot's camera, and other environmental parameters (e.g., place, heading, distance to object) were collected and saved. The "entrance_hall" was used as a default background. In the perturbation experiments, this was replaced with the "mountains" background, which was a desert scene with mountains in the distance. During the local cue perturbation experiments, either the green cylinder or both the green and blue cylinders were rendered invisible using the transparency setting in Webots. The Khepera's distance sensors still detected the object, but they were not visible to the camera. The same random number generator seed was used on all simulation runs to ensure that the robot's trajectory was the same in each condition. The software used to run the simulation is available on GitHub (10.5281/zenodo.7121464; 10.5281/zenodo.7114757).

VAE Construction and Latent Variable Analysis. VAEs (20) were constructed to transform between TDVs and FPVs. Briefly, the VAE design is as follows. The perspective transformation model used in the preliminary results is based on standard VAE training whose loss includes a reconstruction loss term and a Kullback-Leibler (KL) divergence term. The reconstruction term optimizes the network so that the input could be reconstructed, while the KL divergence term is used to constrain the latent representation close to the prior distribution. To promote stable training, we used KL annealing to gradually increase the weight of the KL term from zero to one (62). The model was trained for 20,000 epochs. In the first 50 epochs, the KL term increased linearly from zero to one. After 50 epochs, the KL term weight was kept at one. After the VAE was trained, TDVs or FPVs were presented to the model. We then could measure the latent variable sensitivity by examining how much each latent variable changes with environmental changes. More details are given in *SI Appendix*.

The VAE's latent variables were analyzed for sensitivity to objects, heading, and place.

Object sensitivity was measured by Pearson's correlation of the latent variable to the distance from the robot to the red, green, and blue cylinders. The distance function was given by Eq. **1**,

$$dCyl_{ti} = \sum_{t=1}^{T} \sum_{i=1}^{3} \|(loc_t - cyl_i)\|, \quad [1]$$

where $\|(loc_t - cyl_i)\|$ is the Euclidean distance between the location of the robot, *loc*, and the location of cylinder, *cyl*. The distance, *dCyl*, was calculated for each *i* cylinder at time *t* with *T* equal to 2,000 time steps. This created a vector of length 2,000 of the distances to each cylinder object (*dCyl*) in the simulation. The sensitivity of each latent variable to the cylinder objects was then given by Eq. **2**,

$$cyl_{ni} = \sum_{n=1}^{N} \sum_{i=1}^{3} corr(lv_n, dCyl_i), \quad [2]$$

where *N* is the number of latent variables, *i* is the cylinder index, and lv_n is the response of the latent variable *n* for the 2,000 time steps. The resulting cyl_{ni} are correlation coefficients for each latent variable to the red, green, and blue cylinder objects.

Head direction sensitivity was measured by Pearson's correlation of the latent variable to a cosine tuning curve with one of 16 preferred directions, which were evenly spaced from zero to 2π. The cosine tuning curve was given by Eq. **3**,

$$rHD_{ti} = \sum_{t=1}^{T} \sum_{i=1}^{16} max(0.0, cos(rot_t - pd_i)), \quad [3]$$

where the expected cosine tuning response for each *i* preferred direction *pd* was calculated based on the robot's heading *rot* at data point *t* with *T* equal to 2,000 time steps. This created a vector of length 2,000 of expected head direction responses for each preferred direction (*rHD*). The sensitivity of each latent variable to head direction was then given by Eq. **4**,

$$hd_{ni} = \sum_{n=1}^{N} \sum_{i=1}^{16} corr(lv_n, rHD_i), \quad [4]$$

where *N* is the number of latent variables, *i* is the preferred direction index, and lv_n is the response of the latent variable *n* for the 2,000 time steps. The resulting hd_{ni} are correlation coefficients for each latent variable for each of the 16 idealized head direction cells.

