SENSORY GEOMETRY SETS THE REFERENCE FRAME OF MULTIMODAL SPATIAL CODES

Anonymous authors
Paper under double-blind review

ABSTRACT

Recurrent networks trained to predict their upcoming sensory input develop spatial representations resembling those of the hippocampus, constructing a world-centred (allocentric) map from first-person input. Such models have almost always been given a single sense, usually vision, yet hippocampal representations draw on many senses whose signals depend differently on position and heading. Here we asked how the available senses determine the reference frame of the learned representation. An agent explored a virtual arena with first-person vision and binaural audio from four speakers; recurrent networks received one or both senses plus self-motion and were trained to predict their input one second ahead. Because egocentric signals change with the direction the agent faces, we measured *spatial allocentricity*: the consistency of each unit's spatial tuning across headings (1, identical at every heading; 0, unrelated). All trained networks became more allocentric, but audition far more than vision (median allocentricity 0.677 ± 0.010 vs 0.363 ± 0.033). This difference reflected the geometry of the signals rather than the senses themselves: what the agent sees depends on heading as well as position, whereas what it hears from the surrounding speakers depends mainly on position. Rearranging the speakers so that sound also depended on heading along one axis reduced auditory allocentricity by 40%, removing interaural cues at test time increased allocentricity. Moreover, with visual input fixed, making the prediction target heading-invariant or heading-dependent moved allocentricity up or down. With both senses, vision resolved the positional ambiguity created by rearranging the speakers without raising allocentricity, and removing vision at test time could raise or lower allocentricity, depending on the geometry of the signals the network had to predict. The reference frame of a learned spatial representation is therefore set by how its signals depend on heading, not by their modality. For predictive models, this suggests that latent invariances are inherited from the geometry of inputs and targets and cannot be read from decoding accuracy alone; for the brain, that place field directionality should track how the cues an animal relies on depend on heading.

1 INTRODUCTION

The mammalian hippocampus is a structure necessary for memory and flexible navigation, allowing an animal to move efficiently within and between environments (Scoville & Milner, 1957; Morris et al., 1982). This capacity is thought to depend on a suite of spatially modulated neurons, including place cells (O'Keefe & Dostrovsky, 1971), grid cells (Hafting et al., 2005), and head direction cells (Taube et al., 1990). A key property of each is its spatial reference frame, i.e. the coordinate system relative to which it encodes location or direction. Place and grid cells represent self-location in a world-centred, or *allocentric*, reference frame, largely invariant to the direction the animal faces. Head direction cells encode heading relative to that same allocentric reference frame. Together, these cells are thought to constitute a cognitive map (Tolman, 1948; O'Keefe & Nadel, 1978), maintaining a stable allocentric reference frame that underpins flexible navigation: recognising a place from a novel viewpoint, planning new routes, or returning home via a shortcut. The senses that inform this map, however, are *egocentric*, encoding the world relative to the animal itself, and so they change with the direction it faces: what an animal sees depends on where it is looking, and the binaural cues it hears depend on the orientation of its head. The underlying mechanisms with which the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

hippocampus constructs a spatial reference frame from these egocentric inputs, each with different characteristics, remains an open question.

Normative models offer one answer: an allocentric representation can be learned from egocentric input through prediction alone. Following the view of the hippocampus as a predictive map (Stachenfeld et al., 2017), recurrent neural networks (RNNs) trained to predict their upcoming sensory input develop units resembling place and head direction cells (Uria et al., 2020; Whittington et al., 2020; Gornet & Thomson, 2024; Levenstein et al., 2024), presumably because position and heading largely determine what the agent will sense next. The spatial representations that emerge are shaped by the statistics of that input. Abrate et al. (2026), for instance, showed that the changing movement statistics of a developing rat pup reproduce the order in which place, head direction, and conjunctive cells appear. Similarly, slowness-based models yield place, head direction, or spatial view cells depending on how the agent moves (Franzius et al., 2007). Yet these models have rarely used more than one sense, and that sense has almost always been vision, while path-integration models have only access to self-motion (Cueva & Wei, 2018; Banino et al., 2018).

In contrast, empirical work amply demonstrated that hippocampal spatial representations are multimodal (Jeffery, 2007). Place fields persist in blind rats (Save et al., 1998), CA1 place cells form in mice navigating by odour alone (Radvansky & Dombeck, 2018), and in mouse medial entorhinal cortex different neurons are spatially tuned to visual cues, auditory cues, or both (Nguyen et al., 2024). Crucially, the senses differ in how their signals depend on position and heading: the visual scene changes with every turn of the head, whereas cues such as odour concentration or the loudness of a distant sound depend on where the animal is. The reference frame of the hippocampal representation might therefore depend on which senses are available, as the strongly directional place cells seen in vision-dominated environments suggest (Acharya et al., 2016). However, the reason and mechanism underlying the integration of different senses into an allocentric reference frame remains unexplored.

Here we trained RNNs to predict their upcoming sensory input from the current sensory observation (whether first-person vision, binaural auditon, or both) and self-motion, as an agent explored a virtual arena. In learned representations, we measured how consistently each unit's spatial tuning was preserved across headings, which we termed *spatial allocentricity*. All networks became more allocentric after training, but audition far more than vision. For audition, allocentricity was largely present in the sensory embedding, whereas for vision it had to be built by recurrence. We traced this difference to the geometry of the signals rather than the senses themselves. What the agent sees depends on heading as well as position, whereas what it hears from the surrounding speakers depends mainly on position. Rearranging the speakers, removing interaural cues, and changing only the prediction target each moved allocentricity in the direction our theory predicts. Within the quadrant arena, allocentricity tracked the *interaction share* of the target – a quantity that measures how a sense depends on position and heading jointly rather than on either alone. The reference frame is therefore set by how the signals a network receives and predicts depend on heading, not by their modality. More generally, the invariances of a predictive model's latent state can be inherited from its inputs and targets, and cannot be read from decoding accuracy alone.

We illustrate our contributions as follows.

- **Signal geometry, not modality, sets the reference frame.** With modality fixed, changing the layout of sound sources or the heading-dependence of the prediction target (both by retraining), or removing interaural cues at test time, moves allocentricity as predicted; auditory networks retain it without recurrence, whereas visual networks need recurrence.
- **A sense's positional and reference-frame roles can dissociate.** A sense can add positional information without raising allocentricity, and removing a sense can raise or lower allocentricity depending on what the network is trained to predict.
- **A customisable audiovisual scene and multimodal hippocampal model testbed.** A predictive model with arenas that vary cue geometry with modality fixed, and oracle targets that vary heading-dependence with input fixed.
- **Place cell directionality should follow cue geometry *in vivo*.** We provide a testable neurobiological prediction that place field directionality should track how the available cues depend on heading rather than which senses are present, and a cue that supplies a missing spatial axis may improve positional accuracy without making place fields less directional.

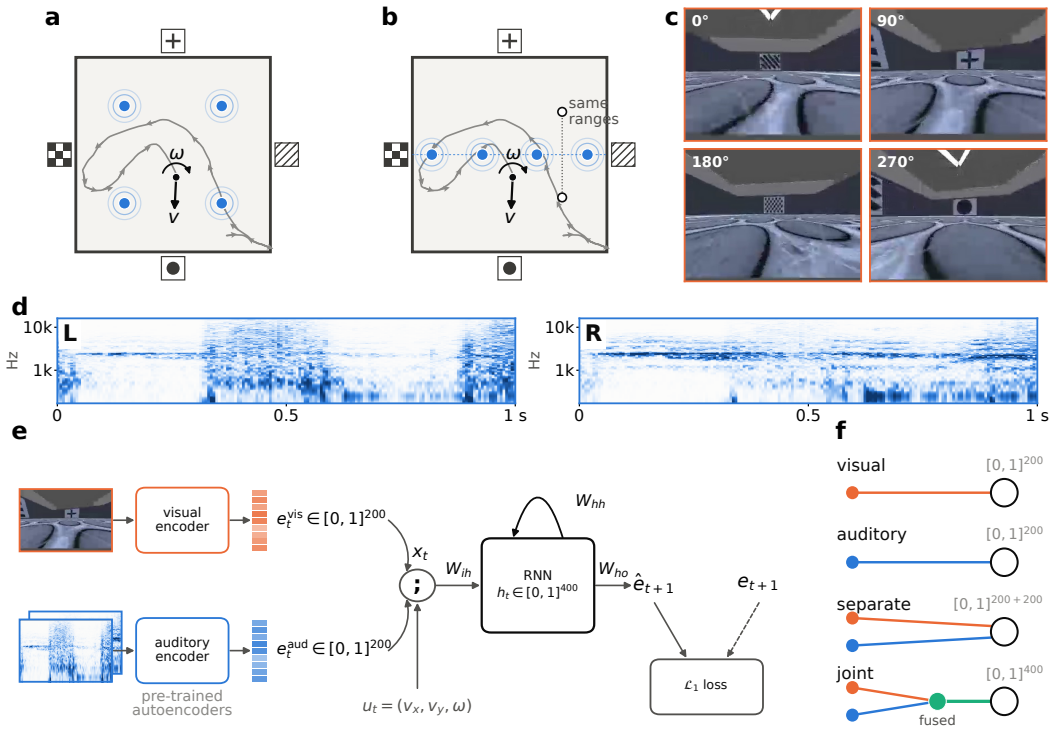

Figure 1: **Virtual environment, sensory input, and predictive model. a** Quadrant arena (2×2 m) with a distinct visual cue on each wall and four invisible speakers (blue), one per quadrant, each looping a different natural sound. Grey line, 30-s sample trajectory. **b** Line arena, identical except that the speakers lie on the horizontal midline; positions mirrored about the line (open circles) are equidistant from every speaker and differ acoustically only in interaural and spectral cues, which depend on heading. **c** First-person views at the four cardinal headings in the centre of the arena: the visual input changes substantially with heading. **d** Left- and right-ear log-frequency spectrograms of a 1-s audio window. **e** Frozen autoencoders map each image and spectrogram to a 200-dimensional embedding (400-dimensional for the joint encoder); a recurrent network receives the embedding(s) with self-motion $\mathbf{u}_t = (v_x, v_y, \omega)$ and is trained with an L_1 loss to predict the embedding 1 s ahead. **f** The four input regimes: visual, auditory, separate (both embeddings concatenated), and joint (fused by a shared encoder).

