Disorganization impairs cognitive maps built from visual inputs

Xiaoyan Wu[1*], Finn Rabe[1], Peizheng Wu[1], Victoria Edkins[1], Yves Pauli[1], Wolfram Hinzen[3,4], Iris E. Sommer[5], Philipp Homan[1,2*]

[1]Department of Adult Psychiatry and Psychotherapy, University of Zurich, Zurich, Switzerland.
[2]Neuroscience Center Zurich, University of Zurich and ETH Zurich, Zurich, Switzerland.
[3]Department of Translation & Language Sciences, Universitat Pompeu Fabra, Barcelona, Spain.
[4]Institucio ´ Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.
[5]Department of Neuroscience, University Medical Center Groningen, Antoni Deusinglaan 2, room 117, Groningen, Netherlands.

*Corresponding author(s). E-mail(s): xiaoyan.wu@bli.uzh.ch; philipp.homan@bli.uzh.ch;

Abstract

The human brain organizes fragmented information into internal maps that support inference of latent relational structure. Vulnerabilities in this mapping process have been hypothesized to contribute to psychosis-spectrum disorders. Whether internal maps arise as readily from visual input as from linguistic description remains unknown. Here, participants learned food items associated with two abstract attributes, presented as images or language, organized along two abstract dimensions within a two-dimensional conceptual space. Participants then completed similarity judgments and reward-learning tasks that required inference of the underlying relational structure. We show that cognitive disorganization, a hallmark of psychosis-spectrum traits, selectively impairs spatial sensitivity and reward generalization when relational structure must be inferred from visual input, but not when matched structure is provided by linguistic descriptions. These findings support the shallow cognitive map hypothesis and

suggest that cognitive disorganization selectively disrupts the transformation of perceptual input into stable relational representations.

Keywords: Cognitive disorganization, Psychosis spectrum, Cognitive maps, Spatial generalization, Computational modeling

Introduction

When exploring a new city, we integrate perceptual fragments such as streets and landmarks into an internal representation that supports navigation in the city. A similar representation might also be formed without direct experience; for instance, one may gain an impression of a city's layout solely from a friend's vivid written description. This may be because linguistic descriptions activate pre-existing knowledge, enabling the formation of an internal structure even without direct perceptual experience. Understanding how such fragmented experiences are organized into a coherent internal structure is central to theories of memory, reasoning and prediction. A widely held view is that the hippocampal–entorhinal system constructs cognitive maps that encode relationships among concepts or events, enabling flexible generalization across physical and abstract spaces [1–5]. These cognitive maps provide a framework that supports the inference of latent relational structure [6], the prediction of expected outcomes [7] and value-guided choice [8, 9].

The brain mechanisms involved in the development and maintenance of cognitive maps appear to be particularly vulnerable in schizophrenia spectrum disorders. A core psychosis spectrum trait, cognitive disorganization, is characterized by difficulties in organizing information and maintaining coherent cognition [10]. In individuals with schizophrenia, language output disturbances are widely considered to reflect disorganized thinking and abnormalities in conceptual and semantic organization (for review, see [11]). The recently proposed shallow cognitive map hypothesis suggests that cognitive disturbances in schizophrenia may reflect disruptions in the hippocampal–entorhinal system that supports relational structure [12, 13]. According to this hypothesis, hippocampal-entorhinal attractor dynamics become less stable, yielding shallower attractor basins and undermining coherent relational structure. Consistent with this hypothesis, individuals with schizophrenia show reduced reliance on semantic proximity when generating words within a category, and the magnitude of this reduction correlates with hippocampal–entorhinal activity [14]. Individuals with schizophrenia also exhibit impaired accuracy in physical-space navigation and diminished grid-like signals in the entorhinal–hippocampal system [15].

