3D Memory Priors Reflect Communicative Efficiency not Statistical Frequency

Thomas Langlois[1,2*], **Thomas L. Griffiths**[1], **Nori Jacoby**[2]

[1] Princeton University. [2] Max Planck Institute for Empirical Aesthetics (MPIEA) * email: thomasalexandrelanglois@gmail.com

An essential function of the human visual system is to encode sensory percepts of complex 3-dimensional objects into memory. Due to limited perceptual resources, the visual system forms internal representations by combining sensory information with strong perceptual priors. We reveal detailed priors in memory for rotations of common everyday objects using data from 1150 respondents over Amazon Mechanical Turk (AMT) engaging in a serial reproduction task, where the response of one participant becomes the stimulus for the next. Successive reconstructions in the serial reproduction of 3D views of common objects reveal systematic errors that converge to stable estimates of the perceptual landmarks that bias memory. By sampling uniformly and densely over all rotations in $SO(3)$ we reveal perceptual landmarks in memory that eluded past experimental approaches. The data challenge explanations based on statistical learning ("Frequency hypothesis"). Instead, we propose that memory priors reflect communicative efficiency rather than natural image statistics. We show that optimizing the Information Bottleneck (IB) trade-off between the complexity and accuracy of view reconstructions using a communication model in which views are represented as distributions over a semantic space produce distortions that align with memory biases.

The human visual system is remarkable for its capacity to encode and retain an extraordinary amount of visual information [1]. At the same time, this capacity belies a very selective allocation of bounded perceptual resources during visual encoding and memory formation [2, 3]. This process often leads to simplified and biased internal representations [4, 5, 6, 7, 8, 3, 9, 10, 11, 12, 13]. A key function of the visual system is to form accurate internal representations of complex 3D visual objects in the world. Because 3D objects can vary widely in appearance depending on viewpoint, the visual system must develop stable and invariant internal representations that are robust to viewpoint changes. At the same time, possessing view-*dependent* representations of objects in space is also critical for supporting visuospatial memory and navigation. Although canonical views [14] in 3D object perception are well-known, biases in memory for object rotations have not been extensively investigated in full $SO(3)$ (where $SO(3)$ corresponds to the group of all possible 3D rotations of an object). Previous work uncovered the perceptual landmarks that anchor visuospatial memory estimates of 2D locations inside images [12] and demonstrated the utility of combining iterative experimental paradigms based on serial reproduction with large-scale crowdsourcing experiments to reveal intricate structure in perceptual representations. We used the same approach to probe perceptual priors in a more complex and more ecologically valid domain, by measuring biases in the reproduction of 3D object views sampled uniformly over the full domain of 3D rotations in $SO(3)$ and by amplifying biases in memory with serial reproduction.

Serial reproduction is an experimental technique that is ideally suited for quantifying biases in visual and auditory perception [12, 13, 15, 16]. It mimics the children's game of telephone, where the response of one participant in an experimental task, such as a visual memory task, becomes the stimulus for a new participant in a chain of participants. Repeating this process multiple times amplifies collective memory biases, and can be shown to converge on shared priors in a Bayesian model of perception under experimentally verifiable assumptions [12]. By iterating a simple memory task, in which participants reproduce rotations of objects over multiple iterations, we reveal collective priors in view-dependent representations of complex 3D objects in unprecedented detail, which show highly structured and intricate patterns of biases towards clear perceptual landmarks. These landmark views deviate from well documented canonical effects in perception for 3D objects, revealing that representations in memory do not necessarily reflect the most "typical" views of the objects [14]. Past work has shown that perception for 2D rotation is biased towards horizontal and vertical "cardinal" orientations, and that these biases match the distribution of local orientations in natural images [17]. However, this work investigated biases for the orientation of 2D oriented Gabor functions, and not complex objects rotated in full 3D. Although traditional accounts of view-dependent effects in 2D and 3D object perception tend to explain biases in perception in terms of a constructive bottom-up inferential process [14, 18, 19] shaped by regularities in natural image statistics ("Frequency hypothesis"), evidence for top-down influence of non-perceptual states, language, and category representations on visual perception and memory is also widespread [20, 21, 22, 23]. These different explanations for the structure of internal priors are illustrated in Fig. 1.

While semantic object-level categories necessarily depend on low-level sensory representations shaped by the statistics of the natural environment, they may also reflect "communicative need" which refers to factors such as capacity constraints and linguistic usage that constrain the structure of semantic representations [24, 25]. Because representations in memory are

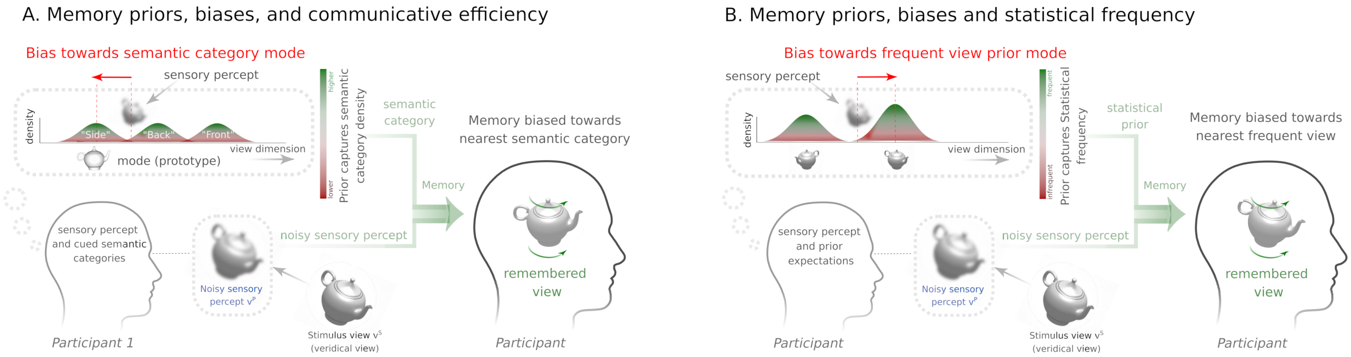


Figure 1 Visual memory priors, frequency hypothesis (typicality), and communicative efficiency hypothesis. A. Communicative efficiency hypothesis. When remembering an object in space, a participant encodes a sensory percept V_t^P (fine-grained memory trace) of the object view, and also possesses semantic visuospatial categories for that same view (e.g. it is a "right profile" view or it is a "tilted" view). Memory for the object view is biased towards the view corresponding to the nearest semantic category center (mode). B. Frequency hypothesis (typicality). When remembering an object in space, a participant encodes the sensory percept V_t^P of the object view, and it is biased in memory towards the nearest prior mode that is proportional to how typical the view is in the visual environment (Frequency hypothesis).

one step removed from the proximal stimulus being remembered, they tend to be systematically biased. Is memory biased towards representations that reflect typicality [14] or natural image statistics ("Frequency hypothesis")? or is it biased towards representations that optimize communicative accuracy and efficiency? In particular, do biases in memory reflect the geometry of latent semantic representations?

Recent work using the Information Bottleneck (IB) principle showed that the emergence of semantic category structure can be predicted from an optimal tradeoff between the efficiency and accuracy of compressed representations of perceptual spaces in multiple domains including color [26, 27, 28]. Here we propose a model that predicts structured priors in memory for rotations of objects as a consequence of Bayesian inference, where noisy sensory percepts are combined with efficient partitions of semantic view representations estimated from near optimal compression [27]. Our results indicate that memory priors in this domain are biased towards simplified visuospatial representations that optimize the trade-off between accuracy and efficiency in information transmission. Our results also show that both biases in memory and biases predicted by our simulations converge towards view representations with lower entropy in the semantic space, a finding that is consistent with the notion that priors reflect communicative efficiency and nameability rather than typicality [14] or statistical frequency.

Results

Revealing visual memory priors for object rotations in $SO(3)$

We ran a series of large scale serial reproduction experiments in which participants reproduced rotations of objects in $SO(3)$ as accurately as possible (Fig. 2A). In each trial, participants saw an object for 4000 ms and were asked to memorize its orientation. Following a 1000 ms retention period without an image, the object was presented in another random orientation, and the participant rotated the object to the remembered position. Only accurate responses within a small margin of error were retained, and participants received a bonus that was proportional to their accuracy in the task (see Methods and SI Appendix). The response orientation then became the stimulus for another participant in the chain.

Fig. 2B illustrates the axis angle representation of a rotation in $SO(3)$. The view from a camera positioned on a sphere pointed towards an object at its center can be described by its azimuth φ and elevation θ on the sphere, and its local orientation α about the position axis (see Fig. 2B for illustration). We used 8 detailed grayscale 3D models of common everyday man-made objects, including a shoe, a teapot, van, clock, camera, coffee maker, motorcycle, and grand piano (see Fig. 2C). The choice of objects was based on the objects used in early studies of canonical effects in 3D perception [14, 29], and the particular models used were chosen because of the level of detail and verisimilitude to their real-life counterparts. We ran 500 chains for each of the objects, and each chain was initialized as a rotation sampled uniformly in $SO(3)$ using the sampling method described in [30]. We ran the chains for 20 iterations based on the results in [12], and observed convergence of the chains by the 11th iteration of the serial reproduction process for most objects (see SI Appendix Fig. S7). Fig. 2D shows positional Kernel Density Estimates (KDEs) of the camera positions in the initial seed iteration, and at iterations 5, 10, 15, and 20 in the chains for four objects (See SI Appendix Fig S1, Fig. S2, Fig. S3, and Fig. S13 for all positional KDE results and raw data for all objects). Fig 2D also

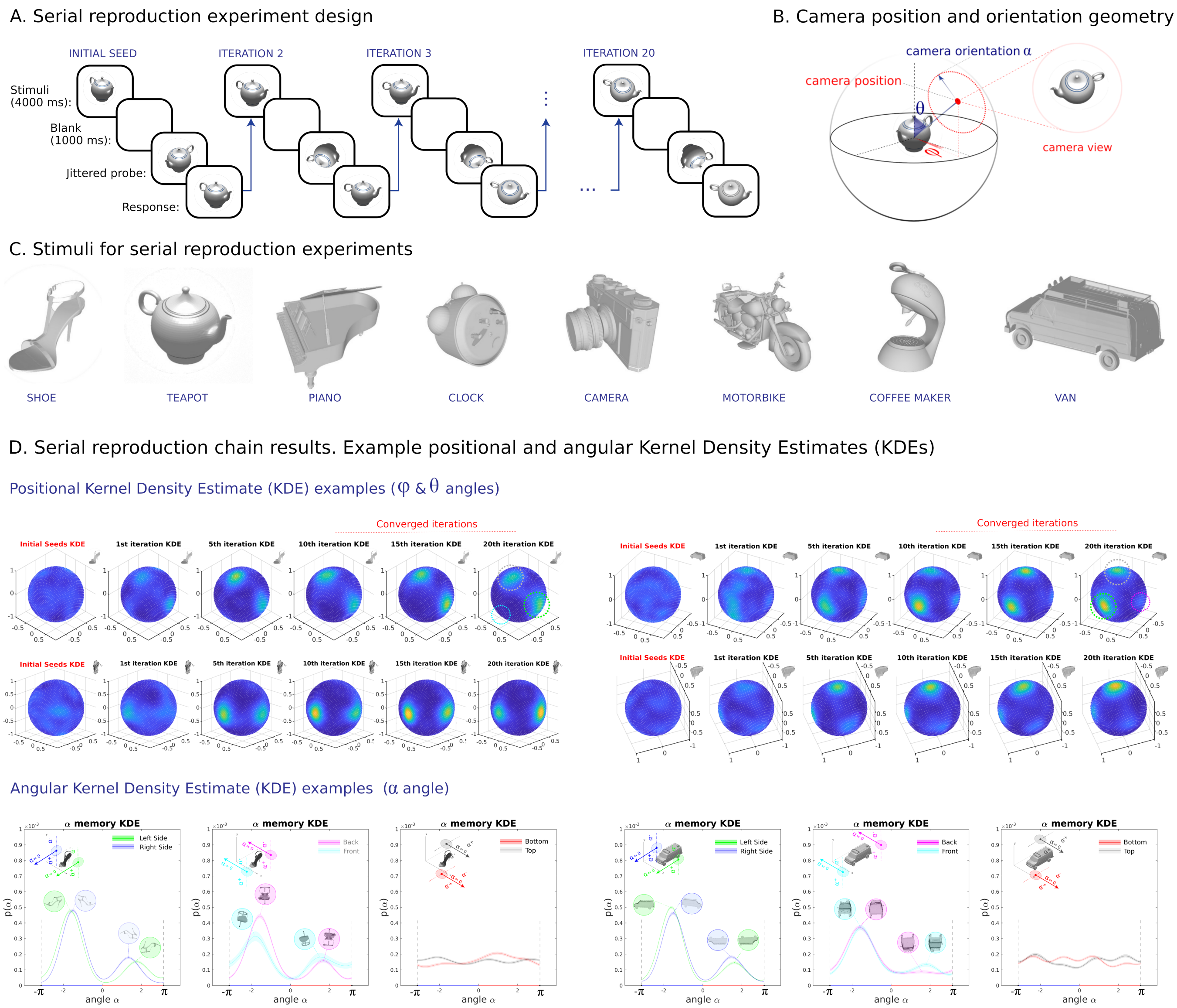


Figure 2 Serial reproduction of views sampled uniformly from the rotation group $SO(3)$. A. Serial reproduction design. Each of the 500 chains in the experiment is comprised of 20 individual nodes. In each node (trial), a unique participant views a rotated object for 4000 ms. Following a blank delay lasting 1000 ms, the object is redrawn on the screen in a completely random rotation. The participant rotates this jittered probe to match the initial stimulus rotation. Participants could use as much time as they needed to complete a trial. if the response was within a small enough margin of error ϵ, this response was routed to a new participant on AMT as the stimulus view. This process was repeated for a full 20 iterations. B. Axis-angle representation of 3D rotations. In polar coordinates, a camera position is defined by the azimuth φ and elevation θ of the camera position on the sphere, and the camera orientation is defined by an angle of rotation α. C. Stimuli. We used 8 grayscale detailed 3D meshes of common objects. D. Memory position (φ & θ) and angular (α) Kernel Density Estimates (KDEs) of the serial reproduction results for four objects. Top rows show KDEs of the memory position results in the seed, 5th, 10th, 15th, and 20th iteration of the process. Insets show the corresponding objects presented in the same orientation as the KDEs for reference. The results show clear convergence towards landmarks. The colored dotted circles illustrate the positional modes with the same color coding scheme used for the angular KDE results shown in the bottom row. Bottom shows KDEs of the memory orientation results in the modes of the convergent positional KDEs. These show biases towards upright and upside-down views for side views, and front vs. back views, but not the top vs. bottom views.

shows KDEs of the local orientations (α angles) for all responses in the final (convergent) iterations of the serial reproduction process inside the modes in the positional KDEs (see SI Appendix Fig. S4, Fig. S5, and Fig. S6 for all KDEs and results of local orientation biases).

The results reveal intricately structured priors emerging by the first few iterations indicating systematic biases in both the remembered view positions and local view orientations in memory. Positional biases revealed a strong tendency for participants to produce rotations towards the primary faces of the objects. In some cases, these were similar to estimates of the canonical perspectives of the objects, although unlike canonical perspectives, memory is also systematically biased towards the top, bottom, front and back faces of the objects (views that are neither the "best" nor the most "typical" views of the objects [14]). The orientation biases (Fig. 2D, SI Appendix Fig. S4, Fig. S5) also show intricacies that are inconsistent with canonical perspectives. In particular, for all the objects we observed a consistent pattern showing a bimodal distribution for side views and the front and back views. Inside the side view positional KDE modes, the strongest angular KDE mode reveals upright orientations of the objects, but the second mode consistently reveals upside-down views for all the objects. Finally, although KDEs of the positional data reveal strong top and bottom views for many of the objects, angular KDEs of the distributions of local orientations inside these modes tended to consistently show a more uniform distribution, indicating that anchors in memory for the top and bottom views did not show as clear a differentiation.