Place cell sensitivity was measured by Pearson's correlation of the latent variable to a 2D Gaussian centered at one of 16 locations, which were evenly spaced across the robot's arena. The Gaussian function was given by Eq. **5**,

$$rPlc_{ti} = \sum_{t=1}^{T} \sum_{i=1}^{16} \frac{exp(-||(loc_t - ctr_i)||)}{\sigma}, \quad \textbf{[5]}$$

where $||(loc_t - ctr_i)||$ is the Euclidean distance between the location of the robot, *loc*, and the centroid of the place cell, *ctr*, and σ was set to 0.33. The response, *rPlc*, was calculated for each *i* place at time *t* with *T* equal to 2,000 time steps. This created a vector of length 2,000 of expected place cell responses for each location (*rPlc*). The sensitivity of each latent variable to place was then given by Eq. **6**,

$$plc_{ni} = \sum_{n=1}^{N} \sum_{i=1}^{16} corr(lv_n, rPlc_i), \quad \textbf{[6]}$$

where *N* is the number of latent variables, *i* is the preferred direction, and lv_n is the response of the latent variable *n* for the 2,000 time steps. The resulting plc_{ni} are correlation coefficients for each latent variable for each of the 16 idealized place cells.

Data, Materials, and Software Availability. Neural network code and simulation software have been deposited in Zenodo (10.5281/zenodo.7121464; 10.5281/zenodo.7114757) (63).

ACKNOWLEDGMENTS. This work was supported by Air Force Office of Scientific Research Contract FA9550-19-1-0306, by the National Science Foundation Information and Intelligence Systems Robust Intelligence (Award 1813785) and by the NSF Neural and Cognitive Systems Foundations Award Information and Intelligence Systems (Award 2024633).