However, it remains unclear whether the cognitive disorganization characteristic of the psychosis spectrum reflects a generalized failure in relational reasoning or a specific vulnerability in transforming fragmented perceptual inputs into coherent internal structures. Convertino et al. [15] showed that schizophrenia is associated with impairments in physical spatial navigation, but the impact on the navigation of the relational structure derived from abstract conceptual knowledge remains untested. Studies of abstract knowledge mapping have typically used visual stimuli such as images or scenes

as the inputs to dimensional spaces [5, 8, 9, 16, 17], which provide no semantically explicit information. In these paradigms, the mapping between objects and their locations in the space is often arbitrary, for example, pairing the length of a duck's neck and legs with a Christmas icon [16]. Other studies have used audiovisual elements as stimuli and assigned novel symbolic labels to them, and similar grid-like neural coding has also been observed for the resulting two-dimensional (2D) feature spaces [18, 19]. These findings suggest that cognitive mapping mechanisms support relational structure in abstract knowledge spaces both when structure is inferred from fragmented perceptual inputs and when it is derived from learned concepts. Yet the degree to which the provision of explicit semantics, as words or texts, facilitates the construction and use of internal relational structure relative to purely visual inputs that do not provide explicit meaning remains unknown. No study has directly compared these two forms of input and examined whether they differentially affect individuals with schizophrenia spectrum disorders.

Schizotypy provides a dimensional expression of psychosis-spectrum traits in non-clinical populations, as measured by the Oxford-Liverpool Inventory of Feelings and Experiences (O-LIFE) [20]. Here, we examine how varying levels of schizotypy influence the construction and use of cognitive maps derived from fragmented inputs presented as image-based or language-based stimuli. Participants learned two graded dimensions by fragmented exposure, associated them with food items, and subsequently used the inferred 2D conceptual map to support similarity judgments and reward-based decisions. We found that when the inputs were image-based, higher cognitive disorganization was associated with reduced spatial sensitivity in the similarity judgments, and reduced spatial generalization in reward-based decisions. No such association was observed when the inputs were language-based. These findings suggest that a central vulnerability in the psychosis spectrum lies in transforming visual inputs into a stable internal structure, whereas the ability to operate on already organized, language-based structure is preserved.