Our results do not show a bias to both vertical and horizontal "cardinal" orientations, so an explanation based strictly on the distribution of local orientations in natural images seems unlikely [17]. A related idea is that our results simply reflect the statistics of object poses, but they deviate significantly from well-documented view dependent effects in 3D object perception and do not appear to reflect views that are judged to be the most typical or commonly experienced ("Frequency hypothesis" [14]). So what explains these patterns in memory? One possibility is that memory is biased towards semantic visuospatial categories that maximize the nameability and communicative efficiency of the visual percepts. Such an explanation would be more in line with the "Maximal information hypothesis" originally proposed by [14]. This hypothesis states that canonical views are views that are maximally informative about the 3D structure of the objects. In a slight reinterpretation of this idea, we propose instead that the landmark views in memory correspond to views that optimize both accuracy and efficiency in *viewpoint* reconstruction via language. In order to do this, we ran an additional series of experiments. In the first experiment, we obtained a lexicon of view-based words. We then ran an n-Alternative Forced Choice (nAFC) experiment for each object in which participants labelled views using a subset of the most frequent view words obtained in the first experiment. Next, we derived estimates of semantic visuospatial representations of views (view spaces), and simulated memory reconstructions of view positions using only the labelling data. In what follows, we describe how we estimated the view spaces for each object, and present simulations of a communication model based on the Information Bottleneck (IB) principle [26], including a quantitative evaluation of its predictions with respect to memory for view position.

Estimating semantic spatial categories & representations

We started by compiling a list of common words used to describe views (see Fig. 3A for an ordered set of the list of the most common visuospatial words that were provided by 50 participants recruited from AMT). We retained only the words that were volunteered at least twice by different participants on AMT. We then grouped these words into 22 unique words (see Fig. 3A-B). In a second experiment, and for each of the 8 objects, we obtained naming distributions from a separate group of 243 participants who labelled 1944 views sampled uniformly in $SO(3)$ for each of the objects (see Fig. 3B for an illustration of the nAFC experiment design, where participants could choose from the 22 word categories). For each of the 1944 views we obtained nAFC responses from 9 unique participants and averaged them for each view. We then defined a regular spherical Fibonacci lattice of J = 324 view positions over the 2-sphere. For each of these J lattice points, we computed a weighted average of the naming distributions for nearby views in the set of 1944 unique views. For the weighted average, we used a fixed Gaussian smoothing kernel with $I_3 \cdot \sigma = 0.2$, for all the objects. We then computed the pairwise similarity between all the J naming distributions for each of the J grid points by computing the Jensen-Shannon Divergence (JSD) between all pairs of normalized naming distributions to estimate the internal "view-space" representations of the objects, where Euclidean distances are proportional to semantic similarity (see Fig. 3D for 3D projection of an example view space, and SI Appendix Fig. S9D-E and Fig. S10C for all object view spaces).

Fig. 3A-B illustrates the experimental pipeline used to estimate view spaces for each of the objects. Fig. 3C shows 3 example naming distributions for 3 views of the shoe object. Similar views (view 1 and view 2 shown in green) yielded a low JSD between their corresponding naming distributions (higher semantic similarity), while JSDs between these similar views and a very different view (comparing view 1 to view 3, and view 2 to view 3) produced higher JSDs (lower semantic similarity, shown in red for view 3). Fig. 3D shows the 3D projection of the 22D view space for the same object, where the pairwise Euclidean distances between all pairs of J views are proportional to the semantic similarity between their corresponding naming distributions. the pairwise semantic similarity between views 1, 2, and 3 shown in Fig. 3C are reflected by the pairwise Euclidean distances in the view space shown in Fig. 3D. SI Appendix Fig. S10C shows 3D projections of the view spaces we estimated for all the objects. We estimated view spaces for all the objects using only the naming frequency data, and they reveal variations in overall geometry depending on the object. In what follows, we describe how we use them to make model-

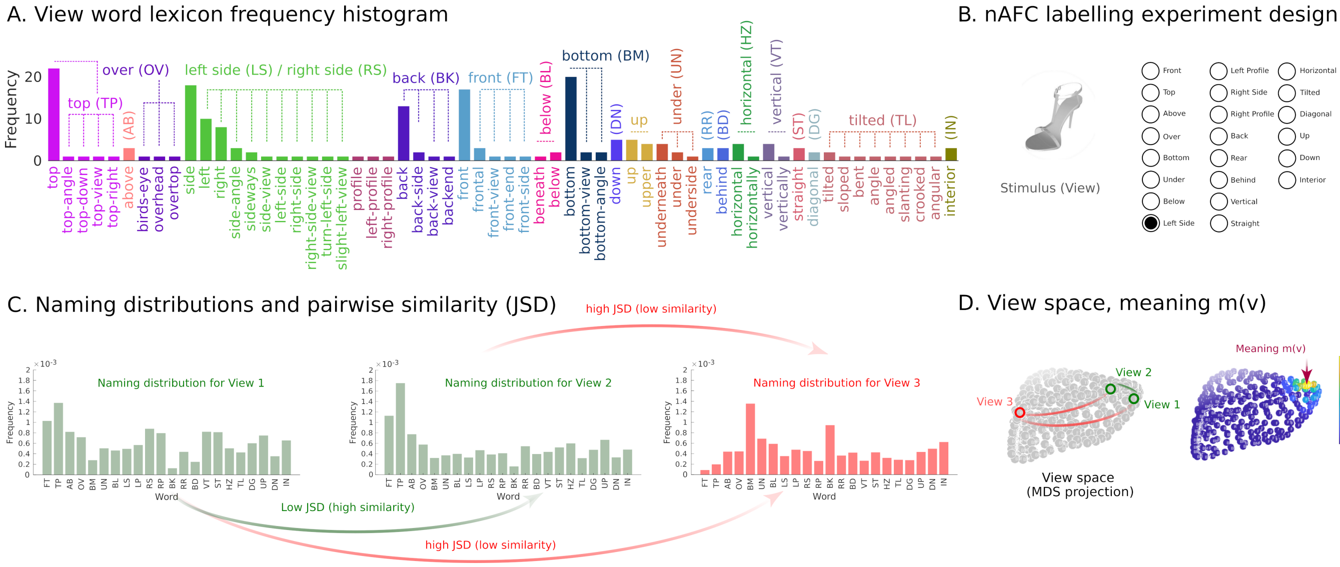


Figure 3 Estimating word-based semantic view spaces. A. Ordered frequency histogram of the most common visuospatial words volunteered by participants on AMT, grouped into 22 basic word categories (we treated “left-profile” and “right-profile” or “left-side” and “right-side” as separate categories). B. Illustration of nAFC labelling experiment, in which participants selected the best word to describe 1944 unique views of an object sampled uniformly in $SO(3)$. C. Computing pairwise semantic similarity between views. We computed the pairwise semantic similarity between object views using the Jensen-Shannon Divergence (JSD). We calculated the JSD between all pairs of normalized naming distributions estimated for 324 regularly spaced view positions over a regular lattice on the 2-sphere. Similar views (views 1 and 2) tended to have similar naming distributions, and produced a lower JSD (higher similarity), while dissimilar pairs of views (view 1 and view 3, or view 2 and view 3) produced higher JSD values, because their corresponding naming distributions were less similar. D. A 3D projection of the full view space for the same object reveals a highly structured internal geometry, where pairwise Euclidean distance is proportional to semantic similarity. A meaning $m(v)$ is defined as an isotropic Gaussian centered on a view v in the space, and captures a person's subjective belief over all views of having perceived a particular view v.

based predictions about how communicating efficiently about perceptual states in the semantic space can lead to biases in serial reproduction.

Priors, meaning, and communicative efficiency

Much work to date has revealed evidence that the structure of perceptual categories are correlated with memory and perception in a variety of visual and non-visual domains [22, 21, 26]. We investigate whether the geometry of language-based semantic representations of the objects derived from the naming data (“view spaces”) predict memory biases on the 2-sphere (biases in reproduction of the view positions). We define a “meaning” as a distribution in a semantic space that represents a perceptual state that a speaker wishes to convey to a listener through language (such as a remembered view of an object). We used a communication model based on the Information Bottleneck developed by [26] to further test if optimal compression of meanings in the view space representations of objects produces biases in reproduction that can account for the systematic memory errors we observed in the serial reproduction experiments.

Model

Communicating about perceptual states

In order to test whether visual memory priors optimize communicative accuracy and efficiency, we start by defining the “meaning” associated with a view v to be the language-based representation associated with a view. Formally, we define it as an isotropic Gaussian centered at v in the internal (semantic) view space using the approach developed by [26], where $m(v) = \exp(-\frac{1}{2\sigma_p^2}||\hat{v} - v||^2)$ is a distribution over the set of all views $\hat{v}$ centered at v which captures a person's internal belief over all views of having perceived a particular view v, where the pairwise distances between possible views are proportional to semantic similarity (see Fig. 3C-D). The parameter σ_p captures the level of perceptual noise in the internal representation, and we assume uniform internal noise across all possible views (fixed σ_p regardless of view in the internal space).

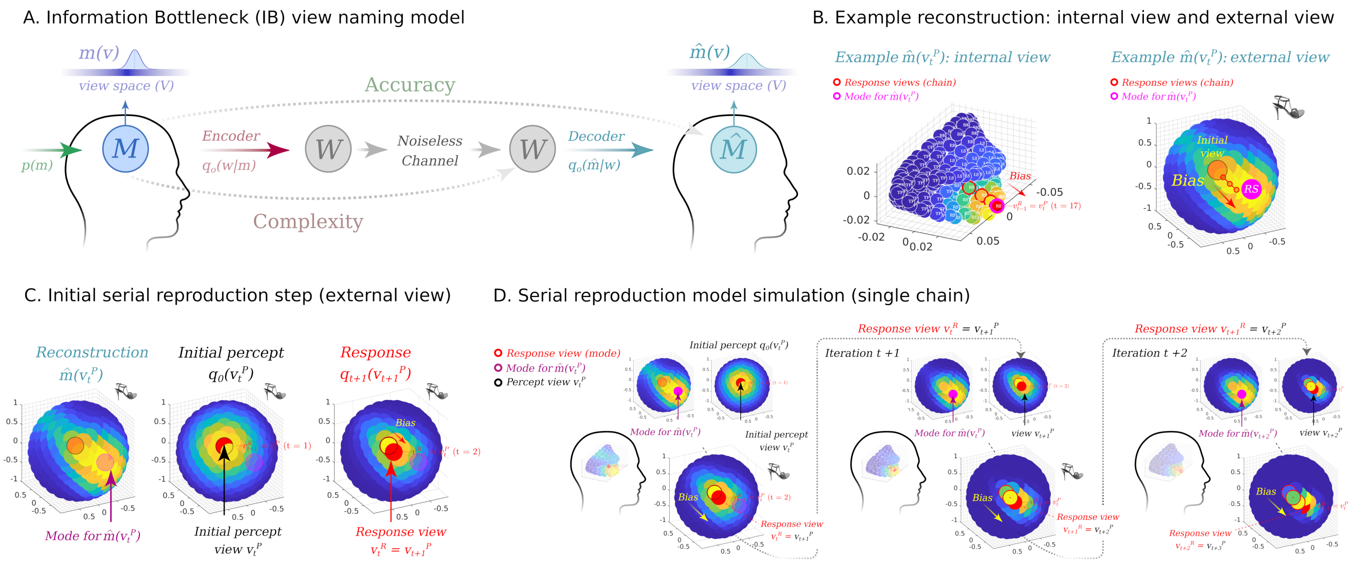


Figure 4 The Information Bottleneck (IB) communication model, and serial reproduction chain simulation. A. The IB view naming model. For a given object o, a speaker conveys a meaning $m(v)$, which is a distribution centered on a view v in an internal semantic space, via a stochastic encoder $q_o(w|m)$. The listener is an idealized Bayesian observer who generates a reconstruction $\hat{m}(v)$ from the speaker's meaning $m(v)$ by inverting the encoder via a decoder $q_o(\hat{m}|w)$. Given a single parameter β, IB specifies the optimal tradeoff between maximizing the accuracy of the reconstructions while also minimizing the complexity of the compressed representation in W. B. Two equivalent views of an example reconstruction $\hat{m}(v)$. Reconstructions $\hat{m}(v)$ can be plotted as distributions in the internal view space (left) as well as over the external (Euclidean) view space on the 2-sphere (right). C-D. Modeling a single reproduction chain. C. The initial percept is modeled as a Gaussian centered on a true stimulus view position v_t^S. This stimulus view position v_t^S is also mapped to its semantic reconstruction $\hat{m}(v_t^S)$, which is determined by the IB communication model. We then model a response ($v_t^R = v_{t+1}^P$) as the mode of the elementwise product between these two distributions ($q_{t+1}(v_{t+1}^P)$). The resulting shift in the view reveals a bias towards the mode of the IB reconstruction for the initial view. The response v_t^R becomes the stimulus for the next simulated participant in the chain. D. Example serial reproduction chain. The beginning of the chain shows the same distributions as in panel C. Subsequently, the process is repeated with the response view from the previous iteration $v_t^R = v_{t+1}^P$. Multiple iterations of this procedure produces biases towards the modes in the IB model's semantic reconstructions for views.

Information Bottleneck (IB) naming model

We adopted a communication model developed by [26]. The model is based on Shannon's classical communication model and involves a speaker and a listener (Fig. 4A). Meanings are represented as distributions over a finite set of possible views $v \in \mathcal{V}$ for an object o. A view perceived by the speaker can be any rotation of an object on the 2-sphere (camera positions) and the view that a speaker wishes to transmit to a listener is a meaning $m(v)$ over $\mathcal{V}$. In practice, each view v in $\mathcal{V}$ corresponds to a point in the 22D semantic view-space representation we estimated using the naming data (see Fig. 3D for an example 3D projection of a view space). As described in the last section, a meaning $m(v)$ is an isotropic multivariate Gaussian distribution with a diagonal covariance matrix Σ_p and mean centered on view v in the internal space. A meaning $m(v)$ can be interpreted as the speaker's subjective belief about the state of the environment, which corresponds to the rotation of an object in space. Communicating a meaning $m \in \mathcal{M}$ indicates that the speaker wishes to communicate a belief that $\mathcal{V} \sim m(v)$.

Although [26] defined a "cognitive source" $p(m)$ that specifies the probability of intended meanings for the speaker, we assume a uniform distribution $p(m)$ over the set of possible meanings $\mathcal{M}$ that the speaker wishes to communicate. This choice makes minimal assumptions about the distribution of intended meanings that a speaker could wish to transmit to a listener, although the model can be extended to capture any systematic variation in the probability of intended meanings by incorporating information about the frequency with which different views of an object tend to be encountered in the world [14, 29], or views that are more likely to be needed based on task demands that might be associated with a given object or perceptual decision-making task. The "cognitive source" $p(m)$ can be interpreted as a distribution over meanings that is fully extrinsic to the speaker and listener, and is intended to capture any systematic variability in the probability of an intended meaning that is based strictly on contextual factors [31] rather than capacity constraints or internal priors.

In the model (see Fig. 4A), the speaker uses a stochastic naming policy $q(w|m)$ to compress a meaning m with a word w from a lexicon W of size $|W|$. In our formulation, the encoder compresses views into semantic visuospatial categories derived from linguistic descriptors. Like [26], we assume an idealized noiseless channel (we focus on modeling perceptual biases rather than noise or errors due to external or contextual factors [31]), and we set no constraints on the lexicon size. When the speaker compresses a meaning m into a word w, the listener infers a meaning $\hat{m}$ based on a decoder $q(\hat{m}|w)$, and we assume an optimal Bayesian listener with respect to the speaker. Fig. 4A illustrates the naming model setup. This model has a single parameter β that controls the tradeoff between accuracy and efficiency in the IB optimization (see section below and SI Appendix for details). Aside from this, the encoder and decoder estimates depend only on the geometry of the semantic view space representations estimated from the pairwise JSDs of the normalized naming distributions, the internal noise parameter σ_p, and the choice of the lexicon size, which determines the upper bound on the set size of W. For all the objects, we did not set constraints on the lexicon size, and allow a one-to-one correspondence between the full set of J views in V and the set of possible words in W (e.g. $|W| = |V|$). In the limit for large values of β, the encoder becomes a $J \times J$ identity matrix when the lexicon size of W equals the total number of views v in $\mathcal{V}$.

The IB efficiency and accuracy tradeoff

In IB [32, 26], the complexity of a lexicon of words $\mathcal{W}$ is measured by the number of bits that are required to represent a speaker's intended meaning $\mathcal{M}$ by $\mathcal{W}$ using the stochastic encoder $q(w|m)$, and it is given by the information rate, which measures the mutual information between the original meanings and the compressed representation (quantization) by W:

$$I_q(M;W) = \sum_m \sum_w q(w|m)p(m) \log \left[\frac{q(w|m)}{q(w)}\right] \tag{1}$$

where $q(w) = \sum_m p(m)q(w|m)$. Complexity is maximized when the encoder is a $J \times J$ identity matrix, where the lexicon size of all the words in $\mathcal{W}$ equals the set size of the views v in $\mathcal{V}$, and each meaning $m(v)$ is mapped to a unique word w. Conceptually, this corresponds to a scenario where a speaker has a unique word to describe each and every possible psychological percept (object views in our case). Such a naming policy maximizes accuracy in information transmission, but at the expense of efficiency, since developing a naming policy that maps every shade of perceptual experience to its own unique word quickly becomes impractical and computationally costly in practice.