1. A. S. Alexander, D. A. Nitz, Retrosplenial cortex maps the conjunction of internal and external spaces. *Nat. Neurosci.* **18**, 1143–1151 (2015).
2. M. L. R. Meister, E. A. Buffalo, Neurons in primate entorhinal cortex represent gaze position in multiple spatial reference frames. *J. Neurosci.* **38**, 2430–2441 (2018).
3. M. Meister, Memory system neurons represent gaze position and the visual world. *J. Exp. Neurosci.* **12**, 1179069518787484 (2018).
4. M. Geva-Sagiv, L. Las, Y. Yovel, N. Ulanovsky, Spatial cognition in bats and rats: From sensory acquisition to multiscale maps and navigation. *Nat. Rev. Neurosci.* **16**, 94–108 (2015).
5. N. Ulanovsky, C. F. Moss, Hippocampal cellular and network activity in freely moving echolocating bats. *Nat. Neurosci.* **10**, 224–233 (2007).
6. D. B. Omer, S. R. Maimon, L. Las, N. Ulanovsky, Social place-cells in the bat hippocampus. *Science* **359**, 218–224 (2018).
7. A. Sarel, A. Finkelstein, L. Las, N. Ulanovsky, Vectorial representation of spatial goals in the hippocampus of bats. *Science* **355**, 176–180 (2017).
8. P. Byrne, S. Becker, N. Burgess, Remembering the past and imagining the future: A neural model of spatial memory and imagery. *Psychol. Rev.* **114**, 340–375 (2007).
9. E. R. Chrastil, S. M. Tobyne, R. K. Nauer, A. E. Chang, C. E. Stern, Converging meta-analytic and connectomic evidence for functional subregions within the human retrosplenial region. *Behav. Neurosci.* **132**, 339–355 (2018).
10. B. J. Clark, C. M. Simmons, L. E. Berkowitz, A. A. Wilber, The retrosplenial-parietal network and reference frame coordination for spatial navigation. *Behav. Neurosci.* **132**, 416–429 (2018).
11. R. A. Epstein, E. Z. Patai, J. B. Julian, H. J. Spiers, The cognitive map in humans: Spatial navigation and beyond. *Nat. Neurosci.* **20**, 1504–1513 (2017).
12. T. Oess, J. L. Krichmar, F. Rohrbein, A computational model for spatial navigation based on reference frames in the hippocampus, retrosplenial cortex, and posterior parietal cortex. *Front. Neurorobotics* **11**, 4 (2017).
13. A. Bicanski, N. Burgess, A neural-level model of spatial memory and imagery. *eLife* **7**, e33752 (2018).
14. E. L. Rounds, A. S. Alexander, D. A. Nitz, J. L. Krichmar, Conjunctive coding in an evolved spiking model of retrosplenial cortex. *Behav. Neurosci.* **132**, 430–452 (2018).
15. T. Danjo, T. Toyoizumi, S. Fujisawa, Spatial representations of self and other in the hippocampus. *Science* **359**, 213–218 (2018).
16. D. Derdikman, E. I. Moser, A manifold of spatial maps in the brain. *Trends Cogn. Sci.* **14**, 561–569 (2010).
17. S. M. Kim, S. Ganguli, L. M. Frank, Spatial information outflow from the hippocampal circuit: Distributed spatial coding and phase precession in the subiculum. *J. Neurosci.* **32**, 11539–11558 (2012).
18. P. E. Sharp, Subicular cells generate similar spatial firing patterns in two geometrically and visually distinctive environments: Comparison with hippocampal place cells. *Behav. Brain Res.* **85**, 71–92 (1997).
19. A. A. Wilber *et al.*, Cortical connectivity maps reveal anatomically distinct areas in the parietal cortex of the rat. *Front. Neural Circuits* **8**, 146 (2015).
20. D. P. Kingma, M. Welling, An introduction to variational autoencoders. *Found. Trends Mach. Learn.* **12**, 307–392 (2019).
21. Cyberbotics, Webots open source robot simulator. http://www.cyberbotics.com/. Accessed 22 September 2022.
22. O. Michel, Webots: Professional mobile robot simulation. *J. Adv. Robotics Syst.* **1**, 39–42 (2004).
23. W. E. Skaggs, B. L. McNaughton, M. A. Wilson, C. A. Barnes, Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. *Hippocampus* **6**, 149–172 (1996).
24. J. L. Kubie, R. U. Muller, E. Bostock, Spatial firing properties of hippocampal theta cells. *J. Neurosci.* **10**, 1110–1123 (1990).
25. T. J. Wills, F. Cacucci, N. Burgess, J. O'Keefe, Development of the hippocampal cognitive map in preweanling rats. *Science* **328**, 1573–1576 (2010).
26. L. M. Frank, G. B. Stanley, E. N. Brown, Hippocampal plasticity across multiple days of exposure to novel environments. *J. Neurosci.* **24**, 7681–7689 (2004).
27. E. C. Tolman, Cognitive maps in rats and men. *Psychol. Rev.* **55**, 189–208 (1948).
28. T. Wolbers, C. Weiller, C. Büchel, Neural foundations of emerging route knowledge in complex spatial environments. *Brain Res. Cogn. Brain Res.* **21**, 401–411 (2004).
29. K. R. Sherrill *et al.*, Hippocampus and retrosplenial cortex combine path integration signals for successful navigation. *J. Neurosci.* **33**, 19304–19313 (2013).
30. H. Zhang, M. Copara, A. D. Ekstrom, Differential recruitment of brain networks following route and cartographic map learning of spatial environments. *PLoS One* **7**, e44886 (2012).
31. S. Lambrey, C. Doeller, A. Berthoz, N. Burgess, Imagining being somewhere else: Neural basis of changing perspective in space. *Cereb. Cortex* **22**, 166–174 (2012).
32. V. Sulpizio, G. Committeri, S. Lambrey, A. Berthoz, G. Galati, Selective role of lingual/parahippocampal gyrus and retrosplenial complex in spatial memory across viewpoint changes relative to the environmental reference frame. *Behav. Brain Res.* **242**, 62–75 (2013).