2 Methods

2.1 Environment and agent

We built two arenas (2×2 m) in Unity3D. Each wall carries a distinct visual pattern, and each of four invisible speakers plays a distinct natural sound in a loop (Mishra et al., 2021). In the *quadrant arena* there is one speaker in each quadrant (Fig. 1a), whereas in the *line arena* the same speakers lie on the horizontal midline (Fig. 1b). The two arenas are otherwise identical. Trajectories were generated with RatInABox (George et al., 2024), which produces smooth, rodent-like exploration with a bias towards walls, and the same trajectories were used in both arenas, so occupancy and the coupling between position and heading are matched. The agent moves at a mean speed of 0.128 m/s, with median rotational velocity of $33°/s$ (s.d. $99°/s$, Appendix A).

The agent sees a first-person 320×240 RGB view with a $120°$ field of view with its direction of travel (Fig. 1c), and hears binaural audio at 44.1 kHz rendered through a head-related transfer function (HRTF; 3D Tune-In Toolkit Cuevas-Rodríguez et al. (2019)), which filters each source according to its direction relative to the head. This way the two ears receive different, frequency-dependent levels (interaural level differences, ILDs) and direction-specific spectral shaping. Sound levels fall

with distance from the speakers, down to a background noise minimum, while reverberation was disabled.

2.2 Sensory input and encoders

Each image was paired with a 1-s binaural audio window centred on the frame (more details in Appendix A). Each channel was converted to a spectrogram by short-time Fourier transform, resampled onto logarithmically spaced frequencies between 20 Hz and 16 kHz, and log-compressed (Fig. 1d), approximating the logarithmic frequency and intensity scales of the cochlea (the inner ear). As the two ears are processed separately, ILDs and direction-specific spectral cues are preserved. However, interaural time differences are dismissed (more details in Appendix B.

Images and spectrograms were encoded by convolutional autoencoders, trained on separate recordings and frozen, into 200-dimensional embeddings bounded in $(0, 1)$, as a stand-in for cortical sensory representations (Appendix B). To test whether the results depend on the network's component where the senses converge, we also trained a joint encoder that fuses both modalities into 400-dimensional embeddings.

2.3 Predictive recurrent network

The predictive network is a single-layer recurrent network of 400 sigmoid units (Fig. 1e). At 1-s intervals, the network received the sensory embedding $\mathbf{e}_t$ and the agent's self-motion $\mathbf{u}_t = (v_x, v_y, \omega)$, its translational and angular velocity accumulated over the following second, and predicted the embedding one second ahead,

$$\mathbf{h}_{t+1} = \sigma\big(\mathbf{W}_{\mathrm{ih}}\,\mathbf{x}_t + \mathbf{W}_{\mathrm{hh}}\,\mathbf{h}_t\big), \qquad \mathbf{x}_t = [\,\mathbf{e}_t;\,\mathbf{u}_t\,], \qquad \hat{\mathbf{e}}_{t+1} = \mathbf{W}_{\mathrm{ho}}\,\mathbf{h}_{t+1}, \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid function, biases are omitted without loss of generality, and $\mathbf{h}_0 = \mathbf{0}$. The network is trained to minimise the L_1 error between $\hat{\mathbf{e}}_{t+1}$ and $\mathbf{e}_{t+1}$. Allocentric information, such as position or heading variables, never enter the loss, except in the target manipulations of Section 3.3. We compared four regimes (Fig. 1f): *visual* and *auditory*, in which the network receives and predicts a single 200-dimensional embedding; *separate* (late fusion), where sensory embeddings are concatenated; and *joint* (early fusion), which uses the fused 400-dimensional embedding. Each regime was trained with five random seeds in each arena, yielding 2 arenas $\times$ 4 regimes$\times$ 5 seeds = 40 trained networks in total, and all analyses use a held-out test recording (Appendix C).

2.4 Spatial allocentricity index

A unit is spatially allocentric if it fires in the same places whichever way the agent faces, as a place cell with an omnidirectional field does (Muller et al., 1994). We call the degree to which this holds *spatial allocentricity*, hereafter *allocentricity*. It measures whether the location of firing is invariant to heading, not how much a unit fires or whether it is egocentric, i.e. heading-dependent changes in rate alone leave it high (see Appendix D), and a low value indicates that the position at which a unit fires changes with heading, as in allocentric conjunctive cells (Sargolini et al., 2006). We divided the arena into 12×12 spatial bins, and heading (taken as direction of travel) into $H = 4$ bins of 90°. A rate map was computed for each unit and heading bin by smoothing activity and occupancy before dividing,

$$r_i^h(b) \;=\; \frac{\big(G_s * S_i^h\big)(b)}{\big(G_s * N^h\big)(b)}, \tag{2}$$

where $S_i^h(b)$ is the summed activity of unit i in spatial bin b while heading lies in bin h, $N^h(b)$ the corresponding occupancy, and G_s a Gaussian kernel of width $s = 1$ bin. Maps were compared only over spatial bins with at least five samples in every heading bin. The allocentricity index a_i of unit i is the mean Pearson correlation ρ between its maps across pairs of heading bins,

$$a_i \;=\; \binom{H}{2}^{-1} \sum_{h<h'} \rho\big(r_i^h,\, r_i^{h'}\big), \tag{3}$$

which is 1 for a unit whose map is unchanged by heading and near zero or negative for one whose map changes with heading (Fig. 2a, Sup. Fig. S1a). A unit tuned to a particular heading direction,

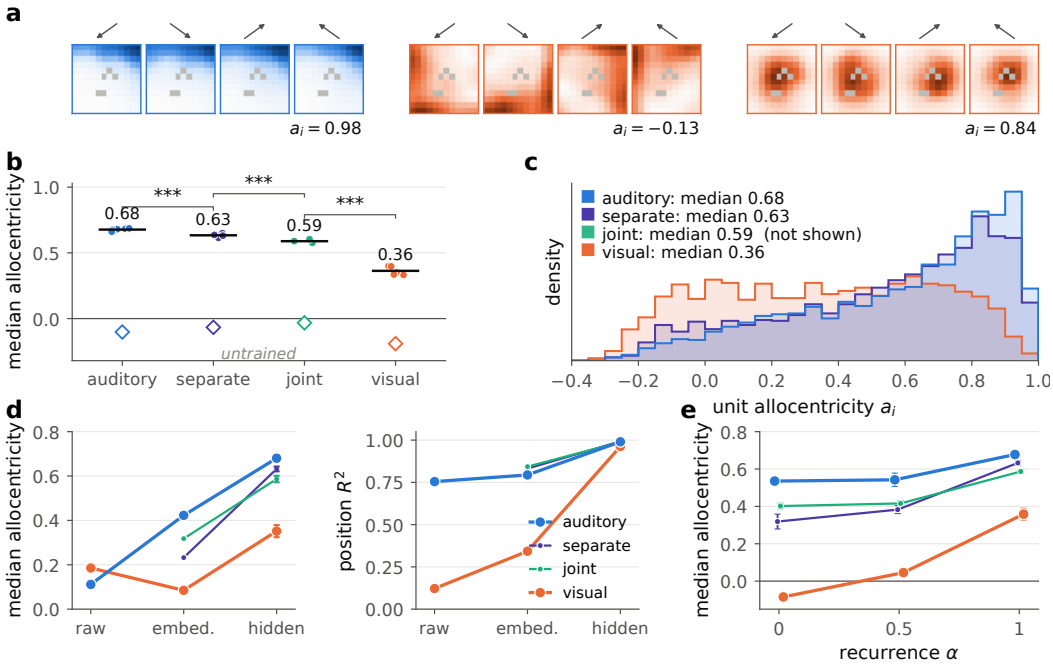

Figure 2: **All trained networks become more allocentric, but audition far more than vision.** **a** Rate maps of three example units in four heading bins (arrows, heading). Left, an auditory unit whose map is unchanged by heading ($a_i = 0.98$); middle, a visual unit whose map changes with heading ($a_i = -0.13$); right, one of the minority of visual units that are heading-invariant ($a_i = 0.84$). Spatial-heading bins not visited 5 times are greyed out. **b** Median allocentricity of each trained network (dots, five per regime; bars, mean) and of untrained networks (open diamonds), which lie near or below zero. **c** Distribution of allocentricity across hidden units; visual units are spread widely, whereas most auditory and separate-regime units are strongly invariant (joint regime omitted for clarity). **d** Median allocentricity (left) and position decoding R^2 (right) at three stages: raw input, sensory embedding, and hidden state. Position becomes decodable from both senses. **e** Median allocentricity when the trained recurrent weights are scaled by α at test time ($\alpha = 0$, no recurrence; $\alpha = 1$, trained network): auditory allocentricity survives without recurrence, visual does not. Quadrant arena throughout. ***, $p < 0.001$ (Welch's two-sample t-test, comparison between adjacent regimes only); error bars, s.d. over five networks.

such as a head direction cell, scores below zero since the index does not measure whether heading is encoded. Units with unreliable maps (split-half reliability below 0.2, see Appendix D) were excluded, and each network is summarised by the median a_i over its units; values are reported as mean $\pm$ s.d. calculated over the five seeds of every condition, with uncorrected t-tests.

3 Results

3.1 Training increased allocentricity, audition far more than vision

After training, networks in all four input regimes (visual, auditory, and the separate and joint multimodal regimes; Fig. 1f) became more allocentric in the quadrant arena, but to significantly different degrees (Fig. 2b). The auditory network was the most allocentric (median $a_i = 0.677 \pm 0.010$), followed by the separate and joint networks (0.633 ± 0.012 and 0.587 ± 0.010), with the visual network being the least allocentric (0.363 ± 0.033). Untrained networks were near or below zero (see Appendix C). The difference held across the population (Fig. 2c, Sup. Fig. S1b,c) – most auditory units were strongly allocentric, whereas visual units spanned the whole range, with only a minority strongly invariant to heading.