At the other extreme, an alternative to maximizing complexity consists in mapping every psychological percept to a single word. While maximally efficient, this strategy comes at the expense of making a speaker unable to convey any meaningful information about different perceptual states to a listener, and results in a total distortion of intended meanings, which minimizes accuracy. In the IB optimization, the Kullback-Leibler (KL) Divergence emerges as the natural distortion measure [32], and the distortion of a meaning $m(v)$ is given by:

$$D[m||\hat{m}] = \sum_v m(v) \log \frac{m(v)}{\hat{m}(v)} \tag{2}$$

where $\hat{m}_w(v) = \sum_m q(m|w)m(v)$ and the expected distortion over all possible meanings $m \in \mathcal{M}$ is given by:

$$\mathbb{E}_q\Big[D[M||\hat{M}]\Big] = \sum_m \sum_w p(m)q(w|m)D[m||\hat{m}_w] \tag{3}$$

The expected distortion of intended meanings is inversely proportional to overall accuracy, which is given by:

$$I(W;V) = \Big(I_q(M;V) - \mathbb{E}_q\Big[D[M||\hat{M}]\Big]\Big) \tag{4}$$

Between the extremes of maximizing accuracy at the expense of a maximally complex naming policy, or using only a single word to describe all possible internal states (maximal compression), the IB framework specifies the naming policy that achieves an optimal tradeoff between accuracy and efficiency, which can be obtained by minimizing the IB Lagrangian for a given value of the tradeoff parameter β (See SI Appendix and Fig. S11 for details):

$$\mathcal{F}_\beta\left[q(w|m)\right] = I_q(M;W) - \beta I_q(W;V) \tag{5}$$

We use the naming model to simulate internal states that are fully specified by *semantic* representations of views over objects determined entirely by linguistic descriptors (Fig. 3). As such, the model describes views v based strictly on words, and not on any visual features of the objects, or a a representation based on pairwise similarity judgments as in [26]. In what follows, we describe how we extended the model to simulate how intended meanings $m(v)$ become biased by iterated reproduction in the semantic space.

Modeling visual memory and serial reproduction

For each chain, we model the first step (trial response) in the serial reproduction experiment in the following way: A participant initially perceives a view v_t^P centered on a Gaussian $q_0(v_t^P)$ with a mean stimulus view position v_t^S, and fixed isotropic noise σ_s (e.g. $q_0(v_t^P) = \mathcal{N}(v_t^S, \sigma_s)$). This same view v_t^P is also mapped to the reconstruction $\hat{m}(v)$ produced by the naming model's reconstruction of $m(v)$ where $v = v_t^P$. Conceptually, $\hat{m}(v_t^P)$ can be interpreted as a simplified *language-based* reconstruction of the meaning $m(v_t^P)$ for the stimulus view v_t^P based on the semantic visuospatial categories that the participant possesses for that object, and that are cued by v_t^P. These categories are determined by the IB compression of meanings in the view space for that object. Note that v_t^P can also be modeled as a random sample from $q_t(v_t^P)$ (e.g. $V_t^P \sim \mathcal{N}(v_t^S, \sigma_s)$).

The participant combines both the initial distribution $q_0(v_t^P)$ over views centered on v_t^P and the language-based reconstruction $\hat{m}(v_t^P)$ by an element-wise product of the two distributions. We then model a memory response v_t^R in a chain as the mode of the resulting distribution $q_t(v_t^P)$. Subsequently, the next step in the chain combines $q_t(v_t^P)$ from the previous step with $\hat{m}(v_{t+1}^P)$, where the new stimulus view $v_{t+1}^P = v_t^R$, which is the response from the previous iteration (Fig. 4D), resulting in a new distribution $q_{t+1}(v_{t+1}^P)$. The response v_{t+1}^R corresponds to the mode of $q_{t+1}(v_{t+1}^P)$, and it becomes the new stimulus v_{t+2}^P in the chain.

With each step, the reconstruction reveals a systematic bias towards the modes in the naming model reconstructions (see Fig. 4B-D for a representative example, SI Appendix S12, and SI Appendix Movie S1 for animations of full chain dynamics for a handful of chains). Formally, after the first iteration, each step in a single chain is modeled as follows:

$$q_{t+1}(v_{t+1}^P) = q_t(v_t^P) \circ \hat{m}(v_{t+1}^P) \tag{6}$$

where the symbol "$\circ$" denotes the element-wise product, and the stimulus view v_{t+2}^P in the next step is equal to the following:

$$v_{t+2}^P = v_{t+1}^R = \arg\max_v \big(q_{t+1}(v_{t+1}^P)\big) \tag{7}$$

This process simulates memory within a single chain as a biased reconstructive process, where the sensory information presented to each subject in the chain, which is captured by the $q_t(v)$ distributions, becomes progressively biased towards compressed semantic categories (specified by the stochastic encoder $q_o(w|m)$).

In order to simulate full KDEs for every step in the serial reproduction process, we repeated this procedure using all the J lattice points v on the 2-sphere as the initial chain seeds, and marginalized over all the $q_t(v)$ distributions at every iteration t. Fig. 5A shows the simulated KDEs for several objects alongside the original memory results.

Model fitting, predictions and results

The model has only two parameters, which we fit to the memory data for each of the objects: the perceptual noise parameter σ_p and the IB tradeoff parameter β. We used the same initial sensory noise parameter σ_s for all objects to avoid introducing additional degrees of freedom to the simulations. For each of the objects, we fit the parameter settings of σ_p and β that maximized the correlation between the memory KDEs (concatenated across all iterations) and the KDEs produced by the simulation, which we also concatenated across all iterations (see Fig. 5A, which shows the simulated memory KDEs for four objects along with the actual memory KDEs. For all the results including permutation test results, see SI Appendix Fig. S13).

In addition to simulating the chain dynamics, the model converges to estimates of the stationary distribution in the memory data that reveal modes in the same locations (see Fig. 5A and SI Appendix Fig. S13). We ran permutation tests that show a high fit between the simulation results and the memory results across all iterations of the chains, and for all objects (SI Appendix Fig. S13). In all cases, the fit was significantly higher than chance, as measured by fitting the simulations to 10000 random rotations of the memory KDEs (r = 0.4 - 0.8, $p < 0.001$).

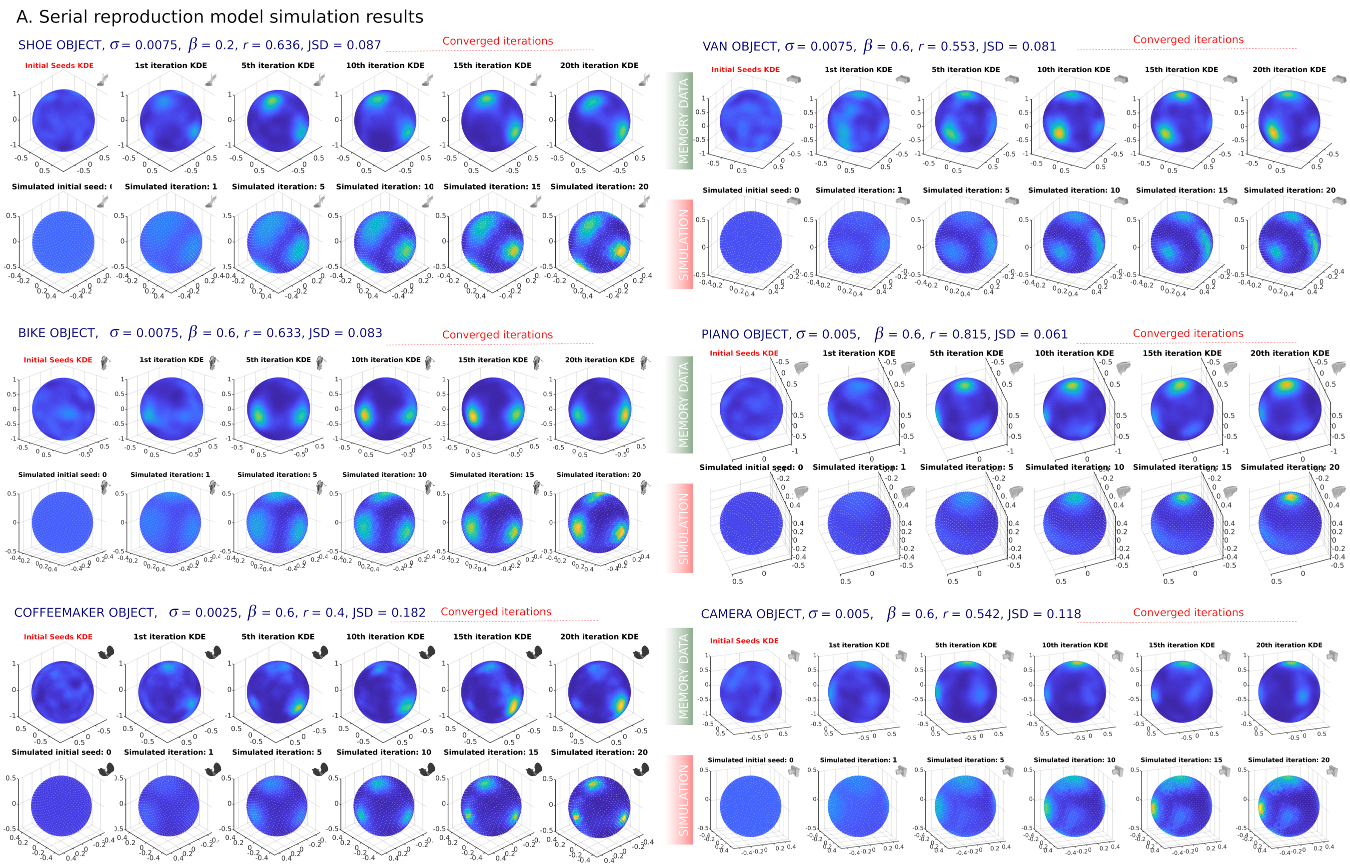


Figure 5 IB Simulation results. A. The first row in each object subpanel shows KDEs of the serial reproduction memory data, the second row immediately below it shows the simulated KDEs produced by the optimized IB serial reproduction model. Examples are shown for six of the eight objects (See SI Appendix for all examples). In all cases the model produced good approximations of the serial reproduction dynamics. Permutation testing shows that the model fits produce predictions of the actual chain KDEs across all iterations that is statistically significantly better than chance in all cases (r = 0.4-0.8, $p < 0.001$). Note that the figure shows the simulation results on a finer spherical lattice grid (containing 1600 lattice points instead of 324, as in Fig. 4). Although we ran the IB simulation with a lower resolution ($J = 324$ lattice points), we used linear interpolation to estimate the results of the chains at a finer resolution and when comparing them to the actual memory KDEs.

Discussion

Summary of the results

We adapted a serial reproduction design [12] to probe shared visuospatial memory priors for the 3D rotation of objects. We overcome serious limitations in past work examining perceptual representations of 3D views by sampling uniformly and densely over the group of all possible 3D rotations (the $SO(3)$ group). The results are not always consistent with canonical effects, but reveal clear structure, with positional modes corresponding to the primary faces of the objects including the top and bottom faces, and modes within the positional modes that consistently reveal a bias towards both upright and upside-down orientations of the objects (for the side views and front and back views). View-dependent effects in 3D perception have traditionally been explained in terms of a constructive bottom-up process that depends on subjective estimates of statistical frequency or typicality [14, 18, 19], but our results appear inconsistent with such an explanation, since many mode views in the memory priors are not necessarily those that are most commonly encountered in the visual environment. There is growing evidence that language and category-level knowledge can influence perception and memory [20, 21, 22], and our results show that memory for 3D rotations of objects are systematically biased towards fixed points that are predicted by optimal compression of language-based representations of object views and tend to correlate with a reduction in entropy of naming distributions (increased nameability, see SI Appendix Fig. S10). These findings suggest a normative explanation for the biases and the structure of the priors, namely that memory contracts towards simplified representations that reflect communicative efficiency by optimizing a trade-off between accuracy and efficiency in information transmission. From an information theoretic perspective, this trade-off is specified by the information bottleneck principle, which has been used to predict the structure of perceptual categories [26, 27, 28]. The structure of the priors align with the geometry of language-based internal representations of the object views and the IB estimates of the optimal stochastic naming policy that best approximates "meanings" that a speaker wishes to communicate to an idealized Bayesian listener. Our model simulates biases in reconstruction towards low entropy regions in the semantic space (See SI Appendix Fig. S12C-D).

Limitations and future directions

Our primary contribution is a detailed empirical examination of visual memory priors for 3D views. We overcome many limitations of past work including non-uniform sampling of 3D view rotations in $SO(3)$ [14]. Running serial reproduction chains of view reconstructions also enabled us to quantify convergent estimates of the fixed points in memory that produce perceptual biases. These experiments reveal consistent patterns that challenge explanations based on statistical learning and are more in line with categorical perception and communicative efficiency. However, due to the dense sampling of the domain space required, one limitation is that we produced estimates for only 8 objects. While we chose examples that were ecologically valid with complex and variagated geometries, the limited number of examples necessarily limits the generalizability of our findings. Similarly, our model results are limited by the lexicon of words that were provided by participants in the word naming task. We elicited visuospatial words because of the nature of the task, but other word choices are possible (such as part-based words, or even other spatial words). Using a forced choice experimental design also limited the number of words participants could choose from. Future work should overcome this limitation by eliciting free-word associations for the different view images. Perhaps that with this data the simulation could be repeated to make predictions on the full 3-sphere (view positions and orientations).

An obvious limitation of our model is that it makes predictions of the memory bias and chain KDEs on the 2-sphere (view positions), but not the full 3-sphere (view positions and orientations). We aim to overcome this limitations in future work with more labelling data. In addition, while the model produces good predictions of the memory biases, the tradeoff parameter values (values for β) do not have a clear interpretation (as they do in other work [26, 28]). IB simply provides an ideal framework for estimating internal category boundaries non-parametrically (via the encoder conditional distribution $q_o(w|m)$) by estimating clusters of meanings in the semantic view spaces we estimated for the objects.

In prior work, we explained perceptual biases and 2D visuospatial priors in terms of efficient coding of natural images [12, 33]. In this work, we formed testable theoretical predictions of the variable precision with which different images regions are encoded, in addition to the memory biases. We explored the relationship between discrimination sensitivity and bias, and made comparisons between the priors and measures of explicit attention via eye-tracking data. However, due in large part to the complexity of $SO(3)$ as a domain of study, we could not make detailed estimates of variations in discrimination accuracy or selective attention which could be mediators of the memory biases. Although attention and encoding precision could be implicated in the biases we observed here, the alignment of the memory priors to semantic categories and variation in attention allocation are not necessarily mutually exclusive processes. Finally, further work could explore a causal manipulation [25] to determine if semantic cues amplify bias by adding precision to category priors.

Author contributions

T.A.L, N.J., and T.L.G. designed research. T.A.L. performed research. T.A.L. and N.J., contributed new reagents/analytic tools; T.A.L. analyzed the data; T.A.L. developed the theory and statistical modeling; T.A.L., N.J., and T.L.G. wrote the paper.

Competing Interests

The authors declare no competing interests.

Participants

All participants were recruited on Amazon Mechanical Turk (Mturk). For the serial reproduction experiments, we used the Dallinger platform for laboratory automation for the behavioral and social sciences [34]. All the experiments were approved by Princeton University's Institutional Review Board (IRB) for Human Subjects under protocol #10859 (Computational Cognitive Science), and all participants provided informed consent. For the serial reproduction experiments, participants were paid a base rate of $0.5 dollars, but could receive a bonus of up to $3.5 for a total of $4.0 to complete the HIT, which contained 105 experimental trials. The bonus was contingent on accuracy in the task (see below). The time required to complete the task was about 30 minutes. Participants could take part only once per experiment, but could take part in more than one experiment. For the labelling nAFC experiment, participants received $0.75 to complete 72 trials. Table S15 presents the exact number of participants in each experiment. The overall number of participants in all experiments was 1944 (243 participants for each objects). We only recruited participants on Mturk who had 95% or more of their completed HITs approved.

Stimuli

The objects we used for the serial reproduction experiments and canonical views (typicality) experiments were grayscale .3DS 3D objects of a shoe, teapot, van, clock, camera, coffee maker, motorcycle, and grand piano. Table S15 for the list of object file names for each of the experiments. For the labelling experiments, stimuli were 1944 images of each object presented in a random orientation. All stimuli for the experiments are available in our open science repository.

Procedure

Transmission chain memory experiments and canonical view experiments were programmed using the Dallinger platform for laboratory automation for the behavioral and social sciences [34]. We provide reproducible code for the Dallinger experiments in the open science repository associated with this paper. The word list experiment and labelling nAFC experiments were programmed using the Amazon Mechanical Turk API.