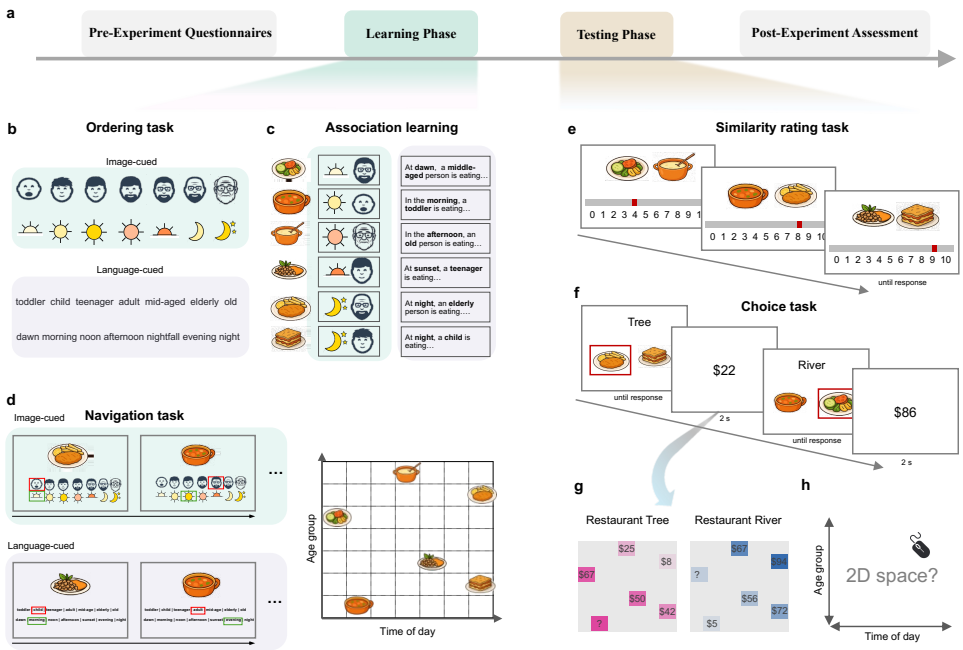

Fig. 1 Overview of the experimental procedure and tasks. a. Schematic timeline of the experiment, including the pre-experiment questionnaires, the learning phase, the testing phase, and the post-experiment assessment. **b.** Ordering task: participants learned the grading of two variables, the age group and time of day, presented either with images (upper panel: faces and sun–moon icons; image-cued condition) or with English words (lower panel: language-cued condition). **c.** Association learning: participants memorized six food items with a specific combination of age group and time-of-day cues. The stimuli of two grading variables were either presented as images (left panel: image-cued condition) or English short sentences (right panel: language-cued condition). **d.** Navigation task: participants were shown a food item and asked to select the correct "age group" and "time of day" associated with it. Each trial required participants to make both selections correctly. **e.** Similarity rating task: participants rated the similarity between pairs of food items on a 0–10 scale based on their learned age–time associations. **f.** Choice task: participants chose between two food items on each trial. After each decision, the value of the selected item was displayed as feedback. Food items were drawn from either the *Tree* or *River* "restaurant," with their values determined by their locations in an unshown 2D space defined by age group and time of day. **g.** Value structures used in the choice task. Each participant completed the task in two value contexts (Restaurant *Tree* in pink, Restaurant *River* in blue), with context–value structure assignment randomized across participants. Values were generated from item coordinates using either Value $= \frac{x+2y}{18} \times 95 - 5$ (pink structure) or Value $= \left(1 - \frac{x+y}{12}\right) \times 95 - 5$ (blue structure). Items marked with question marks were never shown in the choice task. **h.** Placement task: participants were shown a 2D map and asked to drag each food item to its estimated location within the map.

Results

The experimental procedure included a learning phase and a testing phase (Fig. 1a). During the learning phase, participants learned six food items and their associated

attributes along two abstract dimensions: the age group of the person who eats the food (seven ordered levels ranging from toddler to old) and the time of day of eating (seven ordered levels ranging from dawn to night). These two dimensions formed a 2D conceptual map. During training, participants were randomly assigned to either an image-cued ($N = 51$, Fig. 1b-d, green-shaded panels) or a language-cued ($N = 60$, Fig. 1b-d, gray-shaded panels) condition, in which stimuli representing the two dimensions were presented in the form of images or text, respectively.

Participants first learned the rank relationships within each dimension by judging the relative order of two levels (e.g., whether "toddler" is younger or older than "adolescent") until performance achieved an accuracy threshold (85%). They then learned the six associations of food attributes through association learning (Fig. 1c), in which the participants selected the correct food item given a specific combination of attributes, and navigation tasks (Fig. 1d) that required identifying the correct attributes for each food item. Training continued until the performance reached the same threshold.

In the subsequent testing phase, we assessed participants' ability to infer unobserved relationships within the 2D conceptual map using a similarity rating task ((Fig. 1e), as well as their ability to use map-based information to support reward generalization using a choice task (Fig. 1f). All participants completed identical tasks and were presented with the same stimuli during the testing phase. All participants completed the experiment in a single session conducted on the same day.

Higher cognitive disorganization reduces spatial sensitivity specifically in visually derived cognitive maps

We first examined whether distances on the unshown 2D map (Fig. 1d, right panel) could predict participants' judgments of similarity between food items (Fig. 1e), and how cognitive disorganization (assessed using the O-LIFE) influences this process. Participants were instructed to judge the similarity between pairs of food items based solely on who ate them and when they were eaten.

A linear mixed-effects model revealed that dissimilarity scores (reverse-scored similarity) increased with 2D distance ($\beta = 0.62$, $95\%CI[0.56, 0.67]$, $p < 0.001$; Fig. 2b). This indicates that participants were able to infer unobserved relationships within the 2D conceptual map, despite never seeing the map directly, and that these inferences guided their similarity judgments.

We next examined whether the ability to infer relationships within the 2D conceptual map was associated with cognitive disorganization. Here the spatial sensitivity was quantified as the Fisher's z-transformed correlation between pairwise 2D distances and dissimilarity scores. We found that spatial sensitivity was negatively correlated with cognitive disorganization in the image-cued condition ($r = -0.33$, $95\%CI[-0.56, -0.06]$, $p = 0.017$), but not in the language-cued condition ($r = -0.08$, $95\%CI[-0.33, 0.18]$, $p = 0.554$, $BF_{01} = 6.45$; $Fisher'sZ = -1.34$, $p = 0.090$; Fig. 2c).