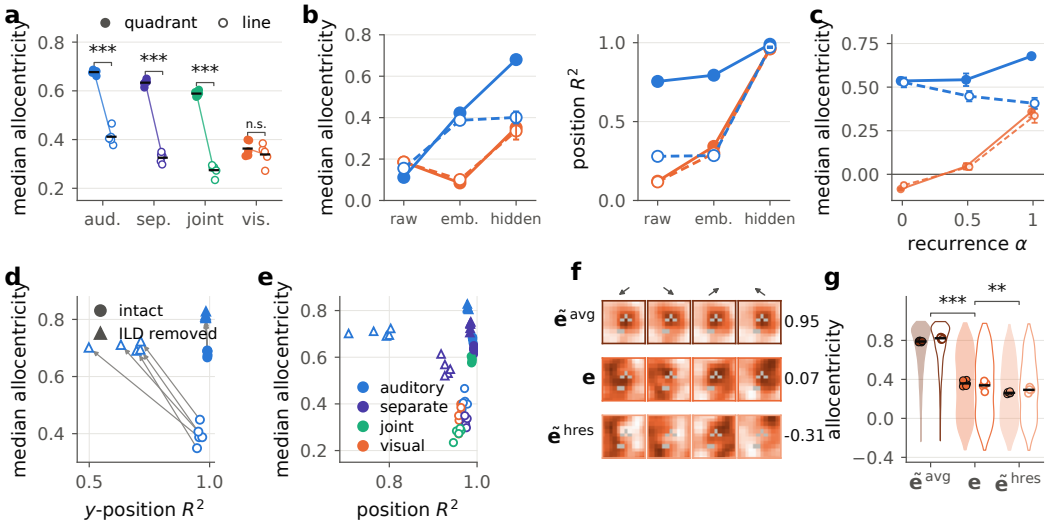

Figure 3: **Allocentricity follows how strongly input and target depend on heading.** **a** Median allocentricity in the quadrant (filled) and line (open) arenas. **b** Median allocentricity (left) and position decoding R^2 (right) at three stages for the auditory (blue) and visual (orange) networks; quadrant (filled, solid), line (open, dashed). **c** Median allocentricity against the recurrence scaling α. **d** Median allocentricity against y-position decoding for auditory networks, intact (circles) and with ILDs removed (triangles); lines join the same network. **e** Median allocentricity against position decoding for all networks, intact and with ILDs removed. **f** One visual embedding dimension under each target, as rate maps in four heading bins; numbers, allocentricity. **g** Allocentricity of units in visual networks trained on each target (violins; dots, network medians), quadrant (filled) and line (open). Welch's t-tests between arenas (a) or adjacent targets (g): ***, $p < 0.001$; **, $p < 0.01$.

To understand how this difference arises, we measured non-linear position decoding (kernel ridge regression; Appendix E) and allocentricity at three stages: raw input, sensory embedding, and hidden state (Fig. 2d). Position could already be decoded from raw spectrograms ($R^2 = 0.755$), but barely from raw images ($R^2 = 0.122$). This is expected, as each speaker plays a distinct sound, and its loudness falls with distance, so the relative loudness of the four sounds uniquely specifies the agent's location. A single view, by contrast, often shows only an unmarked stretch of wall and cannot on its own precisely inform the agent about where it is located. Allocentricity, however, was low for both raw inputs (vision 0.186, audition 0.111), because both individual pixels and spectrogram bins change with the direction the agent faces. In the sensory embeddings, allocentricity rose for audition (0.423) but not vision (0.085), with the joint embedding lying in between (0.318). Sound carries cues that depend mainly on position, such as the overall loudness of each source, whereas a single $120°$ view shows different parts of the arena at different headings. A feedforward transformation of the auditory input was therefore sufficient to produce a substantially allocentric representation.

Since allocentricity can be read directly from the auditory input, recurrence should matter less for audition than for vision. To test this, we scaled the recurrent weights by a factor α at test time (from $\alpha = 0$, no recurrence to $\alpha = 1$, original trained RNN; Fig. 2e). Without recurrence, the auditory network remained allocentric (0.535), whereas the visual network did not (-0.085). Visual allocentricity therefore depends on recurrence, presumably to integrate successive views with self-motion.

3.2 The geometry of the sound alters auditory allocentricity

To test the hypothesis that auditory allocentricity arose from the arrangement of the speakers rather than from audition itself, we moved the speakers onto the horizontal midline (the line arena; Fig. 1b). Positions mirrored about the line when equidistant from every speaker, so which side of the line the agent is on can be recovered only by combining interaural cues with heading. As a result, the auditory allocentricity should fall. As expected, the trained auditory network loses 40% of its

allocentricity (to 0.412 ± 0.033; Fig. 3a), approaching vision, while the visual network's results remain unchanged (0.339 ± 0.042).

The drop reflected how the network recovered position. The auditory input lost much of its positional information in the line arena ($R^2 = 0.280$ for the spectrogram and 0.285 for the embedding), yet position was decoded almost perfectly from the hidden state ($R^2 = 0.971$; Fig. 3b, Sup. Fig. S2b), which suggests that the network combined sound with heading, at the cost of allocentricity. This is supported by recurrence: from $\alpha = 0$ to the trained network, auditory allocentricity rose in the quadrant arena (from 0.535 ± 0.019 to 0.677) but fell in the line arena (from 0.527 ± 0.028 to 0.412; Fig. 3c), where sound had to be combined with heading to recover position.

Conversely, removing the heading-dependent ILD cue by replacing the left and right spectrogram channels with their mean raised allocentricity. In the line arena, allocentricity rose from 0.412 to 0.704 even though y-position R^2 fell from 0.957 to 0.650 (Fig. 3d); in the quadrant arena, where loudness alone specifies position, it rose from 0.677 to 0.813 with position R^2 almost unchanged (0.991 to 0.984). Position decoding was therefore a poor guide to allocentricity. In all 20 networks tested with interaural cues removed, allocentricity rose while y-position decoding fell (5 seeds $\times$ 2 models $\times$ 2 arenas; Fig. 3e; Appendix F). Auditory allocentricity thus depended on the arrangement of the speakers rather than on audition itself: it fell when position had to be recovered by combining sound with heading, and rose when the heading-dependent cue was removed.

3.3 The prediction target alone can set allocentricity

In every network so far, the target was the next value of the input: a visual network received vision as input and predicted vision as output. Input and target therefore depended on heading in the same way, and either could have set allocentricity. To test the influence of each, we retrained visual networks with the same input but an oracle target, built from the agent's true position and heading by averaging the embedding within spatial and heading bins (Appendix G). The *heading-averaged* target, $\tilde{\mathbf{e}}^{\,\mathrm{avg}}$, is the mean embedding at the agent's next position across all headings, so it ignores heading. The *heading-residual* target, $\tilde{\mathbf{e}}^{\,\mathrm{hres}}$, is the mean embedding at that position and heading minus this average, so it keeps only information about how the view changes with heading at each position (Fig. 3f).

Allocentricity was dictated by the target (Fig. 3g). The heading-averaged target raised visual allocentricity from about 0.363 to 0.789 ± 0.009 in the quadrant arena, above even the intact auditory network, whereas the heading-residual target lowered it to 0.262 ± 0.011; similar results were obtained with the line arena (Appendix G). Binning alone cannot explain the rise, because the heading-residual target was also a binned average, yet it lowered allocentricity. Thus, with the input fixed, allocentricity was set by how strongly the target depended on heading. The effects of modality (Section 3.1) and speaker arrangement (Section 3.2) are consistent with this principle: in each case, the less heading-dependent signal produced the more allocentric representation. The fall is expected from the linear readout: any interaction between position and heading in the target must be carried by units whose spatial tuning changes with heading. The rise is not, because a linear readout could equally combine such units into a heading-invariant output; nor does the readout explain why position and allocentricity changed independently (Section 3.2). The readout also does not require what the unimodal networks showed: a hidden state more allocentric than the embedding it predicted (vision 0.085 to 0.363; audition 0.423 to 0.677; Fig. 2d).

3.4 Supplying position does not make a representation allocentric

Removing ILDs showed that allocentricity can rise while positional information falls (Section 3.2). Multimodal networks in the line arena allowed the converse test: can a sense add positional information without making the representation more allocentric? When the speakers are on the midline, sound does not specify which side of the line the agent is on, but vision can, and the networks used it to do so. Removing vision at test time, by replacing its input with the mean, cost more y-position R^2 than removing audition (a fall of 0.101 versus 0.046; $p = 0.001$, paired t-test; Fig. 4b), the reverse of the quadrant arena (0.012 versus 0.063; $p < 0.001$). Removing ILDs at test time reduced y-position R^2 in the multimodal network far less than in the auditory network (by 0.067 versus 0.307; $p = 0.004$; Fig. 4c, Sup. Fig. S2c). Yet the multimodal networks were less allocentric than the auditory network (separate 0.315 ± 0.029, joint 0.275 ± 0.013, auditory 0.412 ± 0.033; Fig. 4a), at or