Transmission chain memory experiments (Fig. 1A)

Participants were shown a rotated object for 4000 ms. The initial object rotations were sampled uniformly in $SO(3)$ using the sampling method described in [30]. After a blank delay lasting 1000 ms, participants were shown the object in a random rotation, and asked to rotate it back to the exact rotation in the initial stimulus phase as accurately as possible (see Fig. 1A). Participants could take as much time as they wanted to match the stimulus rotation, and could move on to the next trial in the experiment once ready by clicking on a "next" button. Once they clicked on the "next button," they were given feedback about their accuracy, along with the monetary bonus for that trial. If their response was within an allowable margin of error, it was then routed by the Dallinger platform to another participant on Mturk who performed the same task. A total of twenty iterations of this "telephone game" procedure were completed for each chain. We terminated each experiment after approximately 24 hours. We ran a total of 500 chains for each object, see Table S15). A typical experiment included 105 trials, and the average time needed to complete the task was about 30 minutes. Table S15 presents the number of participants in each experiment.

Statistical Analysis

The Jensen-Shannon Divergence (JSD)

We used the Jensen-Shannon Divergence (JSD) for statistical comparisons, convergence analyses, and to evaluate the IB communication model simulation fit to the positional memory Kernel Density Estimates (KDEs). The JSD of two distributions P and Q is defined by the following:

$$JSD(P,Q) = \frac{1}{2}KL\left(P \mid\mid M\right) + \frac{1}{2}KL\left(Q \mid\mid M\right)$$

where $M = \frac{1}{2}(P+Q)$ and $KL\left(P_1 \| P_2\right)$ is the Kullback-Liebler (KL) divergence:

$$KL\left(P_1 \| P_2\right) = \int_s P_1(s) \log_2 \frac{P_1(s)}{P_2(s)} ds$$

The JSD is symmetric, and bounded between 0 and 1. It is equal to 0 when $P_1 = P_2$.

Transmission Chain Convergence Analysis

JSD distance between KDEs at each iteration and the initial seed KDE

For serial reproduction experiments, it is critical to evaluate whether the chains converge to fixed points in the iterated process [12]. One way to do this is to compare the Kernel Density Estimates (KDEs) of the data at each iteration of the chains to the KDE of the data in the initial seed (one can also compare the KDEs at each iteration to the KDE of the data in the final iteration). We can say that the data have converged if we can establish that the JSD between a KDE of the data in iteration t and the KDE of the initial seeds is not significantly different from the JSD between a KDE of the data at iteration $t + 1$ and the KDE of the initial seeds. This indicates that the change in responses between iteration t and iteration $t + 1$ has reached a plateau. In order to evaluate whether the differences in JSD are significant, we used bootstrapping to generate 1000 KDEs of the data at each of the 20 iterations of the chains by resampling all of the data in each iteration with replacement 1000 times, and generating a KDE for each bootstrapped sample using the same KDE technique described in the Methods section. We then obtained the JSDs between each of the 1000 KDEs for each iteration t and its corresponding KDE at iteration 0 (initial seed). Thus, for each iteration t we obtained 1000 JSD values comparing the 1000 bootstrapped sample KDEs at iteration t to the same bootstrapped sample KDEs of the 0th initial seed data. We used the standard deviation of these JSD values to measure the variability in JSD values for each comparison, and we used paired t-tests for the comparisons (there were 20 total comparisons). We used the Bonferroni correction to correct for multiple comparisons. We observed that the data converged before the 11th iteration of the serial reproduction procedure for all 3D objects (when changes ceased to be significant) except one which converged by the 17th iteration (clock object). For the remaining objects, changes in JSD were only significant up to the comparison between JSDs between the seed data and the 10th iteration ($p < 0.0001$). For all comparisons (model fit, and comparisons between the data across experiments), we used KDEs of the data aggregated across all the converged iterations (e.g. we fit a KDE to the data for all iterations between iteration 11 and 20 if we observed convergence after the 10 iteration). We show the results of this convergence analysis in Fig. S7.

JSD distance between subsequent iteration pairs

Another method for evaluating convergence of the chains is to evaluate whether there is an iteration at which the data distribution at iteration $t + 1$ ceases to change significantly when compared to the data distribution at iteration t. In order to evaluate this, we used the same procedure described above, except that instead of comparing the KDEs at each iteration to the initial iteration KDE, we compare each of the 19 KDE pairs (e.g. KDE at iteration 1 versus iteration 2, then KDE at iteration 2 versus iteration 3, etc). We found that for all objects the data distributions cease to change significantly after iteration 10 except for the clock object (which reached convergence at iteration 17).

Kernel Density Estimation

We chose to produce KDEs for the positional data (azimuth ϕ and elevation θ) and angular orientation data (angle α) separately. We made this choice to produce visualizations of the results that are as intuitive as possible. Although we could have produced KDEs of the results on the full 3-sphere, they would have been difficult to visualize in a way that is intuitive or easy to interpret.

Positional Kernel Density Estimation

We started by generating a fine regular Fibonacci lattice over the unit 2-sphere [35]. For each point on the lattice nearest to the position of a response in a given chain and for a given iteration (defined by its azimuth ϕ and elevation θ), we added a 2D Gaussian Kernel with standard deviation $I_3 \cdot \sigma = 0.0125$ centered on that lattice point. We repeated this process for all chain responses, and for each iteration, and summed over all the Gaussian Kernels (chain responses) to produce the final KDE for each iteration (see SI Fig. S1, SI Fig. S2, and SI Fig. S3 for all positional KDEs for three objects, along with quiver plots of the raw data in full $SO(3)$ for all chain iterations and initial seeds). SI Fig. S13 shows all positional KDEs for all objects.

Angular Kernel Density Estimation

We started by aggregating the data across all convergent iterations (see convergence analysis section). We then used a spherical k-means clustering algorithm (with $K = 6$) to group the responses in the positional modes we observed in the positional KDEs. We chose $K = 6$ because we observed 6 distinct modes for all the objects, although they varied in density for different objects. Following this, we computed the angle of each response vector in each cluster relative to one of the standard basis vectors.

We used a different reference basis vector depending on the kth centroid. For centroids corresponding to the side views of an object, we computed the angular differences between response vectors and a basis vector pointing from the center of mass of the object towards its front (see schematic illustrations and results in Fig. 1 of the main text, and SI Appendix Fig. S10-12). This measured the distribution of angular responses (local camera orientation angles α) relative to an orientation of the object with its front face pointing vertically upwards. This revealed consistent bimodal responses in the data indicating that memory is not only biased towards left and right frontal faces of the objects —inside these modes they are consistently biased towards both an *upright* orientation as well as an *upside-down* orientation in these right-side and left-side modes. Note that the choice of using a basis reference vector pointing from the center of mass of the object towards its front is arbitrary, and we could have computed the angular difference between the response vectors and a different reference vector, such as a vector centered at the object's center of mass and pointing vertically upward. Such a choice would have simply resulted in a circular permutation of the plotted line (1D KDE) along the x-axis, but it would not have changed the shape of the line itself. We chose the reference basis vector that allowed the two distinct modes (present for all the objects) to be clearly visible (without one being truncated at the edges of the plots).

Once we computed all the angular differences, for each of the K clusters we computed KDEs of the data using a fixed kernel width $\sigma = 0.5$ for each datum (angle), and summed across all the kernels. We then normalized the data to obtain the final KDE for each of the K clusters. Fig. 2D shows two examples, and Fig. S10-12 show KDEs for all the objects, and for all $K = 6$ clusters. The errorbands for each of the KDEs were estimated by computing 100 bootstrapped KDEs from all the angular difference data by resampling all the angular difference data in each cluster 100 times with replacement, and repeating the KDE procedure for each sample and taking the standard deviation of the distribution of KDE values for each angle between $-\pi$ and π. Note that the amount of data varied between clusters, yielding more reliable estimates for clusters containing more data (the most dense modes in the response data).

Bootstrapping Kernel Density Estimates (KDEs)

For the convergence analyses, we computed 1000 bootstrapped KDEs using the following procedure: We resampled all the serial reproduction data for all chains in a given iteration with replacement, and computed a Gaussian kernel centered at the point with a diagonal covariance matrix (using the same procedure for generating the positional KDEs described in the last section). The final KDE was calculated by summing all of the Gaussian kernels and normalizing. We repeated this process for each set of bootstrapped data.

Semantic view space estimation

In order to estimate the semantic view spaces for each of the objects, we used the naming distributions we obtained for each of the 1944 views of the object sampled uniformly in $SO(3)$. For each of the 1944 rotation views $x \in \mathcal{X}$, we averaged the responses from 9 unique participants on AMT who completed the 22-alternative forced choice naming task, in which they selected a view word from a list of 22 view words that they judged to be the best word to describe the object view. We then normalized the naming distribution (averaged over the responses from the 9 unique participants) for each view x. Next, we created a regular lattice of 324 camera positions on the 2-sphere (324 grid points $g \in \mathcal{G}$). For each grid point g, we computed the weighted sum of nearby naming distributions, where the weight was proportional to the density under a 2D Gaussian smoothing kernel centered on that grid point g. In other words, the weight of a nearby naming distribution $x \in \mathcal{X}$ was equal to $G(x, g, \Sigma)$, the Gaussian probability density (weight) evaluated at a view x with mean g and diagonal covariance matrix Σ. We used a fixed kernel width $I_3 \cdot \Sigma = 0.2$ for all views in $\mathcal{G}$, for all objects. We then normalized the weighted sum of nearby naming distributions for each of the grid points $g \in \mathcal{G}$. Finally, we computed all pairwise JSDs between each pair of the resulting (average) naming distributions for all the points on the grid. This gave us a measure of the pairwise semantic similarity between views on the grid. As described above, we then computed each $m(v)$ which was defined by: $G(x, v, \Sigma_p)$ (e.g. the Gaussian probability density with mean v and diagonal covariance matrix Σ_p in the semantic space). Each view space is 22-dimensional (since the view word lexicon in the naming task consisted of 22 view words), but for visualization purposes we show the 3D projection of these 22D view spaces in SI Fig. S10. In practice, and for each object, we used the distance matrix containing all pairwise JSDs between the 22D naming distributions to define the view space support for all meanings $m(v)$ for the simulations. Fig. SI Fig. S10 reveals that the geometry of the 3D projection of the semantic space varied significantly from object to object. The matrix containing each of the J meanings $m(v)$ became the input to the IB communication model for each object. Note that the distribution of meanings varied greatly from object to object and depended on the unique geometry of the view space for that object.

Model

The Information Bottleneck (IB) Model (Fig. 4, and SI Fig. S11)

Notation

We used the same notation used by [26]. We denote random variables using capital letters (e.g. M to describe meanings, and V to describe views), and lower case letters represent samples, such as: $m \in \mathcal{M}$ or $v \in \mathcal{V}$. We refer to the domains (or support) for each using calligraphic letters (e.g. $\mathcal{M}$ and $\mathcal{V}$ for meanings and views, respectively). As in [26], each element m of the finite set of distributions in $\mathcal{M}$ (e.g. each $m \in \mathcal{M}$) is a function that takes a point v in a view space as an argument (see section describing the procedure for estimating view spaces for each object). The function $m(v)$ is defined by: $G(x, v, \Sigma_p)$, which is the Gaussian probability density with mean v and diagonal covariance matrix Σ_p. Hence, $m(v)$ specifies the probability of a view v according to m. Another way to understand $m(v)$ is to think about it in terms of conditional probabilities ($m(v) = p(v|m)$).

The Information Bottleneck (IB) Method

The IB objective function is given by the following:

$$\mathcal{F}_\beta \left[q(w|m)\right] = I_q(M;W) - \beta I_q(W;V) \tag{8}$$

where $I_q(M;W)$ is the complexity term, which measures the mutual information between meanings M and the quantization (compressed representation) W, and $I_q(W;V)$ is the accuracy term. The β parameter specifies the complexity / accuracy tradeoff. Higher values of β maximize the accuracy, while lower values of β lead to more compressed efficient representations (and lower accuracy). Given a value of β, the IB method [32] iteratively updates the following set of self-consistent equations until convergence (when Equation 5 is minimized):

$$q_\beta(w|m) = \frac{p(w)}{\mathcal{Z}(m,\beta)} \exp\Big(-\beta KL\Big[m(v)|\hat{m}(v)\Big]\Big) \tag{9}$$

$$q_\beta(w) = \sum_m q_\beta(w|m)p(m) \tag{10}$$

$$\hat{m}_w(v) = \sum_m m(v) q_\beta(m|w) \tag{11}$$

where $\mathcal{Z}(m,\beta)$ is the normalization factor. Because IB is a non-convex problem, a difficulty of the IB method is that it can converge to sub-optimal fixed points. Several approaches exist to avoid this problem [26]. Two common approaches are deterministic annealing and reverse deterministic annealing. In the latter approach, the IB curve is evaluated by starting the minimization of Equation 5 with a very high value of β and then decreasing it by small increments. For each decrease, the optimization is initialized with the solution from the converged result using the β value from the previous optimization. In practice, we found that initializing the minimization for each value of β with an identity matrix provided solutions that were equivalent to those obtained via reverse deterministic annealing. For each object, we searched for solutions using β values in the $[0 - 10]$ range, and found that the optima tended to be for solutions using β values that were less than one. SI Appendix Fig. S11A shows example $\hat{m}(v)$ solutions (meaning reconstructions) for an example object (shoe) in both the external view (on the 2-sphere lattice) and in the "internal" view space representation (the 3D MDS projection of the true view space used in the simulations). SI Appendix Fig. S11B shows the information plane for the same object. The y-axis shows the normalized accuracy, and the x-axis shows the normalized complexity (see main text). The curves in the information plane show the results of using two optimization algorithms: the standard Lagrangian minimization algorithm described above and used in past work [26] (red line), as well as an agglomerative form [36] that has been used in the past as a way of initializing the former minimization algorithm for a given lexicon size K (blue line). For our simulations, we always used a lexicon size that allowed a one-to-one correspondence between the full set of J views in V and the set of possible words in W (e.g. $|W| = |V|$). We therefore only used the Lagrangian minimization to solve for the encoder $q_o(w|m)$. We show both solutions in SI Appendix Fig. S11B to highlight the superiority of using the self-consistent equations over the agglomerative approach. For each object we fit the full serial reproduction model results to the positional memory KDEs. The only parameters of the model were the IB β parameter, and the internal perceptual noise parameter σ_p. In the next section, we describe the serial reproduction model.

Serial Reproduction Model (Fig. 4)

We model an initial view percept v_t^P centered on a Gaussian $q_0(v_t^P)$ with a mean stimulus view position v_t^S and fixed isotropic noise σ_s (e.g. $q_0(v_t^P) = \mathcal{N}(v_t^S, \sigma_s)$). This same view v_t^P also has a semantic representation in the form of the reconstruction $\hat{m}(v)$ produced by the naming model's reconstruction of $m(v)$ where $v = v_t^P$. The participant combines both the initial (Gaussian) distribution $q_0(v_t^P)$ over views centered on v_t^P and the language-based reconstruction $\hat{m}(v_t^P)$ by an element-wise product of the two distributions. We then define a memory response v_t^R in a chain as the mode of the resulting distribution $q_t(v_t^P)$. Each step in the chain combines $q_t(v_t^P)$ from the previous step with $\hat{m}(v_{t+1}^P)$, where the new stimulus view $v_{t+1}^P = v_t^R$, which is the response from the previous iteration (see Fig. F4D in the main text). This results in a new distribution $q_{t+1}(v_{t+1}^P)$. The response v_{t+1}^R again corresponds to the mode of $q_{t+1}(v_{t+1}^P)$, which becomes the new stimulus v_{t+2}^P in the chain. Each step in a single chain is modeled as follows:

$$q_{t+1}(v_{t+1}^P) = q_t(v_t^P) \circ \hat{m}(v_{t+1}^P) \tag{12}$$

where the stimulus view v_{t+2}^P in the next step is equal to the following:

$$v_{t+2}^P = v_{t+1}^R = \arg\max_v \big(q_{t+1}(v_{t+1}^P)\big) \tag{13}$$

We simulated full KDEs by repeating this procedure using all the J grid points v on the 2-sphere as the initial chain seeds, and marginalizing over all the $q_t(v)$ distributions at each iteration t. SI Appendix Fig. S12A shows a blow-up of panel D in Fig. 5 of the main text. This panel illustrates the serial reproduction model for a single chain. We obtained simulated KDEs by repeating this process for all $J = 324$ lattice points on the 2-sphere.