A similar dissociation between conditions was also observed for the association between spatial sensitivity and the total O-LIFE score: a significant correlation was observed in the image-cued condition but not in the language-cued condition (see Supplementary Fig. A3a). In contrast, other personality measures, such as the Big Five,

showed no associations with spatial sensitivity (Supplementary Fig. A3 and Fig. A7), indicating that the link between spatial sensitivity and schizotypy-related cognitive disorganization is selective rather than reflecting general personality variation.

Furthermore, in the image-cued condition, sensitivity to 2D distance was positively correlated with mental rotation performance ($r = 0.34$, $95\%CI[0.07, 0.56]$, $p = 0.015$), linking spatial sensitivity to spatial transformation ability. No such relationship was observed in the language-cued condition ($r = 0.02$, $95\%CI[-0.24, 0.27]$, $p = 0.908$, $BF_{01} = 7.69$; $Fisher'sZ = 1.71$, $p = 0.044$; see also Supplementary Fig. A7).

In brief, participants' similarity judgments reflected distances within an unobserved 2D conceptual map. In the image-cued condition, higher levels of cognitive disorganization were associated with lower spatial sensitivity, and higher spatial sensitivity was associated with better spatial transformation ability. These associations were not observed in the language-cued condition.

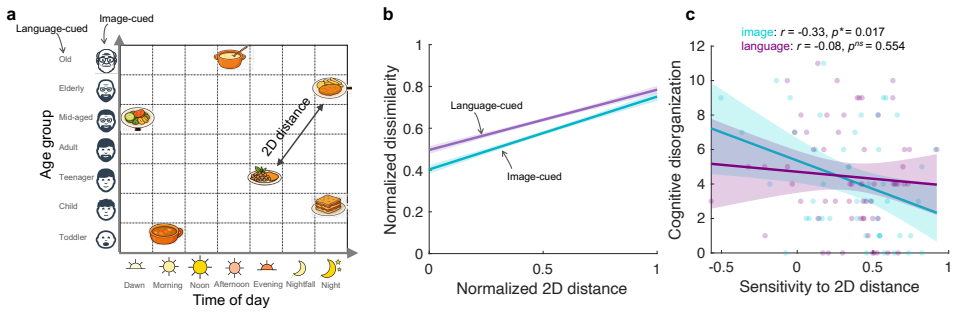

Fig. 2 Results of the similarity rating task. **a.** Example of the 2D age–time abstract space used to define pairwise distances between food items. The solid arrow illustrates the 2D Euclidean distance. **b.** Results of the similarity rating task. The dissimilarity scores (computed as 10 - similarity), min–max normalized to the [0, 1] range, increased with spatial distance in both the image-based and language-based conditions. **c.** Correlation between cognitive disorganization scores (O-LIFE subscale) and sensitivity to 2D distance. Sensitivity to 2D distance was quantified as the Fisher z-transformed correlation between pairwise 2D distances and the corresponding dissimilarity scores, with higher values reflecting stronger spatial sensitivity. For panels b and c, lines show separate linear fits for the image (cyan) and language (purple) conditions, with shaded areas representing 95% confidence intervals. Significance levels: $^*p < 0.05$; ns = not significant.

Spatial memory predicts reward learning performance specifically in visually derived cognitive maps

If participants form an internal spatial representation from sequentially fragmented inputs, better memory for this space should support more effective use of spatial information for problem solving. To this end, participants completed a choice task in which they selected one of two food items and received a reward for the chosen option (Fig. 1f). They were instructed to choose the option with the higher reward. Unknown to participants, the rewards associated with the food items were systematically related to their locations in the 2D space (Fig. 1g).