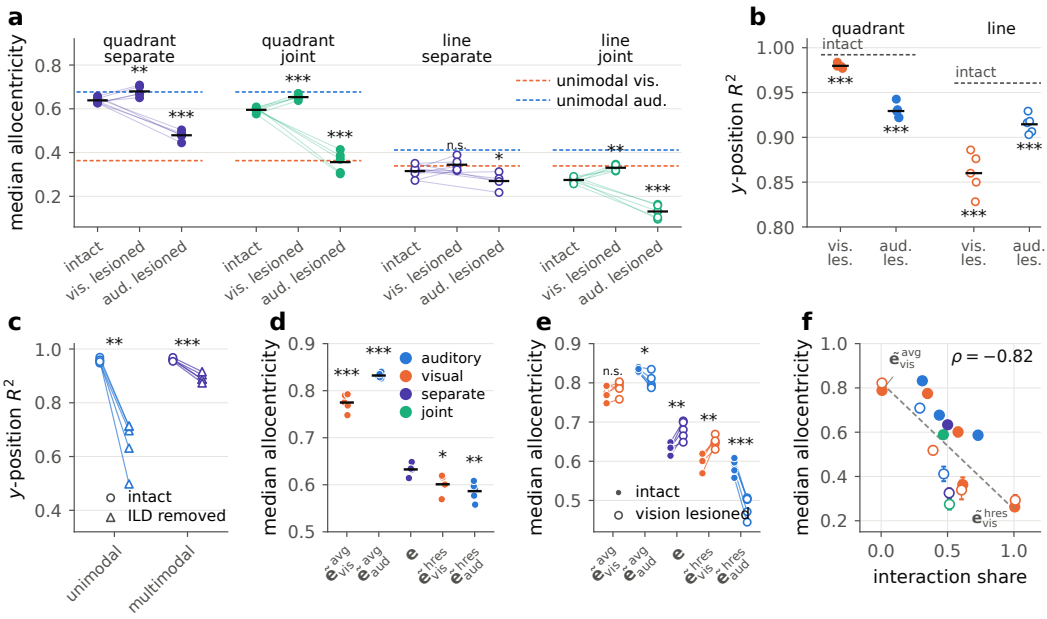

Figure 4: **In multimodal networks, allocentricity follows the prediction targets, not positional information.** **a** Median allocentricity of multimodal networks, intact and with vision or audition removed; lines join the same network; dashed lines, unimodal visual (orange) and auditory (blue) levels. **b** y-position decoding in separate model after removing vision (orange) or audition (blue), both arenas; dashed lines, intact. **c** y-position decoding in line arena, intact (circles) and with ILDs removed (triangles). **d** Median allocentricity of separate multimodal networks (quadrant arena) with one sense's target heading-averaged ($\tilde{\mathbf{e}}^{\mathrm{avg}}$), heading-residual ($\tilde{\mathbf{e}}^{\mathrm{hres}}$), or untransformed ($\mathbf{e}$); subscripts, the sense whose target was transformed. **e** The same networks, intact (filled) and with vision removed (open). **f** Median allocentricity of each condition (mean $\pm$ s.d. over five networks) against the interaction share of its target; dashed line, least-squares fit. Paired t-tests against intact (a, b, c, e), Welch's t-tests against **e** (d); *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

below the visual level (0.339 ± 0.042). Removing vision did not detectably change the allocentricity of the separate network ($p = 0.08$), although it raised that of the joint network ($p = 0.005$). Vision thus supplied the positional information that sound could not without raising allocentricity. This follows from Section 3.3: the networks still had to predict the heading-dependent visual embedding, so allocentricity reflected how the signals depended on heading, not how much position they carried.

3.5 What the network must predict determines the effect of removing vision

In the quadrant arena, audition set the reference frame of the multimodal networks. Their allocentricity was much closer to that of the auditory network than of the visual network (Fig. 2b); removing vision at test time raised it to about the auditory level (separate network, 0.679 ± 0.023, $p = 0.006$; joint, 0.653 ± 0.012, $p < 0.001$), whereas removing audition lowered allocentricity (separate, 0.472 ± 0.024; joint, 0.356 ± 0.049; both $p < 0.001$; Fig. 4a). Is this a genuine sensory difference arising because audition provides panoramic information, or is it a consquence of the arrangement of our sensory cues? The line arena favoured the second explanation, but was not decisive as moving the speakers also weakened the positional information in sound. We therefore trained separate multimodal networks in the quadrant arena with unchanged inputs but with either the visual or the auditory target heading-averaged or heading-residual (Section 3.3). If this is a genuine modality difference, removing vision should always do the same, otherwise it should depend on the targets.

The targets shifted the network's allocentricity as in Section 3.3. Relative to untransformed targets (0.633 ± 0.012), a heading-averaged target raised allocentricity in both visual (0.775) or auditory (0.832) and a heading-residual target lowered allocentricity (0.601 and 0.586; all $p < 0.02$; Fig. 4d).

The effect of removing vision now depended on the targets (Fig. 4e; Appendix G). It raised allocentricity with untransformed targets ($+0.047$, $p = 0.006$) or a heading-residual visual target ($+0.048$, $p = 0.008$), had no detectable effect with a heading-averaged visual target ($+0.013$, $p = 0.12$), and lowered it with a heading-averaged (-0.028, $p = 0.011$) or heading-residual auditory target (-0.101, $p < 0.001$), the latter matching removal of audition (-0.090). The same removal could thus raise, leave unchanged, or lower allocentricity, depending only on the targets.

To place all conditions on one axis, we split each target's variance across spatial and heading bins into parts that depend on position alone, heading alone, and the two jointly, the last being the *interaction share* (see Appendix H). The interaction share is near 0 for the heading-averaged target and near 1 for the heading-residual target. Across all conditions, allocentricity fell as the target's interaction share rose, regardless of modality (Spearman $\rho = -0.82$, $p < 0.001$; Fig. 4f), and the relationship also held among untransformed targets alone ($\rho = -0.71$, $n = 8$, $p = 0.047$). The interaction share, which is measurable before training, thus is a strong predictor of allocentricity.

4 Discussion

In this work, we asked what sets the reference frame of a spatial representation learned by prediction. The answer was not the modality of the input but how strongly the signals the network received and predicted depended on heading. Rearranging the speakers, removing ILDs, and changing only the prediction target each moved allocentricity as this principle predicts. In our networks, the invariances of the latent state followed the geometry of the signals the network received and predicted, and the target alone was enough to change them. Changing only the target altered the allocentricity, and whether removing vision raised or lowered allocentricity depended on what the network had to predict. As in self-supervised learning, where views set what a representation ignores (Tian et al., 2020; von Kügelgen et al., 2021), and in normative grid-cell models (Sorscher et al., 2023; Schaeffer et al., 2022), the objective shapes the invariances a network learns. Decoding alone did not reveal these invariances: removing ILDs raised allocentricity while y-position decoding fell in every network, and vision supplied the missing position without raising allocentricity (Sections 3.2 and 3.4). Invariance must therefore be measured directly.

Our results give a normative reason why *in vivo* place field directionality should depend on the cues an animal relies on and how they vary with heading. Place cells are strongly directional in vision-dominated virtual reality (Acharya et al., 2016) but largely omnidirectional in open arenas (Muller et al., 1994). Directionality also depends on behaviour (Markus et al., 1995) and on non-visual cues absent in virtual reality (Ravassard et al., 2013), but our account suggests that part of the difference reflects how strongly the dominant cues depend on heading, consistent with local cues on a linear track, sensed whichever way the animal runs, making place fields less directional (Battaglia et al., 2004). It predicts more directional place fields when an animal relies on cues that depend on heading, such as a landmark visible only from some directions, and less directional fields when its cues specify position regardless of heading. The prediction concerns directionality rather than stability: without vision, place fields can drift (Quirk et al., 1990; Markus et al., 1994).

Limitations. We used two minimalistic arenas without reverberation and a single-layer RNN. Whether the principle holds in richer scenes and larger world models is untested. The senses were not equivalent: step-wise audition came from four fixed, distinct sources over a 1-s window, whereas vision was a single 120° image. These asymmetries may inflate the difference between the senses but do not affect the manipulations within a sense. Removing a sense or the ILDs and scaling recurrence were test-time manipulations involving inputs outside the training distribution; the speaker and target manipulations, which involved retraining, supported the same principle. The interaction share is a property of the target alone, hence doesn't capture effects that arise from how position must be recovered from the input, such as the speaker arrangement. Finally, the joint network was consistently less allocentric than the separate one, which remain unexplained.

Conclusion. Networks trained to predict their sensory input were never told which reference frame to use; they inherited one from how the signals they received and predicted depended on heading. For predictive models, the choice of sensors and targets therefore shapes the invariances of the latent state, in ways that decoding accuracy alone does not reveal. For the brain, the directionality of place fields should likewise depend on how the cues an animal relies on vary with heading.

REFERENCES

Marco P. Abrate, Laurenz Muessig, Joshua P. Bassett, Hui Min Tan, Francesca Cacucci, Thomas J. Wills, and Caswell Barry. From movement to cognitive maps: recurrent neural networks reveal how locomotor development shapes hippocampal spatial coding. In *International Conference on Learning Representations*, 2026.

Lavanya Acharya, Zahra M. Aghajan, Cliff Vuong, Jason J. Moore, and Mayank R. Mehta. Causal influence of visual cues on hippocampal directional selectivity. *Cell*, 164(1–2):197–207, 2016.

Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J. Chadwick, Thomas Degris, Joseph Modayil, et al. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557(7705):429–433, 2018.

Francesco P. Battaglia, Gary R. Sutherland, and Bruce L. McNaughton. Local sensory cues and place cell directionality: additional evidence of prospective coding in the hippocampus. *Journal of Neuroscience*, 24(19):4541–4550, 2004.

Christopher J. Cueva and Xue-Xin Wei. Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. In *International Conference on Learning Representations*, 2018.

María Cuevas-Rodríguez, Lorenzo Picinali, Daniel González-Toledo, Carlos Garre, Ernesto de la Rubia-Cuestas, Luis Molina-Tanco, and Arcadio Reyes-Lecuona. 3d tune-in toolkit: An open-source library for real-time binaural spatialisation. *PloS one*, 14(3):e0211899, 2019.

Mathias Franzius, Henning Sprekeler, and Laurenz Wiskott. Slowness and sparseness lead to place, head-direction, and spatial-view cells. *PLoS Computational Biology*, 3(8):e166, 2007.

Tom M. George, Mehul Rastogi, William de Cothi, Claudia Clopath, Kimberly Stachenfeld, and Caswell Barry. RatInABox, a toolkit for modelling locomotion and neuronal activity in continuous environments. *eLife*, 13:e85274, 2024.

James Gornet and Matt Thomson. Automated construction of cognitive maps with visual predictive coding. *Nature Machine Intelligence*, 6:820–833, 2024.

Torkel Hafting, Marianne Fyhn, Sturla Molden, May-Britt Moser, and Edvard I. Moser. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806, 2005.

Kathryn J. Jeffery. Integration of the sensory inputs to place cells: What, where, why, and how? *Hippocampus*, 17(9):775–785, 2007.

Daniel Levenstein, Aleksei Efremov, Roy Henha Eyono, Adrien Peyrache, and Blake Richards. Sequential predictive learning is a unifying theory for hippocampal representation and replay. *bioRxiv*, pp. 2024.04.28.591528, 2024.

Etan J. Markus, Carol A. Barnes, Bruce L. McNaughton, Victoria L. Gladden, and William E. Skaggs. Spatial information content and reliability of hippocampal CA1 neurons: effects of visual input. *Hippocampus*, 4(4):410–421, 1994.