Canonical views ("typicality" estimates)

We make explicit comparisons between the results of our memory experiments and estimates of the canonical views (e.g. "typical") of the same objects. To measure the canonical views for our objects, we used the same user interface that participants used for the memory reconstruction in the serial reproduction experiments, but instead of matching a response to a stimulus view from a random probe view (Fig. 2), 688 participants were instructed to orient the object to the "best or most typical" view for that object as in [14]. The results show some similarities to the results from the memory experiments, but with notable differences. Overall, participants did not choose views from the top or bottom of the objects, although this was a consistent finding in the memory data. In addition, participants did not tend to choose upside-down views of the objects, although these were clearly present for all objects in the serial reproduction memory data. Instead, participants oriented the objects to modes over the side views, and in some cases, the front views of the objects. SI Appendix Fig. S14 shows the raw results and KDEs from the canonical views experiment.

References

[1] Zoya Bylinskii, Phillip Isola, Constance Bainbridge, Antonio Torralba, and Aude Oliva. Intrinsic and extrinsic effects on image memorability. *Vision research*, 116:165–178, 2015.

[2] Steven J Luck and Edward K Vogel. The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657):279–281, 1997.

[3] Weiwei Zhang and Steven J Luck. Discrete fixed-resolution representations in visual working memory. *Nature*, 453(7192):233–235, 2008.

[4] David C Knill and Whitman Richards. *Perception as Bayesian inference*. Cambridge University Press, 1996.

[5] Yair Weiss, Eero P Simoncelli, and Edward H Adelson. Motion illusions as optimal percepts. *Nature Neuroscience*, 5(6):598, 2002.

[6] Alan A Stocker and Eero P Simoncelli. Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, 9(4):578, 2006.

[7] Daryl Fougnie, Jordan W Suchow, and George A Alvarez. Variability in the quality of visual working memory. *Nature Communications*, 3:1229, 2012.

[8] Paul M Bays, Raquel FG Catalao, and Masud Husain. The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, 9(10):7–7, 2009.

[9] Xue-Xin Wei and Alan A Stocker. A bayesian observer model constrained by efficient coding can explain'anti-bayesian'percepts. *Nature neuroscience*, 18(10):1509–1517, 2015.

[10] Nora S Newcombe and Janellen Huttenlocher. *Making space: The development of spatial representation and reasoning*. MIT Press, 2003.

[11] Ronald Van Den Berg, Hongsup Shin, Wen-Chuang Chou, Ryan George, and Wei Ji Ma. Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, 109(22):8780–8785, 2012.

[12] Thomas A Langlois, Nori Jacoby, Jordan W Suchow, and Thomas L Griffiths. Serial reproduction reveals the geometry of visuospatial representations. *Proceedings of the National Academy of Sciences*, 118(13):e2012938118, 2021.

[13] Nori Jacoby and Josh H McDermott. Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction. *Current Biology*, 27(3):359–370, 2017.

[14] Stephen Palmer. Canonical perspective and the perception of objects. *Attention and performance*, pages 135–151, 1981.

[15] Michael L Kalish, Thomas L Griffiths, Stephan Lewandowsky, et al. Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin and Review*, 14(2):288, 2007.

[16] Stefan Uddenberg and Brian J Scholl. Teleface: Serial reproduction of faces reveals a whiteward bias in race memory. *Journal of Experimental Psychology: General*, 147(10):1466, 2018.

[17] Ahna R Girshick, Michael S Landy, and Eero P Simoncelli. Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature neuroscience*, 14(7):926–932, 2011.

[18] Jonathan Sammartino and Stephen E Palmer. Aesthetic issues in spatial composition: Effects of vertical position and perspective on framing single objects. *Journal of Experimental Psychology: Human Perception and Performance*, 38(4):865, 2012.

[19] Stephen E Palmer, Karen B Schloss, and Jonathan Sammartino. Visual aesthetics and human preference. *Annual review of psychology*, 64:77–107, 2013.

[20] Gary Lupyan, Rasha Abdel Rahman, Lera Boroditsky, and Andy Clark. Effects of language on visual perception. *Trends in cognitive sciences*, 24(11):930–944, 2020.

[21] Naomi H Feldman, Thomas L Griffiths, and James L Morgan. The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. *Psychological review*, 116(4):752, 2009.

[22] Erik A Wing, Ford Burles, Jennifer D Ryan, and Asaf Gilboa. The structure of prior knowledge enhances memory in experts by reducing interference. *Proceedings of the National Academy of Sciences*, 119(26):e2204172119, 2022.

[23] Jonathan Winawer, Nathan Witthoft, Michael C Frank, Lisa Wu, Alex R Wade, and Lera Boroditsky. Russian blues reveal effects of language on color discrimination. *Proceedings of the national academy of sciences*, 104(19):7780–7785, 2007.

[24] Noga Zaslavsky, Charles Kemp, Naftali Tishby, and Terry Regier. Communicative need in colour naming. *Cognitive neuropsychology*, 37(5-6):312–324, 2020.

[25] Lewis Forder and Gary Lupyan. Hearing words changes color perception: Facilitation of color discrimination by verbal and visual cues. *Journal of Experimental Psychology: General*, 148(7):1105, 2019.

[26] Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31):7937–7942, 2018.

[27] Noga Zaslavsky, Charles Kemp, Naftali Tishby, and Terry Regier. Color naming reflects both perceptual structure and communicative need. *Topics in cognitive science*, 11(1):207–219, 2019.

[28] Noga Zaslavsky, Terry Regier, Naftali Tishby, and Charles Kemp. Semantic categories of artifacts and animals reflect efficient coding. *arXiv preprint arXiv:1905.04562*, 2019.

[29] Elad Mezuman and Yair Weiss. Learning about canonical views from internet image collections. *Advances in neural information processing systems*, 25, 2012.

[30] Xavier Perez-Sala, Laura Igual, Sergio Escalera, and Cecilio Angulo. Uniform sampling of rotations for discrete and continuous learning of 2d shape models. In *Robotic vision: Technologies for machine learning and vision applications*, pages 23–42. IGI Global, 2013.

[31] Julie A Charlton, Wiktor F Młynarski, Yoon H Bai, Ann M Hermundstad, and Robbe LT Goris. Environmental dynamics shape perceptual decision bias. *PLOS Computational Biology*, 19(6):e1011104, 2023.

[32] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

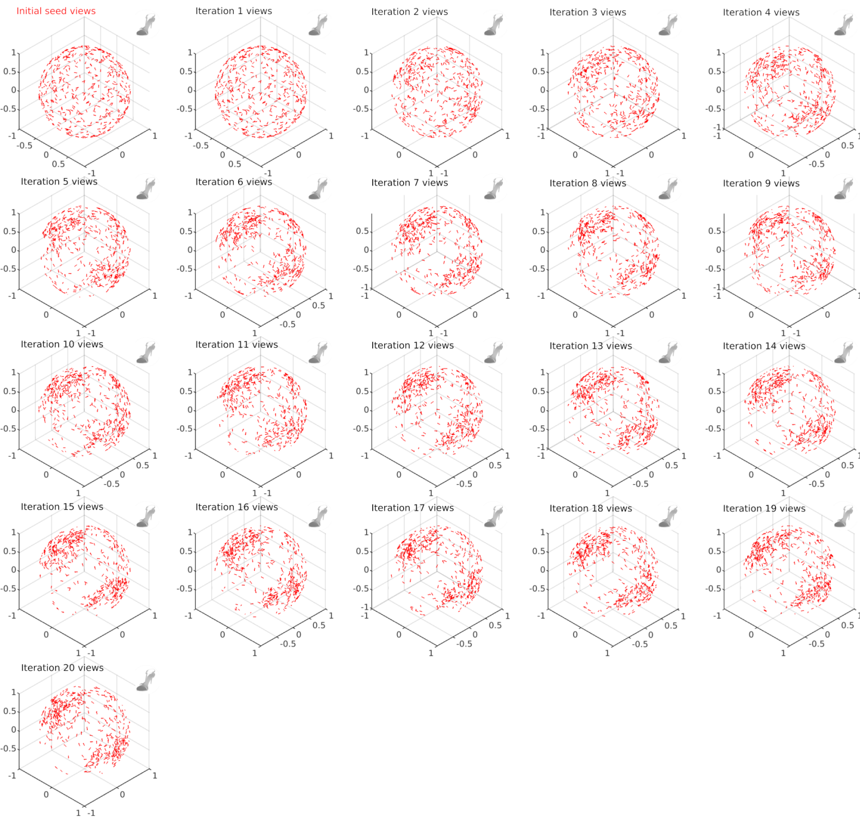
[33] Thomas Langlois, Nori Jacoby, Jordan W Suchow, and Thomas L Griffiths. Uncovering visual priors in spatial memory using serial reproduction. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, 2017.

[34] Morgan T.J.H. Lall V.H. Hamrick J.B. Meylan S.C. Mitchell A.P. Wilkes M. Glick D.I. De La Guardia C. Snyder J.M. Kleinfeldt S.E. Griffiths T.L. Suchow, J.W. Fully automated behavioral experiments on cultural transmission through crowdsourcing. *Collective Intelligence 2019*, 2019.

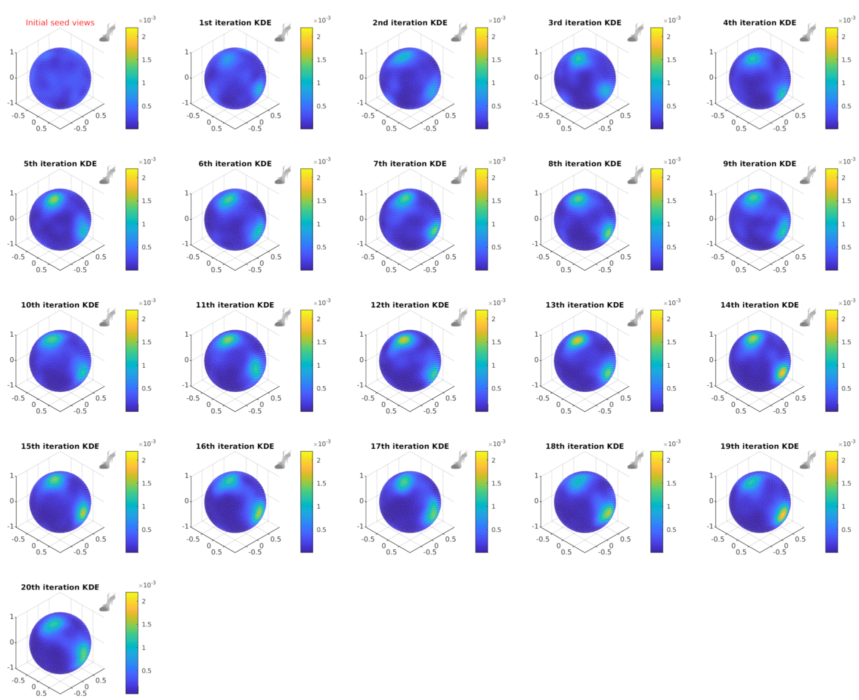
[35] Álvaro González. Measurement of areas on a sphere using fibonacci and latitude–longitude lattices. *Mathematical Geosciences*, 42:49–64, 2010.

[36] Noam Slonim and Naftali Tishby. Agglomerative information bottleneck. *Advances in neural information processing systems*, 12, 1999.

A. Random initial seed and serial reproduction results for all iterations

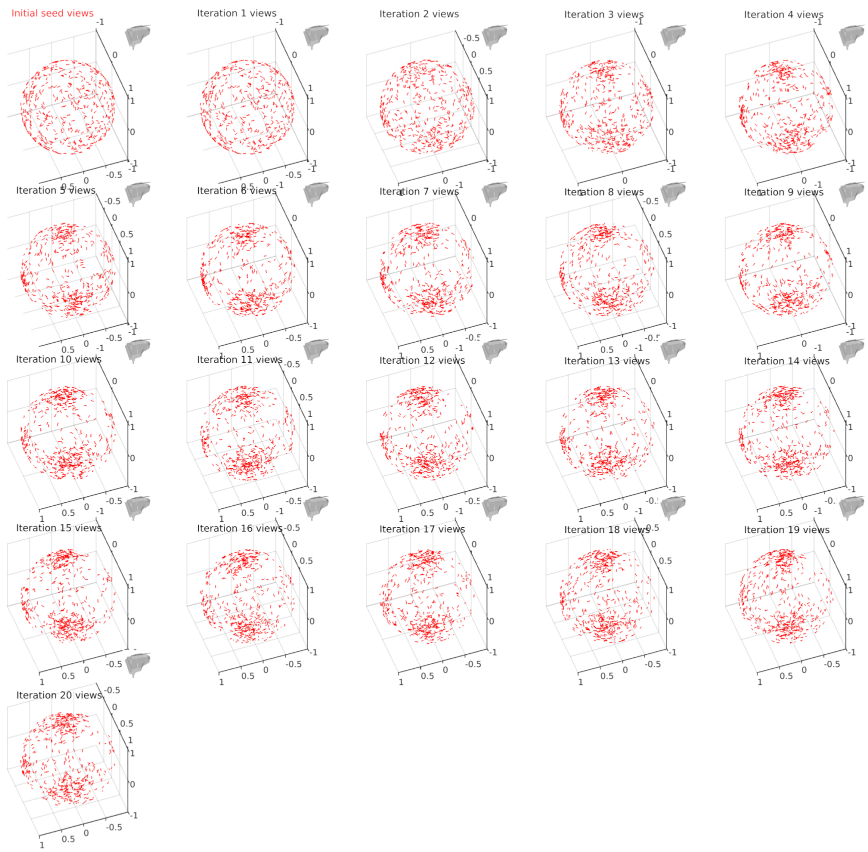


B. Random initial seed positional KDE and serial reproduction positional KDEs for all iterations

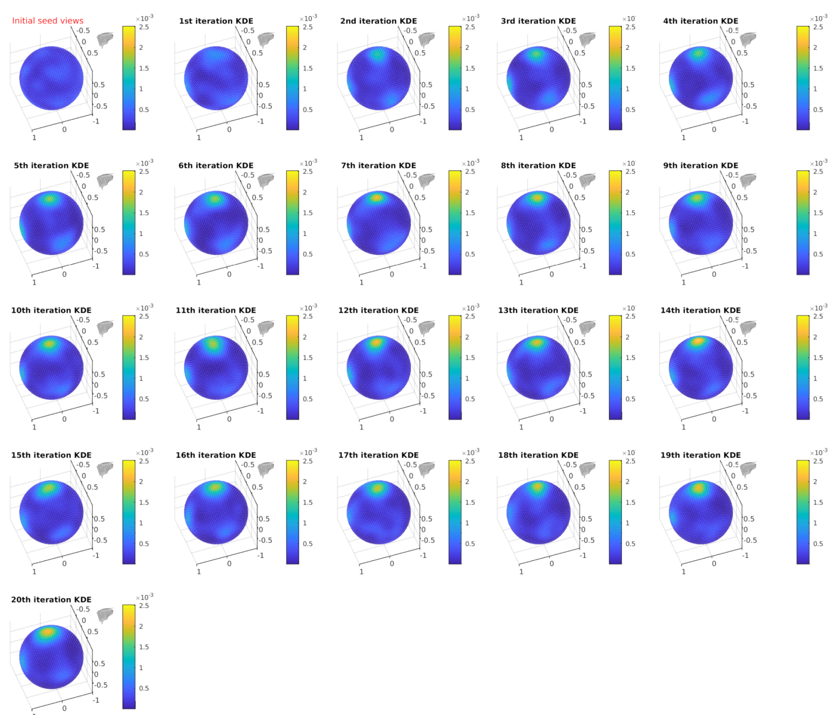


Extended Data Figure 1 Example full serial reproduction chain results and KDEs for the shoe object. A. shows quiver plots of the raw data, and B. shows the positional KDEs. Thumbnails in the upper right of each of the subplots indicates the orientation of the object for reference.