We found that participants in both the image-cued and language-cued conditions improved their decisions over time (Fig. 3a). At the end of the experiment, participants were asked to locate all food items within a 2D space in a placement task (Fig. 1h). Accuracy was quantified as the distance between participants' placed locations and the true coordinates, defined as the 2D placement error (Fig. 3b). Participants were generally able to place the food items close to their correct locations, and no difference was observed between the image-cued and language-cued conditions ($t(109) = 0.79$, $95\%CI[-0.68, 1.57]$, $p = 0.432$, $BF_{01} = 7.68$; see Supplementary Fig. A4). These results indicate that, despite never seeing the 2D map directly and reporting no explicit awareness of a map-based strategy, participants were nonetheless able to reconstruct the hidden spatial structure with considerable accuracy.

A key question arises whether more accurate spatial memory for the food items (i.e., lower 2D placement error) predicts better performance in the choice task. The answer is yes: higher accumulated reward in the choice task was associated with lower 2D placement error ($r = -0.42$, $95\%CI[-0.62, -0.16]$, $p = 0.002$), but this relationship emerged only in the image-cued condition (language-cued condition: $r = -0.16$, $95\%CI[-0.40, 0.09]$, $p = 0.212$, $BF_{01} = 3.44$; $Fisher's Z = 1.46$, $p = 0.07$; Fig. 3c).