Etan J. Markus, Yu-Lin Qin, Brian Leonard, William E. Skaggs, Bruce L. McNaughton, and Carol A. Barnes. Interactions between location and task affect the spatial and directional firing of hippocampal neurons. *Journal of Neuroscience*, 15(11):7079–7094, 1995.

Ambika P Mishra, Nicol S Harper, and Jan WH Schnupp. Exploring the distribution of statistical feature parameters for natural sound textures. *PLoS One*, 16(6):e0238960, 2021.

Richard G. M. Morris, Paul Garrud, J. N. P. Rawlins, and John O'Keefe. Place navigation impaired in rats with hippocampal lesions. *Nature*, 297(5868):681–683, 1982.

Robert U. Muller, Elizabeth Bostock, Jeffrey S. Taube, and John L. Kubie. On the directional firing properties of hippocampal place cells. *Journal of Neuroscience*, 14(12):7235–7251, 1994.

Duc Nguyen, Garret Wang, Talah Wafa, Tracy Fitzgerald, and Yi Gu. The medial entorhinal cortex encodes multisensory spatial information. *Cell Reports*, 43(10):114813, 2024.

John O'Keefe and Jonathan Dostrovsky. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, 34(1):171–175, 1971.

John O'Keefe and Lynn Nadel. *The Hippocampus as a Cognitive Map*. Clarendon Press, Oxford, 1978.

Gregory J. Quirk, Robert U. Muller, and John L. Kubie. The firing of hippocampal place cells in the dark depends on the rat's recent experience. *Journal of Neuroscience*, 10(6):2008–2017, 1990.

Brad A. Radvansky and Daniel A. Dombeck. An olfactory virtual reality system for mice. *Nature Communications*, 9:839, 2018.

Pascal Ravassard, Ashley Kees, Bernard Willers, David Ho, Daniel Aharoni, Jesse Cushman, Zahra M. Aghajan, and Mayank R. Mehta. Multisensory control of hippocampal spatiotemporal selectivity. *Science*, 340(6138):1342–1346, 2013.

Francesca Sargolini, Marianne Fyhn, Torkel Hafting, Bruce L McNaughton, Menno P Witter, May-Britt Moser, and Edvard I Moser. Conjunctive representation of position, direction, and velocity in entorhinal cortex. *Science*, 312(5774):758–762, 2006.

Etienne Save, Arnaud Cressant, Catherine Thinus-Blanc, and Bruno Poucet. Spatial firing of hippocampal place cells in blind rats. *Journal of Neuroscience*, 18(5):1818–1826, 1998.

Rylan Schaeffer, Mikail Khona, and Ila Fiete. No free lunch from deep learning in neuroscience: A case study through models of the entorhinal-hippocampal circuit. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

William Beecher Scoville and Brenda Milner. Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry*, 20(1):11–21, 1957.

Ben Sorscher, Gabriel C. Mel, Samuel A. Ocko, Lisa M. Giocomo, and Surya Ganguli. A unified theory for the computational and mechanistic origins of grid cells. *Neuron*, 111(1):121–137, 2023.

Kimberly L. Stachenfeld, Matthew M. Botvinick, and Samuel J. Gershman. The hippocampus as a predictive map. *Nature Neuroscience*, 20(11):1643–1653, 2017.

Jeffrey S. Taube, Robert U. Muller, and James B. Ranck, Jr. Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. *Journal of Neuroscience*, 10(2):420–435, 1990.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Edward C. Tolman. Cognitive maps in rats and men. *Psychological Review*, 55(4):189–208, 1948.

Benigno Uria, Borja Ibarz, Andrea Banino, Vinicius Zambaldi, Dharshan Kumaran, Demis Hassabis, Caswell Barry, and Charles Blundell. The Spatial Memory Pipeline: a model of egocentric to allocentric understanding in mammalian brains. *bioRxiv*, pp. 2020.11.11.378141, 2020.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

James C. R. Whittington, Timothy H. Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy E. J. Behrens. The Tolman-Eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263, 2020.

AI USE STATEMENT

In this work, we used generative AI tools (Claude, Anthropic) to help formulate and check the analytical derivations in the appendices, to help design analyses, to give feedback on methodology and on the interpretation of results, and to help write analysis code, which the authors checked by hand. We did not use them to generate data, which were recorded in Unity, or to develop the models or hypotheses; the other tasks requiring disclosure are not applicable to this work. Additionally, we used them to search and summarise the literature, check citations, suggest a structure for the paper, and draft and edit text for readability. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All authors adhere to the ICLR Code of Ethics. This work uses only simulated agents and involves no human or animal subjects or personal data.

REPRODUCIBILITY STATEMENT

Appendix A specifies the arenas, speaker positions, trajectories, and recordings; Appendix B the sensory encoders; Appendix C the recurrent network, optimiser, seeds, and checkpoint selection; and Appendices D, E, and G the allocentricity index, position decoding, and oracle targets. Code, Unity scenes, trajectories, and trained models is provided in the supplementary codes, and further details will be provided when camera ready.

A ENVIRONMENT, STIMULI, MOTION, AND TRAJECTORIES

Arena and cues. The agent moves in a square arena rendered in a game engine. Positions are reported throughout in arena units, with the arena spanning 50×50 units (1 unit = 4 cm) and the origin at its centre, so coordinates run from -25 to $+25$; all analyses normalise positions to $(0, 1)$ before binning. The four walls carry distinct visual patterns, so the environment has no rotational symmetry and a view alone is in principle sufficient to disambiguate heading. Four loudspeakers are placed on the floor of the arena. In the quadrant configuration there is one per quadrant, at $(-20, -15)$, $(15, 10)$, $(15, -15)$ and $(-15, 10)$, so their bearings from a given point differ widely and the set of distances to them varies across the arena in two dimensions. In the line configuration the same four speakers are placed collinearly along the x-axis, at $(-20, 0)$, $(-7, 0)$, $(7, 0)$ and $(20, 0)$. Speakers are invisible to the agent. The two arenas are otherwise identical, with the same walls, rendering, speakers and sounds, so every difference reported between them is a difference in the geometry of the auditory cues, and the visual input is unchanged. Figure 1a,b draws the speakers schematically at the quadrant centres; the coordinates above are the rendered ones.

Trajectories. Trajectories are generated with `RatInABox` (George et al., 2024), whose agent model produces smooth, temporally correlated exploration with realistic speed and turning statistics rather than uniform random sampling. The agent is simulated at $\Delta t = 0.02$ s with a mean speed parameter of 0.06 arena widths per second, rotational velocity standard deviation 1.0 rad/s, rotational coherence time 0.08 s, and thigmotaxis 0.6, the last of which biases the agent towards walls as real rodents do. The speed parameter is expressed in arena widths, so in the 2 m arena it corresponds to 0.12 m/s; the rotational parameter is angular and is not rescaled. The realised statistics are a speed of 3.2 ± 1.9 arena units per second (mean $\pm$ s.d.), equivalent to 0.128 ± 0.076 m/s, and an angular velocity whose median absolute value is 33° per second and whose standard deviation is 99° per second. Head direction is recorded explicitly, whereas the allocentricity index uses the smoothed direction of travel.

The two arenas use identical trajectory and the agent visits the same places in the same order in both, so occupancy, the coupling between position and heading, and the shared-support mask of Appendix D are matched exactly, and no comparison between arenas can be attributed to differences in behaviour.

Visual stimuli. At each simulated instant the engine renders the agent's first-person view as a 320×240 RGB image from its position and heading, with a fixed 120° field of view and the camera aligned with the direction of travel. Because the trajectories are two-dimensional, the camera is additionally given a vertical tilt that varies smoothly over time, capped at 15° above and below the horizontal, so that the view pitches up and down as an animal's head would rather than remaining rigidly level. Images are scaled to $(0, 1)$ before encoding.

Auditory stimuli. The four speakers play different natural recordings, a running kitchen faucet, hair brushing, a squawking parrot and applause, each looped continuously, so that the sources are distinguishable by spectrotemporal content as well as by location. The signal reaching the agent is rendered binaurally with the 3D Tune-In Toolkit in High Quality mode, using the head-related transfer function of subject `IRC_1008` of the IRCAM LISTEN database (human-measured, 128-sample HRIRs at 44.1 kHz) together with the toolkit's near-field level-difference compensation. Reverberation was disabled so the rendering is anechoic. Interaural time differences were enabled, the toolkit storing HRIRs with the delay removed and applying the measured ITD separately, interpolated between directions, rather than synthesising it from a spherical head model. The two channels therefore differ in level, in arrival time and in spectral shape, as they would at two ears of a head of finite size, although only the level and spectral differences survive the magnitude spectrogram the encoder receives (Appendix B).

Audio is sampled at 44.1 kHz and consumed in one-second windows centred on the image, so each window carries 0.5 s of sound recorded before the image and 0.5 s recorded after. The audio input is therefore slightly anticipatory relative to the frame it accompanies. Successive input windows tile the recording without overlap, and no window overlaps the target one step ahead, so the lookahead does not give the network access to its own prediction target.

Each source is attenuated by a logarithmic rolloff, its level falls with the inverse of distance (-6 dB per doubling) from a reference distance of 1 arena unit, and stops attenuating at 25 units (half the

arena width). Loudness varies with distance throughout the arena and, unlike the interaural cues, does so independently of the agent's heading. For ILD removal, applied only at test time, the two channels of the log-magnitude spectrogram are replaced by their mean.

Recordings. Each arena provides seven recordings with distinct trajectories:

- Two 1-hour recordings used for training the autoencoders;
- Three 2-hour recordings used for training the RNNs;
- Two 1-hour recordings used for testing the trained RNNs.

As motion is continuous, no position and heading pair is ever sampled twice exactly: all 1.8×10^5 samples of a recording occupy distinct positions.

B SENSORY ENCODERS

The agent's sensory input is encoded by convolutional autoencoders trained before the predictive network and frozen thereafter. We used three autoencoders: a visual one and an auditory one for the separate-encoder regimes, and a joint one whose two branches are fused into a single code. Every encoder ends in layer normalisation followed by a sigmoid, so embeddings lie in $(0, 1)$ and are comparable in scale across modalities.