A. Random initial seed and serial reproduction results for all iterations

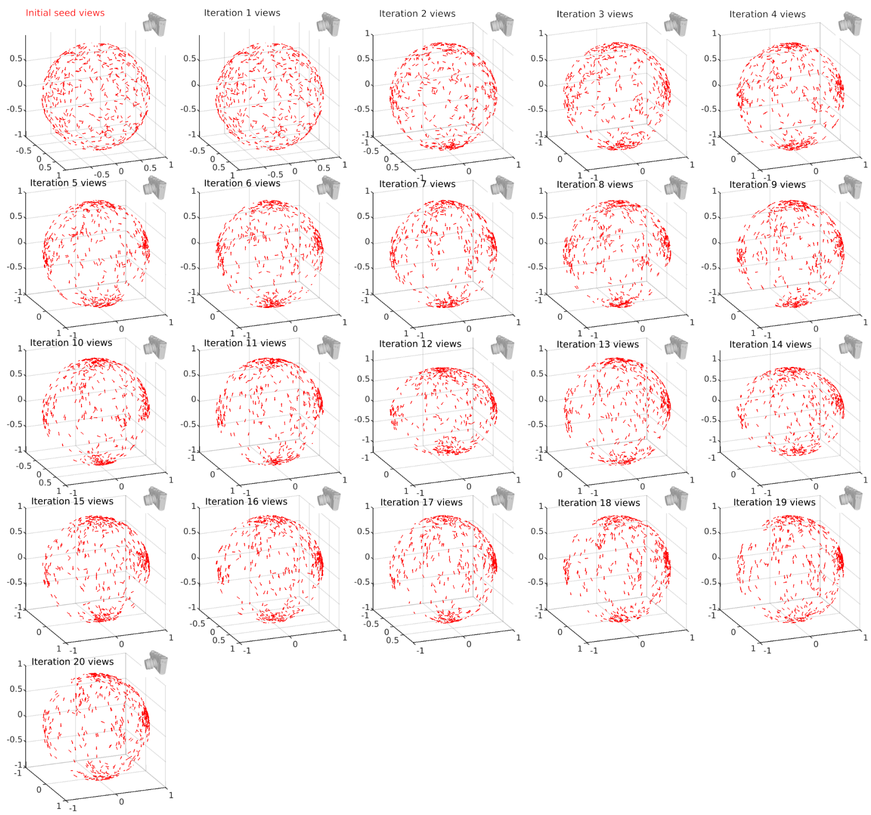


B. Random initial seed positional KDE and serial reproduction positional KDEs for all iterations

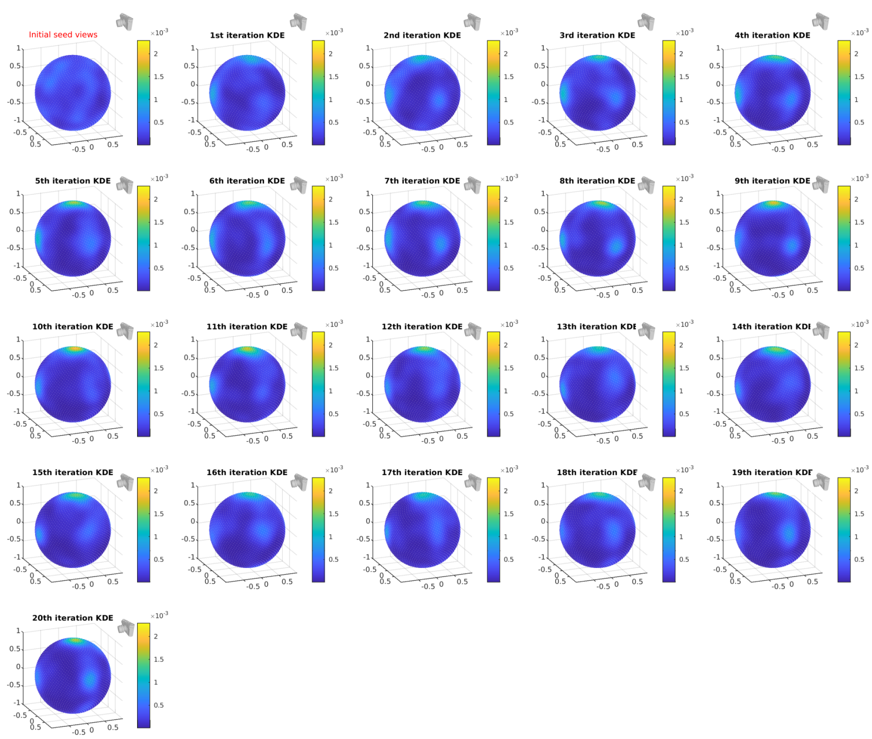


Extended Data Figure 2 Example full serial reproduction chain results and KDEs for the piano object. A. shows quiver plots of the raw data, and B. shows the positional KDEs. Thumbnails in the upper right of each of the subplots indicates the orientation of the object for reference..

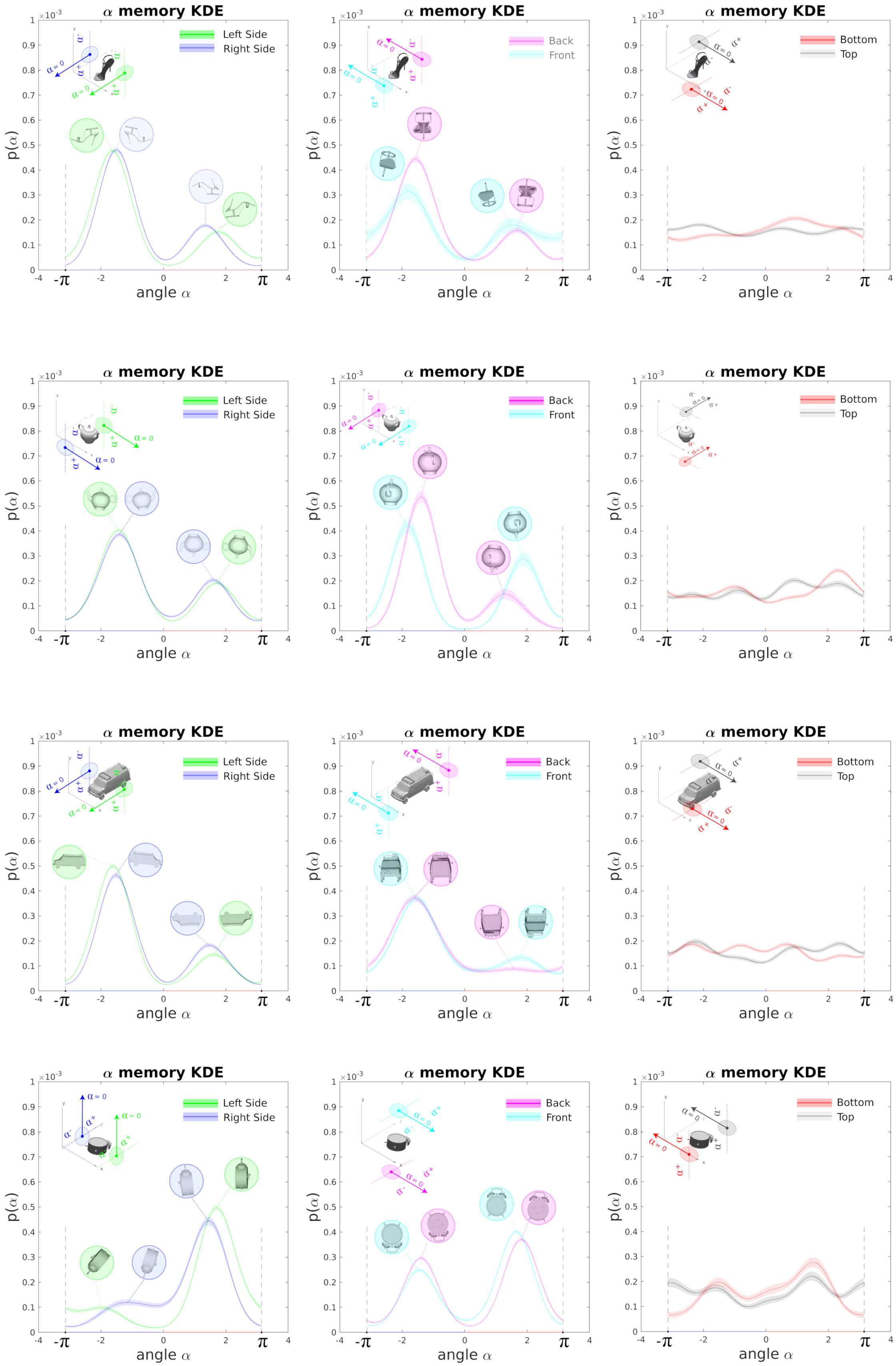
A. Random initial seed and serial reproduction results for all iterations



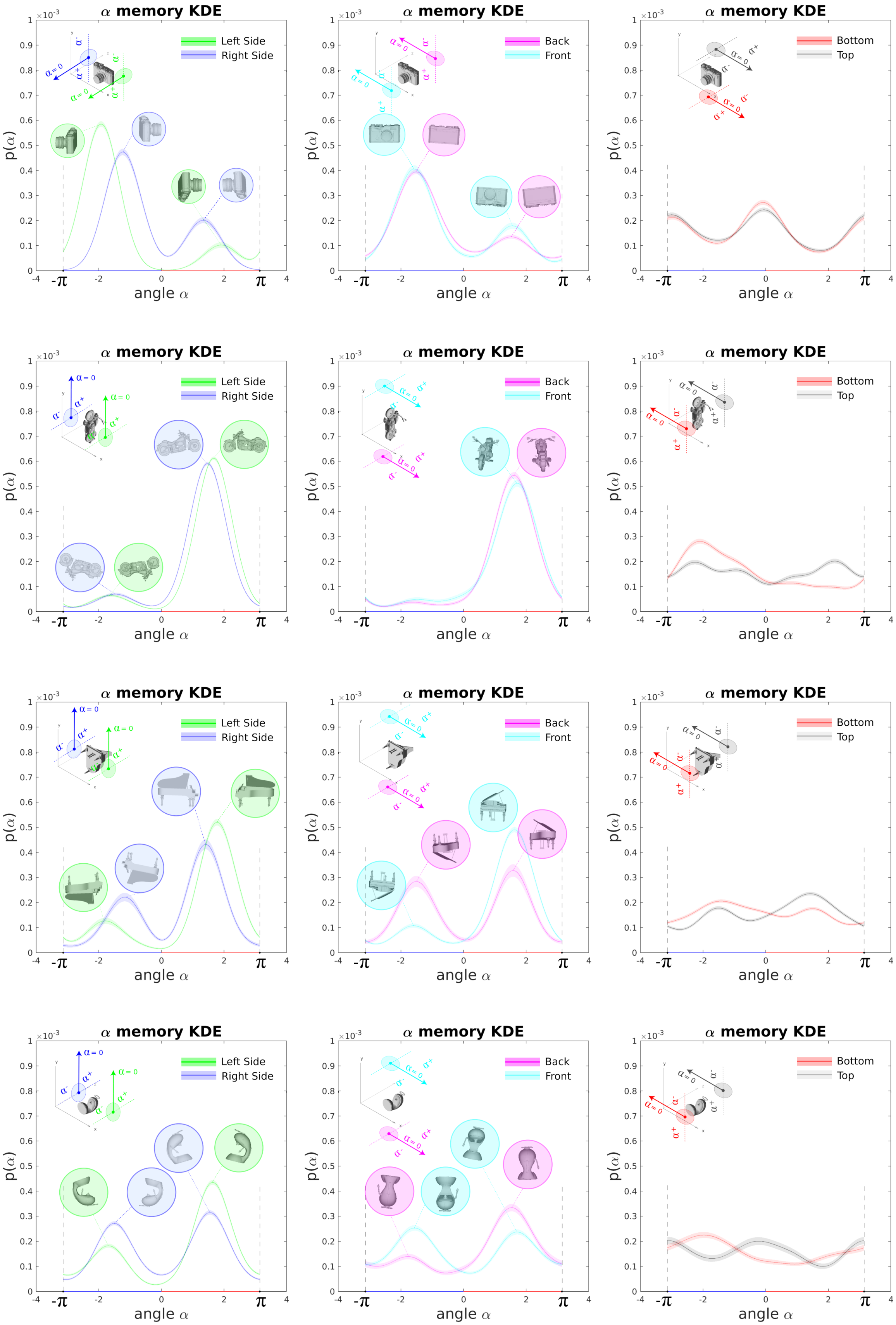
B. Random initial seed positional KDE and serial reproduction positional KDEs for all iterations



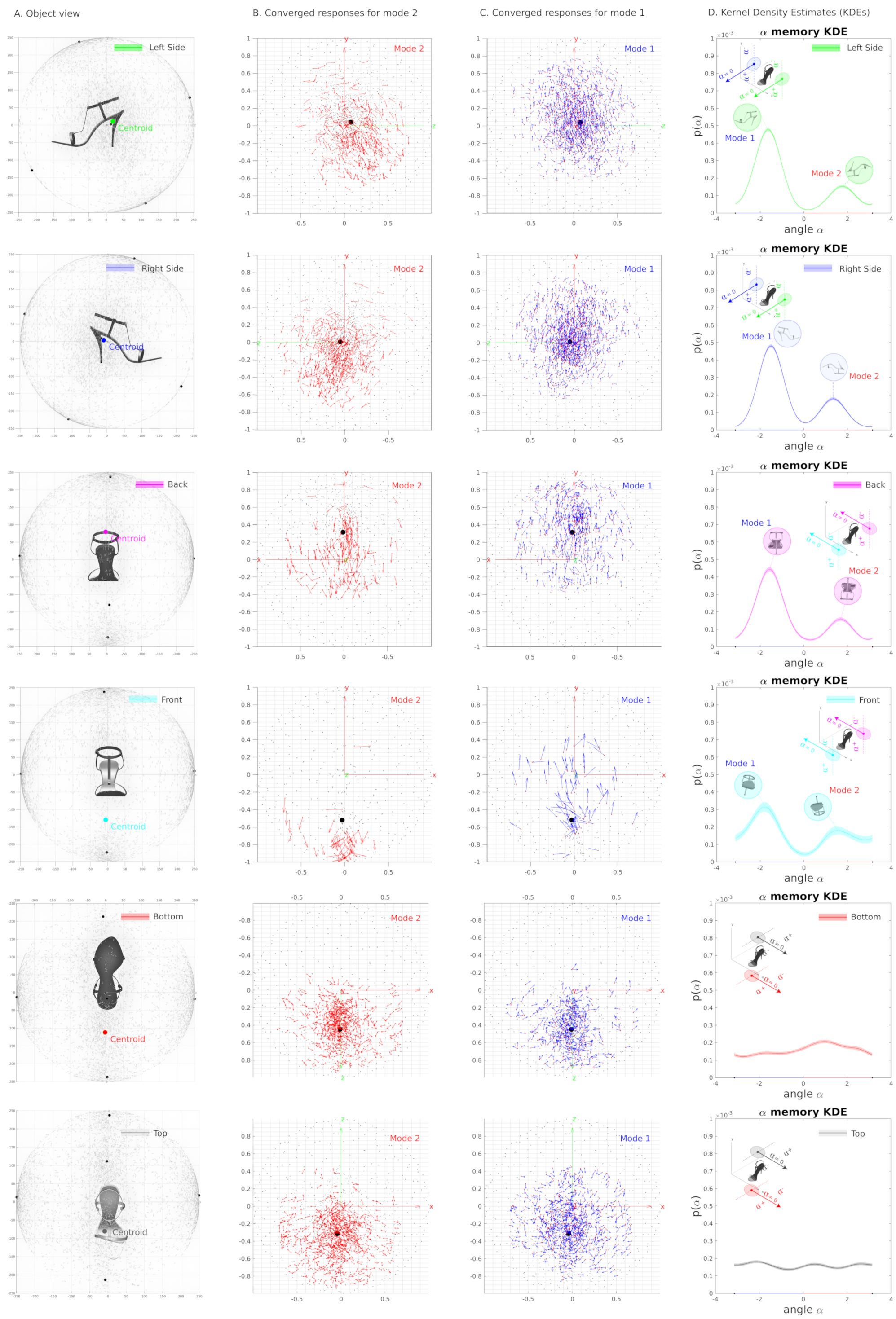
Extended Data Figure 3 Example full serial reproduction chain results and KDEs for the camera object. A. shows quiver plots of the raw data, and B. shows the positional KDEs. Thumbnails in the upper right of each of the subplots indicates the orientation of the object for reference.



Extended Data Figure 4 Orientation biases.