33. V. Sulpizio, G. Committeri, S. Lambrey, A. Berthoz, G. Galati, Role of the human retrosplenial cortex/parieto-occipital sulcus in perspective priming. *Neuroimage* **125**, 108–119 (2016).
34. M. Stangl *et al.*, Boundary-anchored neural mechanisms of location-encoding for self and others. *Nature* **589**, 420–425 (2021).
35. P. L. St Jacques, K. K. Szpunar, D. L. Schacter, Shifting visual perspective during retrieval shapes autobiographical memories. *Neuroimage* **148**, 103–114 (2017).
36. P. L. St Jacques, A. C. Carpenter, K. K. Szpunar, D. L. Schacter, Remembering and imagining alternative versions of the personal past. *Neuropsychologia* **110**, 170–179 (2018).
37. P. Y. Jacob *et al.*, An independent, landmark-dominated head-direction signal in dysgranular retrosplenial cortex. *Nat. Neurosci.* **20**, 173–175 (2017).
38. D. A. Nitz, Spaces within spaces: Rat parietal cortex neurons register position across three reference frames. *Nat. Neurosci.* **15**, 1365–1367 (2012).
39. A. S. Alexander *et al.*, Egocentric boundary vector tuning of the retrosplenial cortex. *Sci. Adv.* **6**, eaaz2322 (2020).
40. E. R. Chrastil, K. R. Sherrill, M. E. Hasselmo, C. E. Stern, There and back again: Hippocampus and retrosplenial cortex track homing distance during human path integration. *J. Neurosci.* **35**, 15442–15452 (2015).
41. E. R. Chrastil, K. R. Sherrill, M. E. Hasselmo, C. E. Stern, Which way and how far? Tracking of translation and rotation information for human path integration. *Hum. Brain Mapp.* **37**, 3636–3655 (2016).
42. A. S. Alexander, D. A. Nitz, Spatially periodic activation patterns of retrosplenial cortex encode route sub-spaces and distance traveled. *Curr. Biol.* **27**, 1551–1560.e4 (2017).
43. J. Cho, P. E. Sharp, Head direction, place, and movement correlates for cells in the rat retrosplenial cortex. *Behav. Neurosci.* **115**, 3–25 (2001).
44. A. S. Alexander *et al.*, Neurophysiological coding of space and time in the hippocampus, entorhinal cortex, and retrosplenial cortex. *Brain Neurosci. Adv.* **4**, 2398212820972871 (2020).
45. S. Toledo *et al.*, Cognitive map-based navigation in wild bats revealed by a new high-throughput tracking system. *Science* **369**, 188–193 (2020).
46. L. Harten, A. Katz, A. Goldshtein, M. Handel, Y. Yovel, The ontogeny of a mammalian cognitive map in the real world. *Science* **369**, 194–197 (2020).
47. A. Pouget, T. J. Sejnowski, Spatial transformations in the parietal cortex using basis functions. *J. Cogn. Neurosci.* **9**, 222–237 (1997).
48. L. H. Snyder, K. L. Grieve, P. Brotchie, R. A. Andersen, Separate body- and world-referenced representations of visual space in parietal cortex. *Nature* **394**, 887–891 (1998).
49. K. Chen *et al.*, Neurorobots as a means toward neuroethology and explainable AI. *Front. Neurorobotics* **14**, 570308 (2020).
50. R. M. Cichy, D. Kaiser, Deep neural networks as scientific models. *Trends Cogn. Sci.* **23**, 305–317 (2019).
51. U. Güçlü, M. A. J. van Gerven, Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* **35**, 10005–10014 (2015).
52. D. L. Yamins, J. J. DiCarlo, Eight open questions in the computational modeling of higher sensory cortex. *Curr. Opin. Neurobiol.* **37**, 114–120 (2016).
53. P. J. Mineault, S. Bhaktiari, B. A. Richards, C. C. Pack, Your head is there to move you around: Goal-driven models of the primate dorsal pathway. bioRxiv [Preprint] (2021). https://doi.org/10.1101/2021.07.09.451701 (Accessed 24 October 2022).
54. P. Bashivan, K. Kar, J. J. DiCarlo, Neural population control via deep image synthesis. *Science* **364**, eaav9436 (2019).
55. D. Shah, S. Levine, Viking : Vision-based kilometer-scale navigation with geographic hints. arXiv:2202.11271 [cs.RO]. https://doi.org/10.48550/arXiv.2202.11271 (Accessed May 3, 2022).
56. S. Hedayati, R. E. O'Donnell, B. Wyble, A model of working memory for latent representations. *Nat. Hum. Behav.* **6**, 709–719 (2022).
57. C. Pandarinath *et al.*, Inferring single-trial neural population dynamics using sequential auto-encoders. *Nat. Methods* **15**, 805–815 (2018).
58. N. Qiang *et al.*, Deep variational autoencoder for mapping functional brain networks. *IEEE Trans. Cogn. Dev. Syst.* **13**, 841–852 (2021).
59. S. D. Vann, J. P. Aggleton, E. A. Maguire, What does the retrosplenial cortex do? *Nat. Rev. Neurosci.* **10**, 792–802 (2009).
60. H. Makino, T. Komiyama, Learning enhances the relative impact of top-down processing in the visual cortex. *Nat. Neurosci.* **18**, 1116–1122 (2015).
61. J. M. Olson, K. Tongprasearth, D. A. Nitz, Subiculum neurons map the current axis of travel. *Nat. Neurosci.* **20**, 170–172 (2017).
62. S. R. Bowman *et al.*, "Generating sentences from a continuous space" in *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, S. Riezler, Y. Goldberg, Eds. (Association for Computational Linguistics, 2016), pp. 10–21.
63. J. Xing, KarlXing/Link-Views-in-Brain: Code for Linking Views in the Brain. Zenodo. https://zenodo.org/record/7121464. Deposited 28 September 2022.