Visual. The agent's camera returns an RGB image of 320×240 pixels at each time step, rendered from its own position and heading. The trunk consists of nine convolutional layers with ReLU activations, of which four have stride 2: $3 \to 16$ (stride 1), $16 \to 32$ (stride 2), $32 \to 32$, $32 \to 64$ (stride 2), $64 \to 64$, $64 \to 64$ (stride 2), $64 \to 32$ (stride 2), $32 \to 32$, and a final $32 \to 8$ layer with a 2×3 kernel and no padding. The four stride-2 stages use 4×4 kernels, hence each cell of the final lattice sees a 16-pixel receptive field. The output is an $8 \times 19 \times 13$ map, flattened to 1976 features and passed through two linear layers to the embedding, then layer-normalised and sigmoid-squashed. The end embedding is 200-dimensional in the unimodal and separate-encoder regimes. The decoder mirrors the trunk with transposed convolutions and is trained to reconstruct the image. Only the encoder is used afterwards.

Auditory. Sound is rendered binaurally through a head-related transfer function, therefore the two channels differ in level and spectral shape as they would at two ears. Each one-second window of the 44.1 kHz stereo waveform is transformed as follows. A short-time Fourier transform with a 512-sample window and 50% overlap gives a magnitude spectrogram of 171 time bins per channel. Frequencies are restricted to 20 Hz–16 kHz in 185 bins, and these are resampled by linear interpolation onto logarithmically spaced centre frequencies spanning the same range, thus low frequencies are represented with finer resolution than high ones. Power is then compressed logarithmically in $10 \log_{10}(\cdot)$. The result is a $2 \times 185 \times 171$ tensor for each input into the autoencoder. Note that interaural timing does not survive this transformation.

That tensor is encoded by five blocks of convolution layers, batch normalisation, ReLU and 2×2 max-pooling, with channels $2 \to 16 \to 32 \to 64 \to 32 \to 16$ and a 2×2 kernel in the first block and 3×3 thereafter. The output is a $16 \times 5 \times 5$ map, flattened to 400 features and passed through a single linear layer to the 200-dimensional layer-normalised and sigmoidal embedding.

Joint. The joint encoder contains the same two trunks, layer for layer, so that a unimodal encoder and the corresponding branch of the joint encoder are the same network. Each branch is flattened and projected linearly to the full 400-dimensional embedding width, and the two 400-dimensional vectors are concatenated and passed through a single linear layer of $800 \to 400$ layer-normalised and sigmoidal embedding. The two modalities are therefore mixed in one shared code with no labelled visual or auditory half.

Note that in the separate-encoder model, a stream can be silenced in embedding space, because the code has a visual half and an auditory half. The joint model's 400-dimensional code has no such split, and removal must instead be applied to the branch features before fusion and the code re-derived. In both, the stream's features are held at their average over the recording, making the two architectures remain comparable under the same manipulation.

Training. All three autoencoders are trained by reconstruction on a dedicated embedding recording. We used cyclic Adam optimiser, with initial learning rate 0.0005, decaying epoch-wise exponentially by a factor of 0.99, and reset to start the next cycle. The start of each cycle to a value that decays by a factor of 0.95 per cycle, and decays exponentially within the cycle, which is then repeated. The visual encoder uses 50 cycles of 10 epochs, batch size 40, initial learning rate 6×10^{-4}, and a binary cross-entropy loss on the image, whose pixels are scaled to $(0, 1)$. The auditory encoder uses 10 cycles of 100 epochs, batch size 80, initial learning rate 5×10^{-4}, and an L_1 loss on the two-channel log-power spectrogram. Note that L_1 is used rather than a squared error because the spectrogram is dominated by a small number of high-energy bins.

C THE PREDICTIVE RECURRENT NETWORK OF THE HIPPOCAMPUS

We used a single-layer vanilla recurrent network, the standard surrogate for hippocampal–entorhinal computation in normative modelling work.

At each time step the network receives the sensory embedding produced by the frozen encoders concatenated with a three-dimensional self-motion vector, $\mathbf{x}_t = [\mathbf{e}_t; \mathbf{u}_t]$ with $\mathbf{u}_t = (v_x, v_y, \omega)$. Here v_x and v_y are the components of translational velocity, accumulated over the interval from t to $t + 1$, and ω is the total change in heading over the same interval. The hidden state is updated as

$$\mathbf{h}_{t+1} = \sigma(W_{ih}\,\mathbf{x}_t + \alpha\, W_{hh}\,\mathbf{h}_t)\,, \qquad \hat{\mathbf{e}}_{t+1} = W_{ho}\,\mathbf{h}_{t+1} \tag{4}$$

with σ the logistic sigmoid, $\mathbf{h}_t \in (0, 1)^{400}$ and $\mathbf{h}_0 = \mathbf{0}$. The biases are absorbed into the weight matrices without loss of generality. We set $\alpha = 1$ throughout training and in every analysis except the recurrence sweep of Figs. 2e and 3c, where it is set at test time, with no retraining, to interpolate between a purely feedforward code ($\alpha = 0$) and the trained network.

The RNN predicts sensory embedding one second ahead. In the separate-encoder regimes the target is the concatenation of the two 200-dimensional embeddings, and in the joint regime, this is the single 400-dimensional joint code; in the unimodal regimes, this is the 200-dimensional embedding of the available modality. The hidden layer is 400 units in every regime.

Whilst the recording is sampled at 50 frames per second, the network runs at 1 Hz. Each recording is decomposed into 50 interleaved streams, one per phase offset. For example, the first sees frames $1, 51, 101, \ldots$, the second frames $2, 52, 102, \ldots$, etc. The 50 streams form the batch dimension so every gradient step sees the same stretch of the recording at 50 different phases, and each stream is itself continuous in time, albeit correlated. With three training recordings the streams of all three are stacked, giving a batch of 150 rows, each row a continuous stream carrying its own hidden state of the same batch size.

Training minimises the L_1 distance between $\hat{\mathbf{e}}_{t+1}$ and the encoder's embedding of the true next input. Gradients are computed by truncated backpropagation through time on non-overlapping windows of 10 steps. The hidden state is initialised to zero once at the start of a recording and then carried from one window to the next, detached at each boundary so that gradients do not propagate beyond the window, but is never reset within a recording. We used an Adam optimiser with an initial learning rate of 3×10^{-4} with gradients clipped at norm 1, in 100 cycles of 10 epochs; the learning rate decays by 0.95 per cycle and by 0.999 per step, stopping at 10^{-6}. Networks are trained on the three two-hour recordings and evaluated after each cycle on the one-hour test recording; the checkpoint with the lowest evaluation loss is kept.

The untrained baselines of Fig. 2b are the same architecture with the same frozen encoders and the same inputs, but with W_{ih}, W_{hh}, W_{ho} and their biases left at their random initialisation. Three initialisations are run per regime and evaluated exactly as the trained networks are, and they establish what the allocentricity index returns before anything has been learned.

D THE SPATIAL ALLOCENTRICITY INDEX

The spatial allocentricity index (shortened to *allocentricity* in the main text) measures whether a unit fires in the same places regardless of which way the agent is facing. Figure S1 shows what it returns on manually constructed units and on the hidden units of the trained networks.

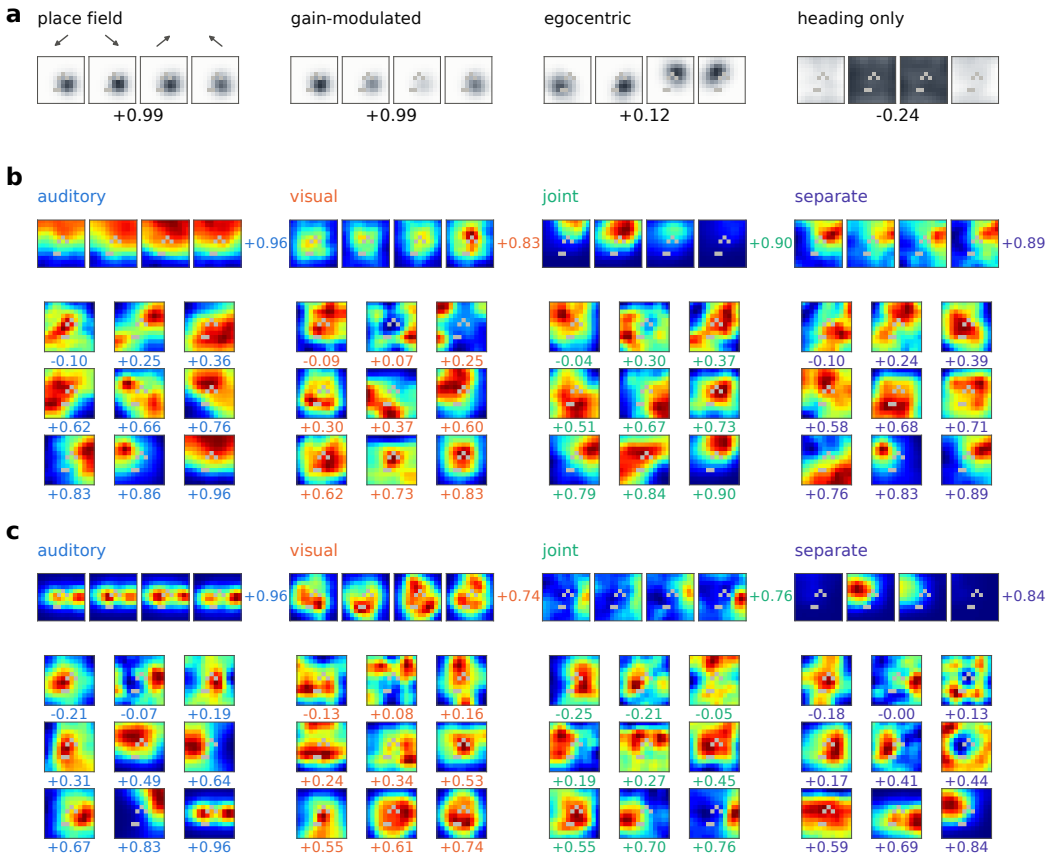

Figure S1: **Examples of allocentricity index and developed rate maps.** **a** Four manually constructed units. A place field fixed in the world scores 0.99. The same field with a heading-dependent gain also scores 0.99 as correlation is taken into account in calculating allocentricity. An egocentric field that moves with the agent scores 0.12. A unit tuned to heading alone, whose maps differ only in overall level, scores -0.24. **b** Nine randomly selected units from each regime in the quadrant arena, shown as heading-averaged (overall) rate maps with their a_i; above each regime, the four heading-specific maps of its most allocentric unit; **c** The same for the line arena.