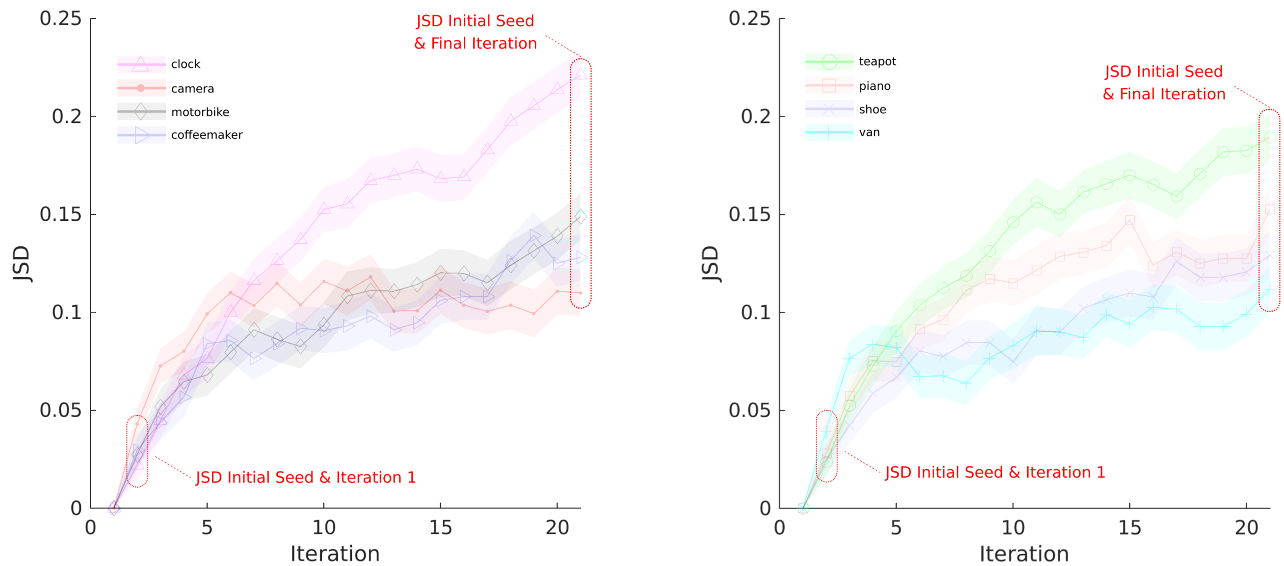


Extended Data Figure 5 Orientation biases.

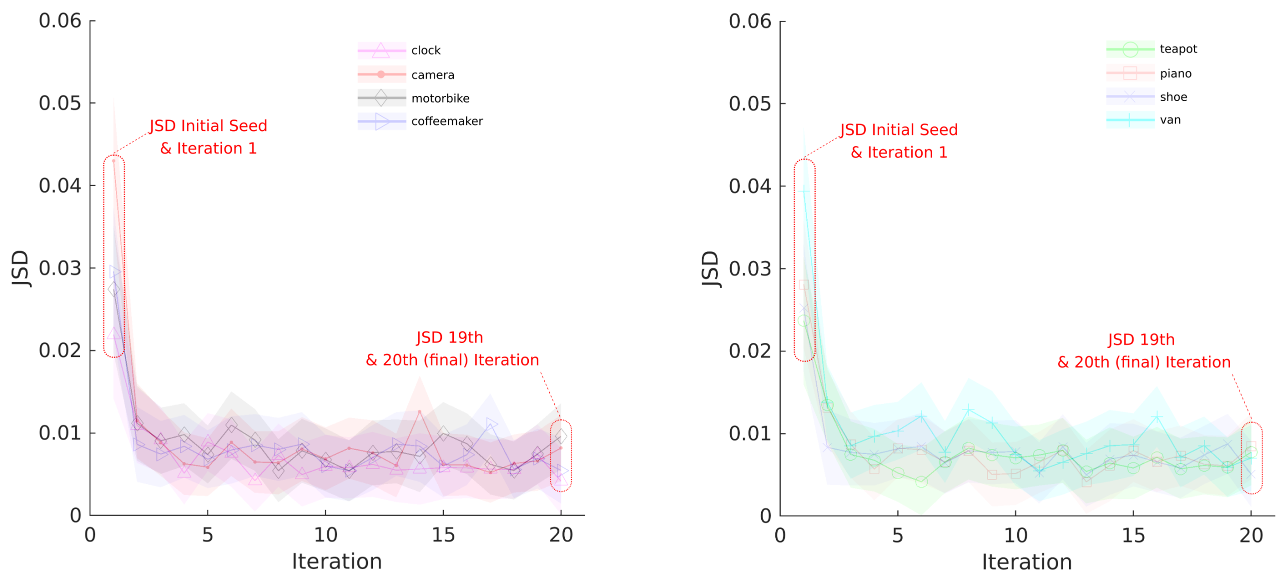


Extended Data Figure 6 Orientation biases in positional mode clusters. A. Thumbnails showing the views and the data as seen from one of the six view positions (centroids of the modes in the convergent serial reproduction chain iterations, which were estimated by aggregating the data across the last 10 iterations of the serial reproduction chains for this object). B. First row: quiver plot with red quivers shows the subset of responses in the cluster that had angle orientations that were negative relative to the $-z$ basis vector, which corresponds to an α angle of 0 (and vertical orientation of the object with the front facing up). These negative orientations are biased towards a view of the object that is upside down. C. First row: Quiver plot with blue quivers showing the remaining subset of responses in the cluster, which all had angle orientations that were positive relative to the $-z$ reference basis vector. These are biased towards a view that is oriented upright. Overall, the distributions of angles show two distinct modes. D. First row: Kernel Density Estimates (KDEs) of the angle data in the cluster, revealing the clear modes (mode 1 and mode 2) for that view cluster (side views). The results in each of the remaining rows show the same subdivision of the data in each of the clusters, along with the KDEs. Schematics in D provide illustrations of the reference vectors used for the analysis.

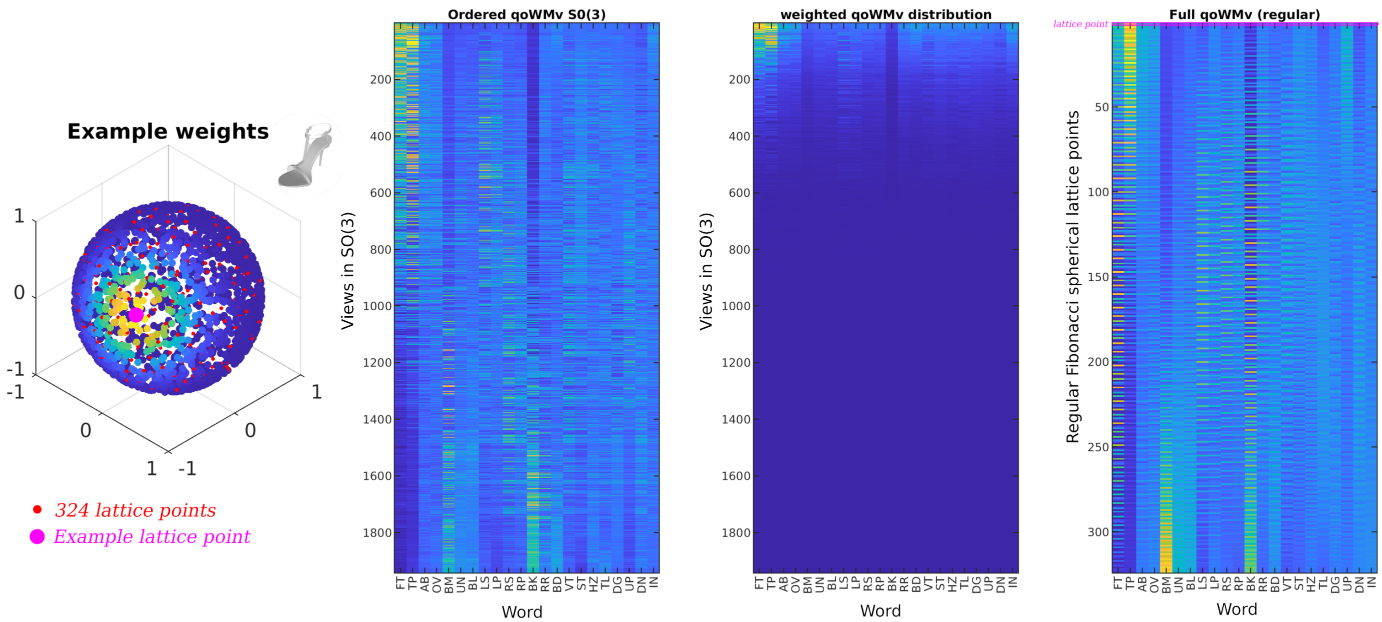
A. Jensen-Shannon Divergence (JSD) between seed and ith iteration bootstrapped KDEs



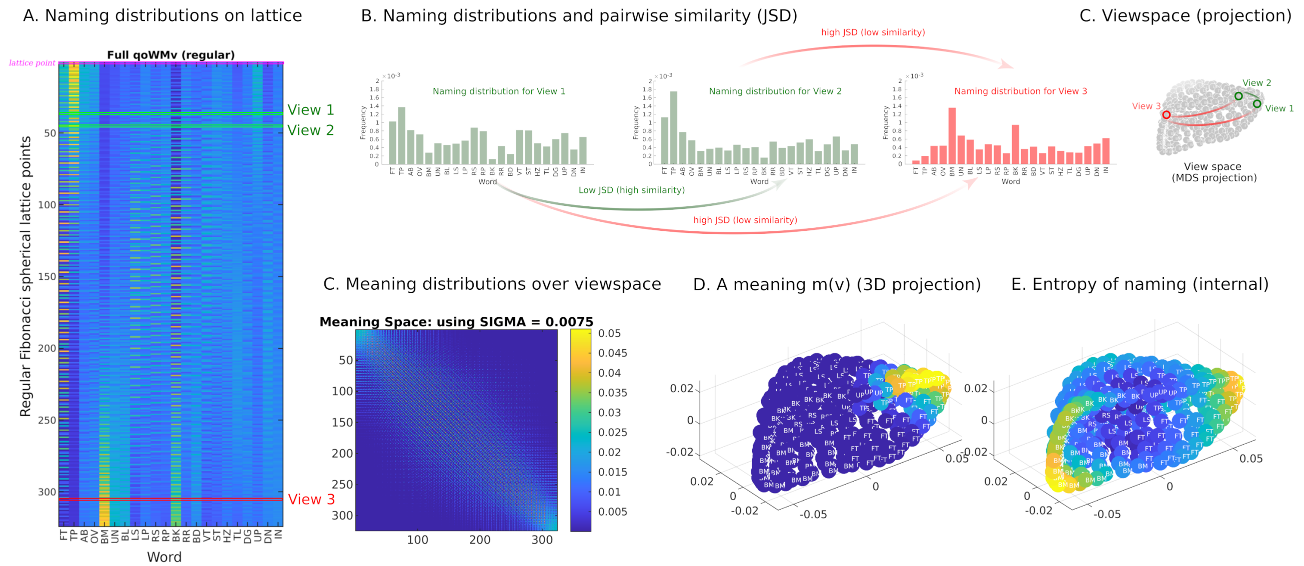
B. Jensen-Shannon Divergence (JSD) between bootstrapped KDEs for iteration i and (i+1)



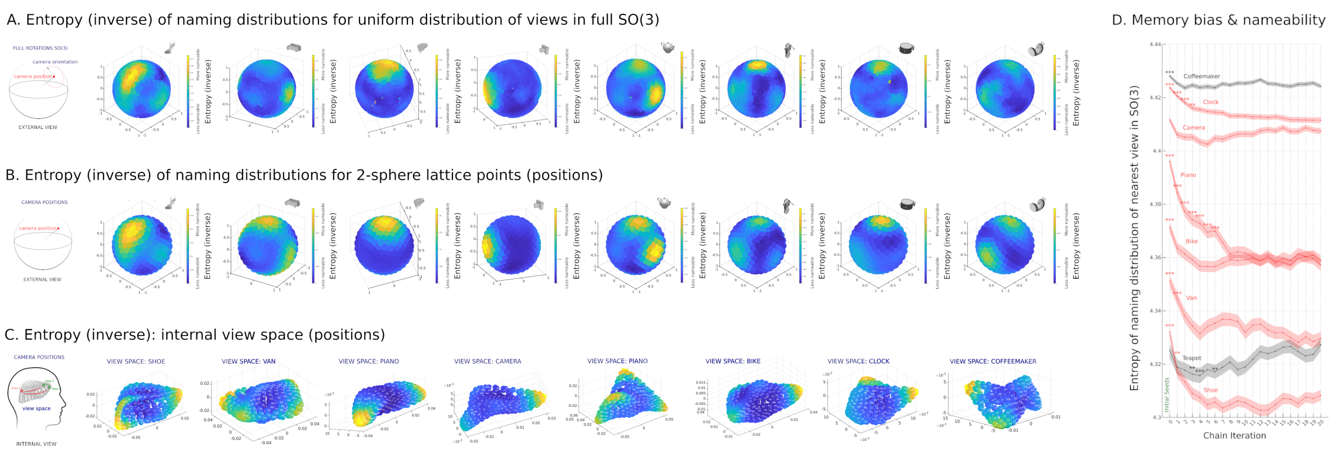
Extended Data Figure 7 Serial reproduction chain convergence analysis. All chains show convergence by the 11th iteration of the process $p < 0.0001$, except for the clock object, which showed convergence by the 17th iteration ($p < 0.0001$).



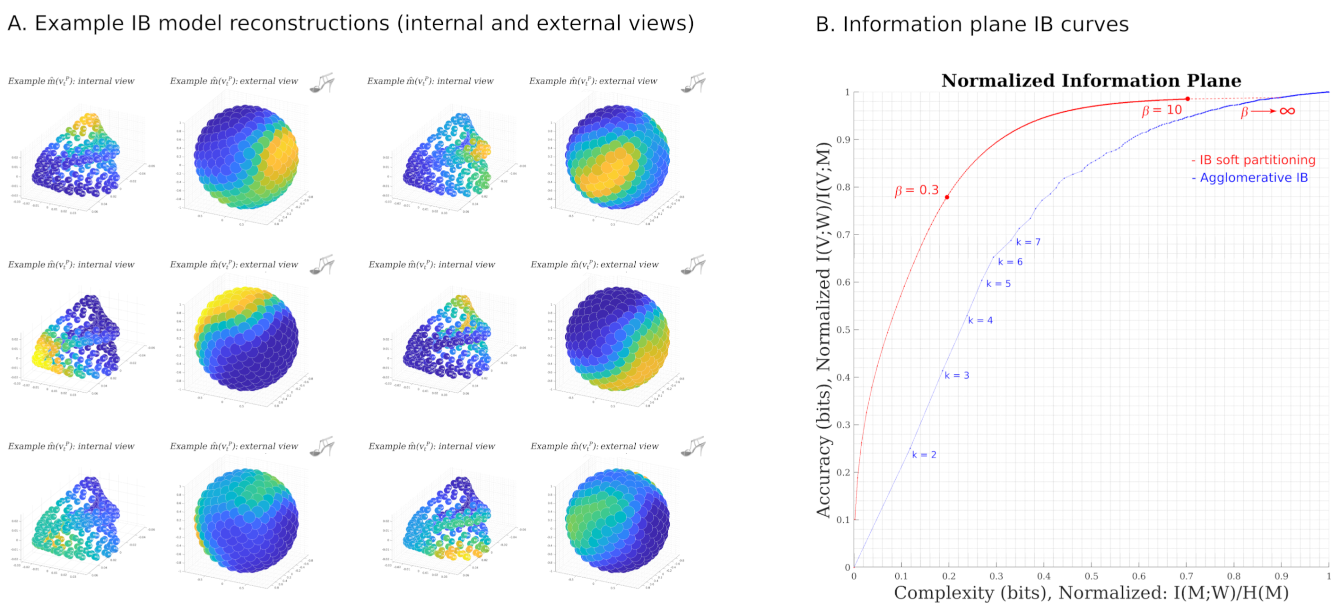
Extended Data Figure 8 Naming distributions for 1944 views in $SO(3)$ and weighted averaging procedure for estimating naming distributions for 324 lattice points on the 2-sphere (view positions). The red dots on the sphere (far left) are 324 evenly distributed spherical Fibonacci lattice points on the 2-sphere. The magenta point shows one of the lattice points and the center of a smoothing kernel used to weigh the naming distributions obtained for nearby views in $SO(3)$, which are also plotted (as dots without the orientation "up" vectors), where the colormap indicates the weight for each view around the example lattice point under the Gaussian smoothing kernel. The matrix in the second column of the figure shows all naming distributions (each row is a distribution) for each of the 1944 uniformly distributed views in $SO(3)$. The acronyms on the x-axis indicate each of the 22 words in the forced-choice naming experiment. The second matrix (second from the right) shows the same matrix of naming distributions weighted according to the smoothing kernel centered on the example lattice point. For each of the 324 lattice points, we averaged all the weighted naming distributions. The resulting weighted average naming distributions are shown for each of the 324 lattice points in the matrix on the far right. The magenta line at the top of the matrix shows the final average result for the example lattice point shown on the sphere on the far left.