Rate maps. The arena is divided into a 12×12 grid of spatial bins and heading into $N_\phi = 4$ bins of $90°$, taken on the smoothed direction of travel. For unit i and heading bin h, the map is the smoothed sum of activity divided by the smoothed occupancy,

$$f_i^h(b) \;=\; \frac{\big[G_\sigma * \sum_{t\in h} a_{i,t}\,\delta_{b(\mathbf{p}_t),\,b}\big](b)}{\big[G_\sigma * \sum_{t\in h} \delta_{b(\mathbf{p}_t),\,b}\big](b)}, \tag{5}$$

with G_σ a Gaussian kernel of $\sigma = 1$ bin, δ an indicator function of the agent's presence in that bin, and $a_{i,t}$ the unit's activity. Note that the numerator and denominator are smoothed separately and divided afterwards.

Support. We consdier a bin entering the comparison iff every heading bin visited it at least five times,

$$\mathcal{B} = \{\, b : n_h(b) \geq 5 \text{ for all } h \,\}, \tag{6}$$

and all maps are restricted to $\mathcal{B}$ (grey cells in Fig. S1). Position and heading are coupled by the trajectory, so each heading bin has a different set of unvisited bins; if unvisited bins are filled with zeros and smoothed, those shared holes are identical across headings and correlate perfectly, inflating the index by an amount that depends on occupancy rather than on the code. On the analysed test recordings, 142 of 144 bins survive the criterion (139 of 144 under the subsampling used for the per-unit analyses), so restricting to $\mathcal{B}$ discards almost nothing.

The Index. We define the per-unit index as the mean correlation between heading-conditioned maps, over the shared support,

$$a_i \;=\; \binom{N_\phi}{2}^{-1} \sum_{h<h'} \rho\Big(f_i^{h}, f_i^{h'}\Big)_{\mathcal{B}}\,, \tag{7}$$

where ρ is the Pearson correlation across bins. Because ρ is invariant to affine rescaling, a_i measures where a unit fires rather than how much, and gain-modulated units would retain a high a_i (Fig. S1a). We defined the index this way because heading-dependent gain is ubiquitous in hippocampal recordings and does not move the represented location, whereas a field that moves with heading does. Figure S1a shows four constructed units.

Noise ceiling. We consider a split-half reliability that is used in Section 2.4 to exclude unreliable units. This serves as a noise ceiling on a_i. A unit whose activity is noisy correlates poorly with itself as well as across headings, so a low a_i on its own does not separate a heading-dependent code from an unreliable one. We cut the recording cut into 20 contiguous blocks, blocks are assigned alternately to two halves, maps are computed separately within each half and correlated, and the result is averaged over heading bins. Blocks are contiguous because consecutive samples are 20 ms (at 50 fps) apart, and a random split would place near-duplicate samples in both halves and give a spuriously high ceiling. Units with a ceiling below 0.2 are excluded, as are units with an undefined a_i, which arises when a unit has no spatial variance in some heading bin. Across the 40 intact networks, between 366 and 399 of the 400 hidden units survive.

The value reported for a network in all four figures is the median over the surviving units,

$$A \;=\; \mathrm{median}_{i\in\mathcal{U}}\, a_i, \tag{8}$$

and values in the text are the mean $\pm$ standard deviation of A over the five seeds of a condition, compared with uncorrected t-tests. The median is used rather than the mean because the distribution of a_i is skewed and multi-modal in most networks (Fig. 2c), and because every regime contains units across the whole range in both arenas (Fig. S1b,c).

E Decoding position from the population code

In Fig. 2, Fig. 3, Fig. 4, we decoded position from the RNN hidden state by kernel ridge regression with an RBF kernel.

Let $\mathbf{h}_t \in \mathbb{R}^{400}$ (although, in the RNN each dimension was restricted to $(0, 1)$) be the hidden state and $\mathbf{y}_t$ the positional target (where y-position is explicitly stated, only the vertical coordinate y_t was used as the target). Samples are subsampled evenly along the recording to $n = 3000$ points, and each unit is standardised.

Writing Z for the standardised states, the Gram matrix $G = ZZ^\top$ is normalised to unit mean diagonal, squared distances are read off it as $d^2_{ts} = G_{tt} + G_{ss} - 2G_{ts}$, and

$$K_{ts} = \exp\!\left(-\frac{d^2_{ts}}{2\sigma^2_{\mathrm{K}}}\right), \qquad \sigma^2_{\mathrm{K}} = \mathrm{median}_{t<s}\, d^2_{ts}, \tag{9}$$

the bandwidth being set by the median distance rather than tuned hence every condition is decoded with the same instrument.

The decoder is fitted on a training set $\mathcal{S}$ and applied to a held-out set $\mathcal{T}$,

$$\boldsymbol{\alpha} = (K_{\mathcal{SS}} + \lambda I)^{-1}\,(\mathbf{y}_{\mathcal{S}} - \bar{y}_{\mathcal{S}})\,, \qquad \hat{\mathbf{y}}_{\mathcal{T}} = K_{\mathcal{TS}}\,\boldsymbol{\alpha} + \bar{y}_{\mathcal{S}}, \tag{10}$$

so the prediction at a test point is a similarity-weighted combination of training positions. The ridge parameter λ is chosen inside the training set by fitting on its first 80% and scoring the last 20% over $\lambda \in \{10^{-6}, \ldots, 10^{1}\}$.

We used five-fold cross-validation. Random folds would leak, since consecutive samples are 20 ms (at 50 fps) apart and a sample's near-duplicate would appear on the other side of the split. Predictions from all folds are assembled and scored once,

$$R^2 = 1 - \frac{\sum_t (y_t - \hat{y}_t)^2}{\sum_t (y_t - \bar{y})^2}, \tag{11}$$

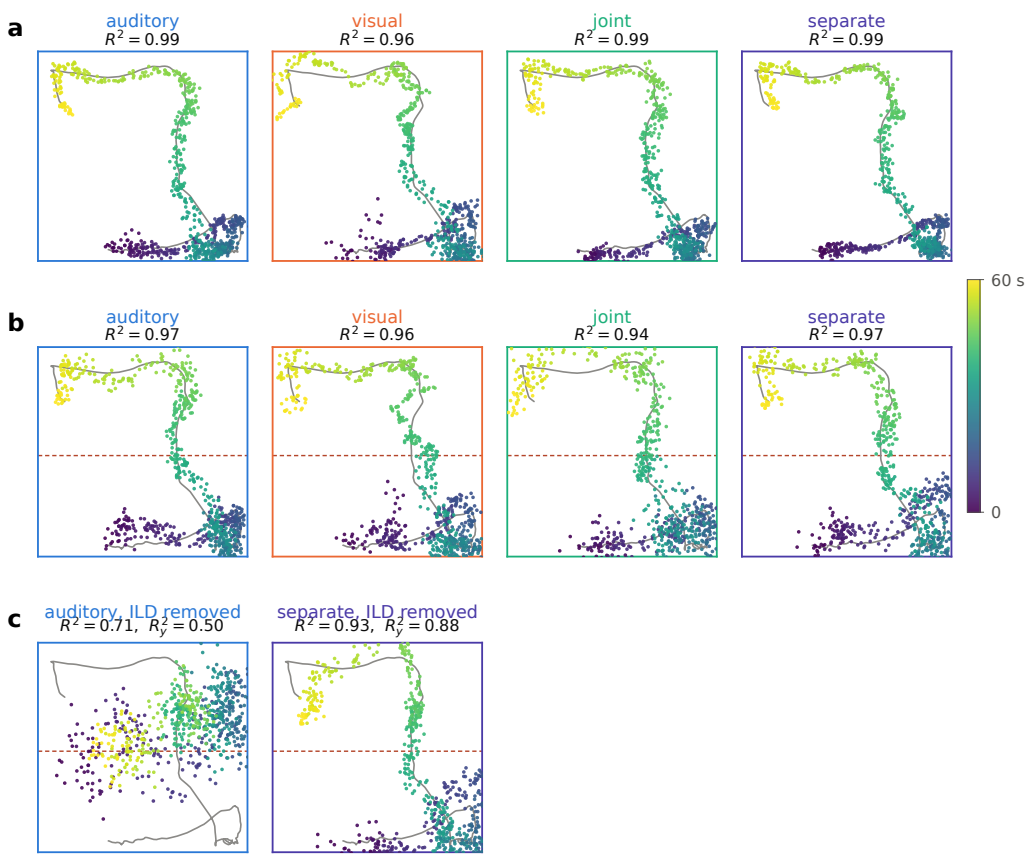

Figure S2: **Example of decoded trajectories.** Grey line, 60 s of the true trajectory; dots, the position decoded from the hidden state with RBF kernel regression, coloured by time. The R^2 above each panel is that appendix's five-fold value over the whole recording Panel borders and titles are coloured by regime; the dashed line marks the speaker line of the line arena. **a** Quadrant arena. **b** Line arena. **c** Line arena with the interaural level difference (ILD) removed at test time, by replacing the two channels of the log-magnitude spectrogram with their mean. In this case, the auditory network loses the axis.

therfore $R^2 = 1$ is exact recovery of position from the population state and $R^2 = 0$ is no better than predicting the mean position.The identical pipeline is applied to the raw sensory input, the encoder embedding, and the hidden state, so the three stages of Fig. 3b are comparable.

Decoded trajectories. Figure S2 as an example decodes the same hidden states to show what the recovered trajectory looks like. No temporal smoothing or Kalman filters are applied.

F Allocentricity and decodability are dissociable

A population can specify position exactly while containing no unit with a heading-invariant spatial field. We provide here in the Appendix:

- an example of a set of egocentric units providing positional information;
- a sketch proof of this claim.

Egocentric Example. Let the arena be a square centred at the origin and let the agent's heading be θ. Consider four units: two carrying the position of a landmark at the origin in the agent's own

reference frame,

$$\mathbf{u} = R_\theta^\top \mathbf{p}, \qquad R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \tag{12}$$

and two carrying heading, $\cos\theta$ and $\sin\theta$. Every unit is a function of egocentric quantities alone; nothing in the population is anchored to the world.

Position is nonetheless recoverable exactly, by undoing the rotation:

$$\mathbf{p} = R_\theta \mathbf{u} = \begin{pmatrix} r_3 & -r_4 \\ r_4 & r_3 \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \end{pmatrix}, \tag{13}$$

a quadratic function of the population activity, well within reach of the non-linear decoder of Appendix E. So $R^2 = 1$.

The allocentricity of those units is low by construction. The rate map of unit $i \in \{1, 2\}$ in heading bin θ_h is the ramp $f^h(\mathbf{p}) = (R_{\theta_h}\mathbf{e}_i)^\top \mathbf{p}$, whose gradient rotates with heading. On a centred square, where $\mathbb{E}[p_1^2] = \mathbb{E}[p_2^2]$ and $\mathbb{E}[p_1 p_2] = 0$, the correlation between two such maps is $\cos(\theta_h - \theta_{h'})$. With four heading bins the six pairs give correlations $\{0, -1, 0, 0, -1, 0\}$ and hence $a_i = -1/3$. Passing a unit of this kind through the estimator of Appendix D, built on the real trajectory, returns -0.33. The heading units have no spatial variance and are excluded by the ceiling criterion. See Appendix D.

Sketch proof of claim. We write the population response as $\mathbf{a} = \boldsymbol{\phi}(\mathbf{u}, \theta)$ with $\mathbf{u} = R_\theta^\top \mathbf{p}$, so that every unit is driven by egocentric position and heading and by nothing else. Let $\boldsymbol{\phi}$ be injective on the range the agent visits, which is the weakest sense in which the population can be said to carry its inputs. Then $(\mathbf{u}, \theta)$ is a function of $\mathbf{a}$, and by equation 13 so is $\mathbf{p} = R_\theta \mathbf{u}$. Position is therefore a deterministic, continuous function of the population activity, and a decoder flexible enough (such as using an RBF kernel) to represent it recovers position exactly in the noiseless case, regardless of the reference frame of the individual units.

G TRANSFORMED TARGETS IN UNIMODAL AND MULTIMODAL NETWORKS

Table 1: Oracle targets in unimodal visual networks. Median allocentricity A, mean $\pm$ s.d. over five networks per condition, in both arenas. The visual input is identical in the two arenas.

Target	Quadrant arena	Line arena
$\tilde{\mathbf{e}}^{\,\mathrm{avg}}$	0.789 ± 0.009	0.822 ± 0.012
$\mathbf{e}$ (untransformed)	0.363 ± 0.033	0.339 ± 0.042
$\tilde{\mathbf{e}}^{\,\mathrm{hres}}$	0.262 ± 0.011	0.293 ± 0.023

Table 2: Oracle targets in multimodal separate networks (quadrant arena). Median allocentricity A, mean $\pm$ s.d. over five networks, intact and the change Δ when one stream is replaced by its time-average (paired t-test against intact).

		Δ when removed	
Target	Intact	Vision	Audition
$\tilde{\mathbf{e}}_{\mathrm{vis}}^{\,\mathrm{avg}}$	0.775 ± 0.018	$+0.013$ ($p = 0.12$)	-0.100 ($p < 0.001$)
$\tilde{\mathbf{e}}_{\mathrm{aud}}^{\,\mathrm{avg}}$	0.832 ± 0.006	-0.028 ($p = 0.011$)	-0.020 ($p = 0.008$)
$\mathbf{e}$ (untransformed)	0.633 ± 0.012	$+0.047$ ($p = 0.006$)	-0.161 ($p < 0.001$)
$\tilde{\mathbf{e}}_{\mathrm{vis}}^{\,\mathrm{hres}}$	0.601 ± 0.020	$+0.048$ ($p = 0.008$)	-0.157 ($p < 0.001$)
$\tilde{\mathbf{e}}_{\mathrm{aud}}^{\,\mathrm{hres}}$	0.586 ± 0.020	-0.101 ($p < 0.001$)	-0.090 ($p = 0.001$)

Here we illustrate how transformed ('oracle') targets were computed from the embeddings of the whole recording.

Let $b(p) \in \{1, \ldots, 20^2\}$ index the spatial bin containing position p (a 20×20 grid, finer than the 12×12 grid of the allocentricity index) and $h(\theta) \in \{1, \ldots, 8\}$ the heading bin containing θ. We

computed the mean embedding in each spatial bin, $\mu(b)$, in each spatial and heading bin, $\mu(b,h)$, and overall, $\bar{\mathbf{e}}$:

$$\mu(b) = \mathbb{E}\big[\mathbf{e} \mid b(p) = b\big], \qquad \mu(b,h) = \mathbb{E}\big[\mathbf{e} \mid b(p) = b,\ h(\theta) = h\big], \qquad \bar{\mathbf{e}} = \mathbb{E}[\mathbf{e}]. \tag{14}$$

The heading-averaged target

$$\tilde{\mathbf{e}}^{\,\mathrm{avg}}_{t+1} = \mu\big(b(p_{t+1})\big) \tag{15}$$

is identical whichever way the agent faces within a spatial bin. The heading-residual target

$$\tilde{\mathbf{e}}^{\,\mathrm{hres}}_{t+1} = \mu\big(b(p_{t+1}),\, h(\theta_{t+1})\big) - \mu\big(b(p_{t+1})\big) + \bar{\mathbf{e}} \tag{16}$$

keeps the part of the embedding that deviates from its positional mean with heading, shifted by the overall mean $\bar{\mathbf{e}}$. As this deviation differs from place to place, the target depends on position as well as heading. In the multimodal networks of Section 3.5, the transform was applied to one sense's target only, and the other target was left unchanged. The results in unimodal visual networks and multimodal networks are reported in Table 1 and Table 2.

H HEADING-POSITION INTERACTION SHARE

Table 3: Position, heading and interaction shares of every prediction target, in both arenas, ordered by interaction share. Shares are noise-corrected and need not sum to exactly one, and a component that is pure noise can come out slightly negative.

	Quadrant arena			Line arena		
Target	pos.	head.	inter.	pos.	head.	inter.
$\tilde{\mathbf{e}}^{\,\mathrm{avg}}$ (visual)	0.99	0.00	0.01	0.99	0.00	0.01
$\tilde{\mathbf{e}}^{\,\mathrm{avg}}_{\mathrm{aud}}$ + vis	0.68	0.01	0.31	0.70	0.01	0.29
$\tilde{\mathbf{e}}^{\,\mathrm{avg}}_{\mathrm{vis}}$ + aud	0.58	0.07	0.35	0.53	0.08	0.39
$\mathbf{e}_{\mathrm{aud}}$ (auditory)	0.47	0.09	0.44	0.43	0.09	0.47
$\mathbf{e}$ (separate)	0.44	0.06	0.50	0.42	0.07	0.51
$\mathbf{e}$ (joint)	0.44	0.09	0.47	0.42	0.07	0.51
$\mathbf{e}_{\mathrm{vis}}$ (visual)	0.37	0.01	0.61	0.38	0.01	0.61
$\tilde{\mathbf{e}}^{\,\mathrm{hres}}_{\mathrm{vis}}$ + aud	0.35	0.07	0.58	0.34	0.08	0.58
$\tilde{\mathbf{e}}^{\,\mathrm{hres}}_{\mathrm{aud}}$ + vis	0.20	0.07	0.73	0.18	0.09	0.73
$\tilde{\mathbf{e}}^{\,\mathrm{hres}}$ (visual)	-0.03	0.02	1.01	-0.03	0.02	1.01

A prediction target can be characterised by how its structure divides between position, heading, and the two jointly ('interaction share'), without training networks. Here use the bins of the allocentricity index, the same 12×12 spatial grid, four heading bins and shared support (Appendix D), and take the mean target vector in each occupied cell, $\mu(b,h)$, its marginals $\mu(b)$ and $\mu(h)$, and the overall mean μ. We can decompose the target as

$$\mu(b,h) \;=\; \mu \;+\; \underbrace{[\mu(b) - \mu]}_{\text{position}} \;+\; \underbrace{[\mu(h) - \mu]}_{\text{heading}} \;+\; \underbrace{[\mu(b,h) - \mu(b) - \mu(h) + \mu]}_{\text{interaction}}, \tag{17}$$

and the variance of each term, summed over embedding dimensions, gives its share of the total.

A target of the form $\mu(b) + \mu(h)$ can be produced by a weighted sum of units tuned to position and units tuned to heading; the interaction term cannot be produced that way. A trained RNN must learn the conjunction of the two, namely, the directional field. The interaction share measures how much of a target obliges the recurrent state to hold position and heading together.

Noise correction. Sampling noise inflates the interaction term, which is the one of interest: with few samples per cell, independent noise appears as apparent structure in $\mu(b,h)$ but not in the marginals. Every term in equation 17 is therefore estimated as the covariance between two halves of the recording rather than as a variance within one. The halves are alternating contiguous blocks, as for the allocentricity noise ceiling, so that samples 20 ms apart do not appear on both sides of the split. Independent noise then contributes zero in expectation, and a target consisting of noise alone

returns a total of approximately zero. Shares are reported as measured, so a component that is pure noise can come out slightly negative, as the position share of $\tilde{\mathbf{e}}^{\,\mathrm{hres}}$ does in Table 3.

Table 3 gives the three shares for every condition in both arenas. The correlation between the interaction share and median allocentricity (Fig. 4f) is unchanged under the analysis choices: Spearman $\rho = -0.82$ with the default bins, -0.82 on a 10×10 grid, -0.82 on a 16×16 grid, -0.81 with six heading bins and -0.79 with eight ($p < 0.001$ throughout).

I COMPUTE AND REPRODUCIBILITY

We constructed the arenas in Unity version 2022.3.22f1 (DX11). For the rendering of the arena and training both the autoencoders and the RNNs, we used a machine with a single CPU of 2.10 GHz base speed (20-core, 28-thread, Intel Core i7-14700F), 128 GB RAM, and a single CUDA (version 13.2) GPU with 16 GB video memory (NVIDIA RTX 4080).