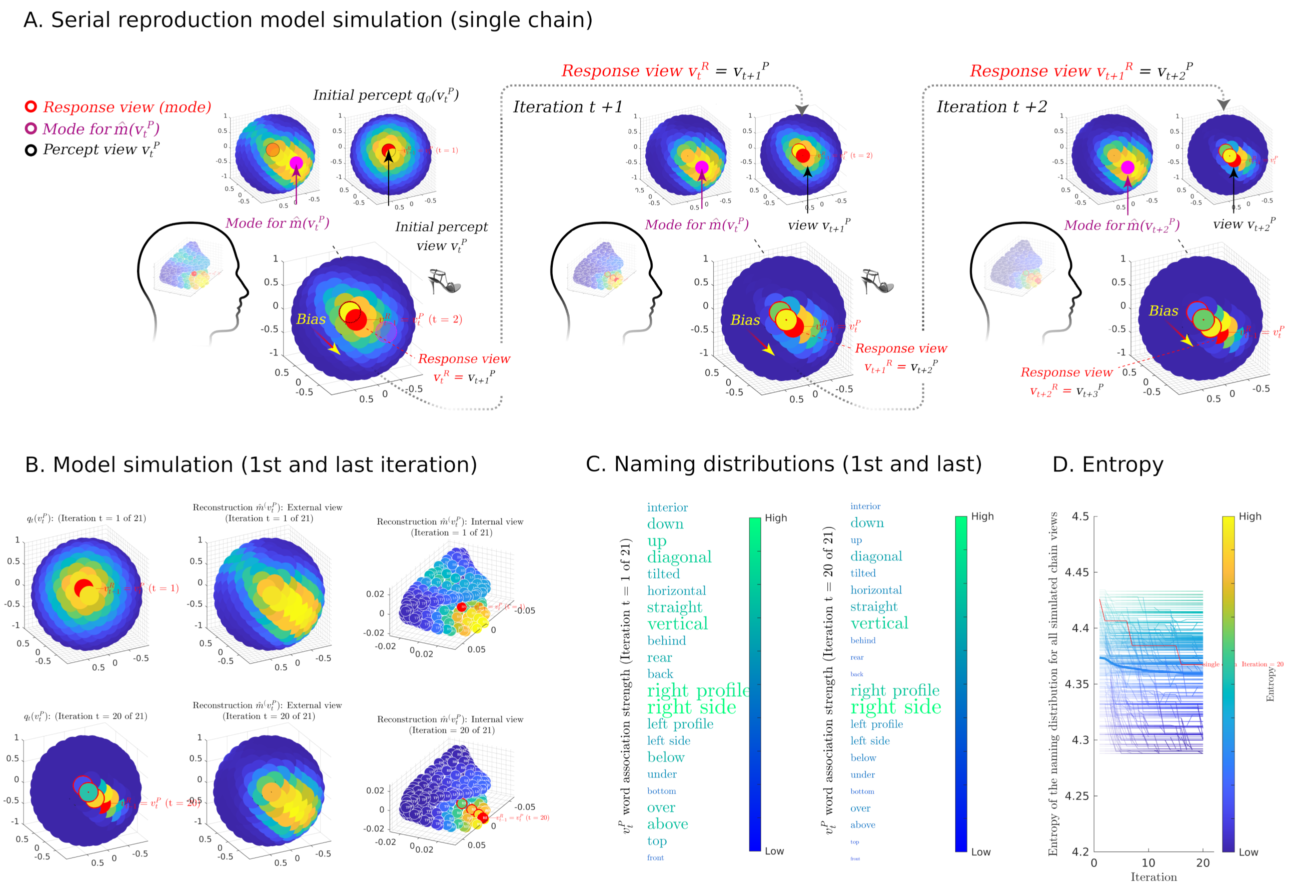
Extended Data Figure 9 View space estimation (shoe object example). A. Naming distribution (each row is a weighted average of all 1944 naming distributions obtained for uniform distribution of views in $SO(3)$) for each of the 324 lattice points on the 2-sphere (view positions). Highlighted rows indicate 3 example views shown in B. B. View space estimation. We computed all pairwise JSDs between each pair of view positions on the 2-sphere. 2 views with similar naming data (view 1 and 2) will be closer in the semantic space than dissimilar views (view 1 and 3 or view 2 and 3). C. 3D MDS projection of a view space (shoe object example). The red and green points illustrate semantic similarity between the 3 example views. C. Meaning distributions in the view space. Each column of the matrix is an isotropic Gaussian with a diagonal covariance matrix $\Sigma = I_3 \cdot 0.0075$ centered at a view v in the semantic space. D. Visualization of an example $m(v)$ in the view space. E. Visualization of the normalized entropy (inverse) of the average naming distributions for each of the 324 spherical Fibonacci lattice points in the internal representation.



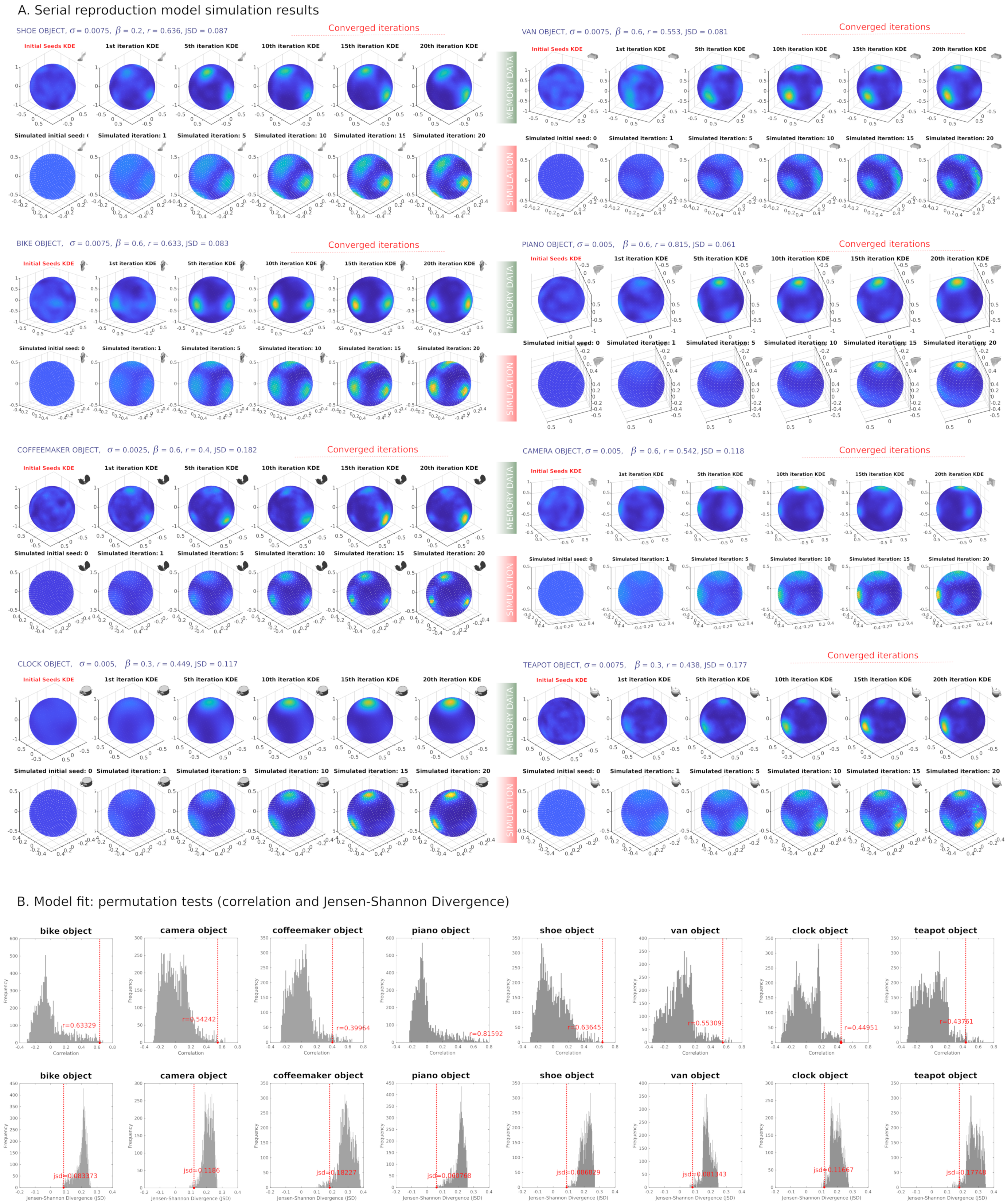
Extended Data Figure 10 Memory bias and nameability. A. Inverse of the normalized entropy of naming distributions for 1944 uniformly distributed views in $SO(3)$ for all objects. Points rather than vectors are plotted to reduce clutter in the subplots. The yellower color indicates views with *lower* entropy (more nameable views), while bluer colors indicate views with *higher* entropy (views that were less nameable). B. Inverse of the normalized entropy for view positions on a spherical Fibonacci lattice over the sphere. We averaged views with a Gaussian kernel weight centered at each view position on the spherical lattice to estimate the nameability of view positions (2-sphere). C. Shows the same as panel B in a 3D MDS projection of the internal view space representation for view positions on the spherical lattice, where pairwise distances are proportional to the JSD between naming distributions for each view position. Initials over each point (view position) show the label mode in the naming distributions for that view position. D. Memory bias and nameability. For each object, for each chain memory response and for each iteration, we computed the entropy of the view in $SO(3)$ that was the nearest neighbor to that memory response, and averaged over all chains. In nearly all cases the results show a decrease between the initial seed distribution and the final iteration when we compared the results in the final iteration to the results in all other iterations including the initial seed ($p < 0.001$ with the Bonferroni correction applied to correct for multiple comparisons). The errorbars were estimated from bootstrapping 1000 random samples of the data with replacement. The red lines show the results for the objects shown in panels A-C. These results show that memory is biased towards views that have lower entropy in naming.



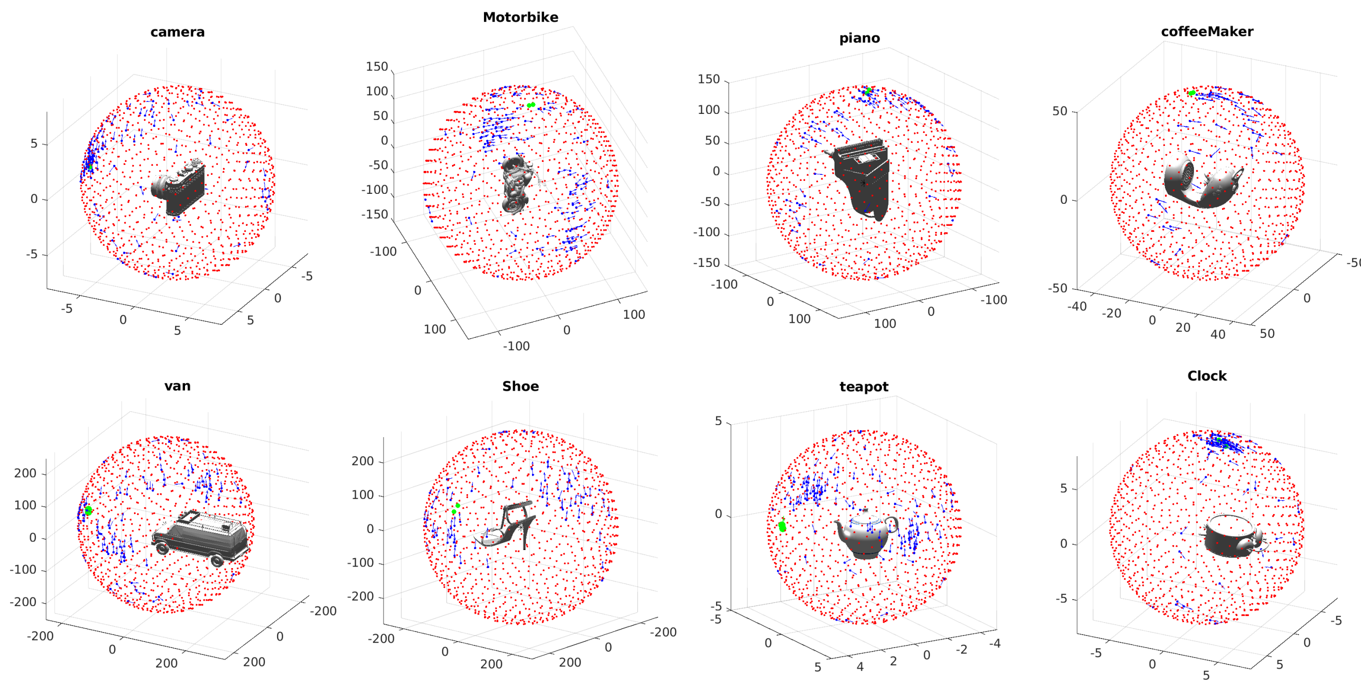
Extended Data Figure 11 Information Bottleneck (IB) information plane and curves and example reconstructions $\hat{m}(v)$ for the shoe object. A. Several example reconstructions (internal and external view representations). These were estimated using a low value of the tradeoff parameter β that results in higher compression of meanings $m(v)$. B. Information plane with curves showing results of (in red) using the self-consistent equations (see methods) to minimize the IB Lagrangian [26, 28, 32], and (in blue) an agglomerative IB algorithm [36]. We used the first method rather than the agglomerative hard partitioning method for our simulations, since they produce much better results: note that the red line is above the blue line, indicating that it achieves a much better complexity and accuracy tradeoff. The red curve was estimated for β values ranging from 0 to 10 and we initialized the algorithm with an identity matrix for each value of β (rather than using reverse deterministic annealing). Higher values of β result in more complex encoders and more accurate reconstructions of meanings. Using IB to compress meanings in the internal view spaces for all the objects allowed us to estimate the boundaries of semantic visuospatial categories that predict memory biases in the serial reproduction experiments using only the word frequency data.



Extended Data Figure 12 IB serial reproduction model. Modeling a single serial reproduction chain. A. Single chain example. The initial sensory percept $q_0(v_t^P)$ and its reconstruction $\hat{m}(v_t^P)$ through the communication model are combined by an element-wise product. The mode of the resulting distribution becomes the stimulus for the next simulated participant in the chain. This produces a bias between the initial view and first view reconstruction (yellow arrow indicates the bias, which is the change in the view position. View position is shown as a red dot on the 2-sphere). We repeated this process for 20 iterations. B. Top row shows the initial percept $q_0(v_t^P)$, semantic reconstruction $\hat{m}(v_t^P)$ in external coordinates, and again in the internal view space. The red dot shows the location of the view v_t^P. The bottom row shows the same results at iteration 20 of the process. Note the formation of a chain of biased responses. C. naming distributions for views v_t^P in the initial seed and iteration 20. The size and colors of the words are proportional to the density of the naming distributions for v_t^P at $t = 0$ and $t = 20$. D. Entropy of the naming distributions for all 324 chains and for all iterations. The chain in red highlights the example shown in A-C. Note the drop in entropy over multiple iterations of the serial reproduction process showing simulated reconstructions towards more nameable views (with lower entropy in naming).



Extended Data Figure 13 All simulation results. A. Serial reproduction KDEs for view position for the seed iteration, and iterations 5, 10, 15, and 20 of the process, and for all objects. Also shown are the simulated KDEs and optimal parameter fits for the perceptual noise parameter σ and the IB tradeoff parameter β. B. Permutation tests evaluating the correlation and JSD of the simulation results to 10000 random rotations of the memory KDEs.



Extended Data Figure 14 Canonical view experiment raw results. Subjective estimates of "best" or most "typical" views for all 8 objects. The results deviate in important ways from the memory results.

Experiment number	Experiment type	Stimulus	Number of participants	Number of chains
1	Memory Serial Reproduction	Camera	139	500
2	Memory Serial Reproduction	Shoe	118	496
3	Memory Serial Reproduction	Teapot	149	500
4	Memory Serial Reproduction	Clock	144	499
5	Memory Serial Reproduction	Van	167	500
6	Memory Serial Reproduction	Bike	157	500
7	Memory Serial Reproduction	Piano	133	499
8	Memory Serial Reproduction	Coffeemaker	143	500
9	Word list	NA	50	NA
10	Image Labelling nAFC	Camera	243	NA
11	Image Labelling nAFC	Shoe	243	NA
12	Image Labelling nAFC	Teapot	243	NA
13	Image Labelling nAFC	Clock	243	NA
14	Image Labelling nAFC	Van	243	NA
15	Image Labelling nAFC	Bike	243	NA
16	Image Labelling nAFC	Piano	243	NA
17	Image Labelling nAFC	Coffeemaker	243	NA
18	Canonical views	Camera	106	NA
19	Canonical views	Shoe	70	NA
20	Canonical views	Teapot	114	NA
21	Canonical views	Clock	80	NA
22	Canonical views	Van	92	NA
23	Canonical views	Bike	115	NA
24	Canonical views	Piano	73	NA
25	Canonical views	Coffeemaker	38	NA

Extended Data Figure 15 Table showing the number of participants, and number of chains (where applicable), for all experiments.