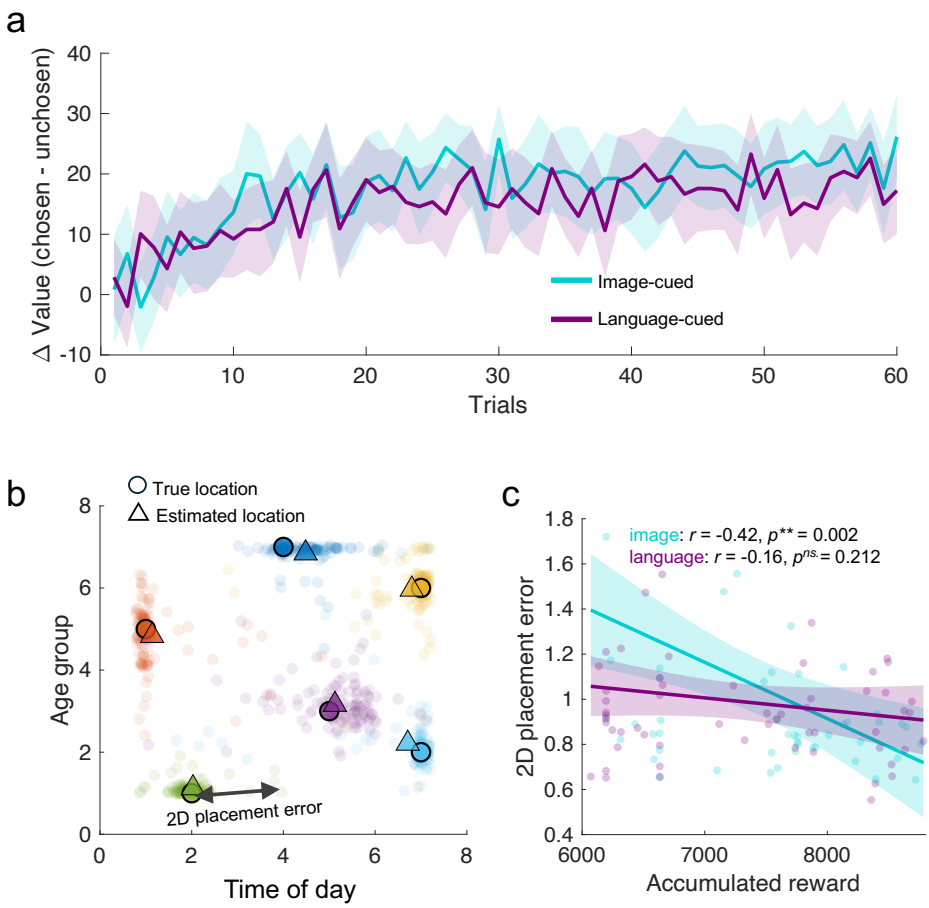

Fig. 3 Results of the choice task. **a.** Trial-by-trial dynamics of decision value differences (Δ value $= V_{\text{chosen}} - V_{\text{unchosen}}$) across 60 trials in two restaurant contexts for the image-cued (cyan) and language-cued (purple) conditions. Each restaurant context comprised two blocks of 30 trials; for visualization, the two blocks corresponding to the same restaurant are concatenated here. Across the experiment, the four blocks (two restaurants $\times$ two blocks) were presented in randomized order. **b.** Comparison of the true locations and participants' estimated locations for the food items within the 2D map from the placement task. Circles indicate the true item positions, and triangles indicate the participants' average estimated positions. Each transparent colored dot represents one participant's estimated location, with different colors denoting different food items (i.e., different locations). The solid arrow illustrates the 2D placement error for one participant in one example food item, i.e., the Euclidean distance between the true and estimated item locations (lower values indicate greater accuracy). **c.** Correlation between accumulated reward (x-axis) in the choice task and accumulated 2D placement errors (y-axis). The accumulated reward refers to the total gains earned during the choice task. Lines show separate linear fits for the image-based (cyan) and language-based (purple) conditions, with shaded areas representing 95% confidence intervals. Significance levels: $^{**}p < 0.01$; $ns.$ = not significant.

Higher cognitive disorganization attenuates spatial generalization specifically in visually derived cognitive maps

We compared four computational models (as specified in the pre-registration) to investigate the mechanisms underlying value-based decisions in the choice task:, including a baseline model, a model-free Q-learning model, a spatial Gaussian process regression (GPR) model, and a hybrid model.

Participants' behavior in the choice task was best described by the Spatial-GPR model (Fig. 4a). This model assumes that reward generalization is governed by similarity between options represented in a latent 2D space, such that rewards associated with previously chosen options are weighted by their similarity to the current option, as previously applied to reward generalization in physical space [8]. In the current study, this model well predicted participants' subjective values for the chosen food items (Fig. 4b–c; see also Supplementary Fig. A2 for posterior parameter distributions and parameter recovery).

Under the assumption of this model, participants should be able to infer the values of items at any location on the map by generalizing from learned spatial relationships. Indeed, the Spatial-GPR model produced value predictions that were well aligned with participants' estimated values and generalized well to food items that were never presented during the choice task (Supplementary Fig. A5). Moreover, participants whose behavior was best described by the Spatial-GPR model showed smaller deviations between model-predicted and estimated values than those described by other models (see also Supplementary Fig. A5, last panel).

This model assumes reward generalization across space via Gaussian process regression, with λ indexing the extent of spatial generalization, larger values indicate broader generalization (Fig. 4d). We next examined how this generalization process was related to cognitive disorganization. The results showed that narrower spatial generalization (smaller λ) was associated with higher cognitive disorganization. Consistent with the findings from the similarity rating task, this effect was observed only in the image-cued condition ($r = -0.31$, $95\% CI[-0.54, -0.04]$, $p = 0.027$), but not in the language-cued condition ($r = -0.09$, $95\% CI[-0.34, 0.17]$, $p = 0.482$, $BF_{01} = 5.99$; $Fisher's Z = 1.18$, $p = 0.120$; Fig. 4e). No associations between λ and other personality measures, such as the Big Five, were observed (see Supplementary Fig. A6 and Fig. A7), indicating that the observed association is selective for cognitive disorganization.

Moreover, λ was positively correlated with mental rotation performance in the image-cued condition ($r = 0.28$, $95\% CI[0.01, 0.52]$, $p = 0.043$), but not in the language-cued condition ($r = 0.22$, $95\%[-0.03, 0.45]$, $p = 0.084$, $BF_{01} = 1.63$; see also Supplementary Fig. A7), mirroring the similarity-rating results and linking spatial generalization in the image-cued condition to individual differences in spatial transformation ability. Individuals with higher cognitive disorganization show reduced spatial sensitivity when internal maps are constructed from visual inputs, and exhibit reduced use of these maps when decisions require spatial navigation and generalization; an effect that is not observed when maps are learned from language.

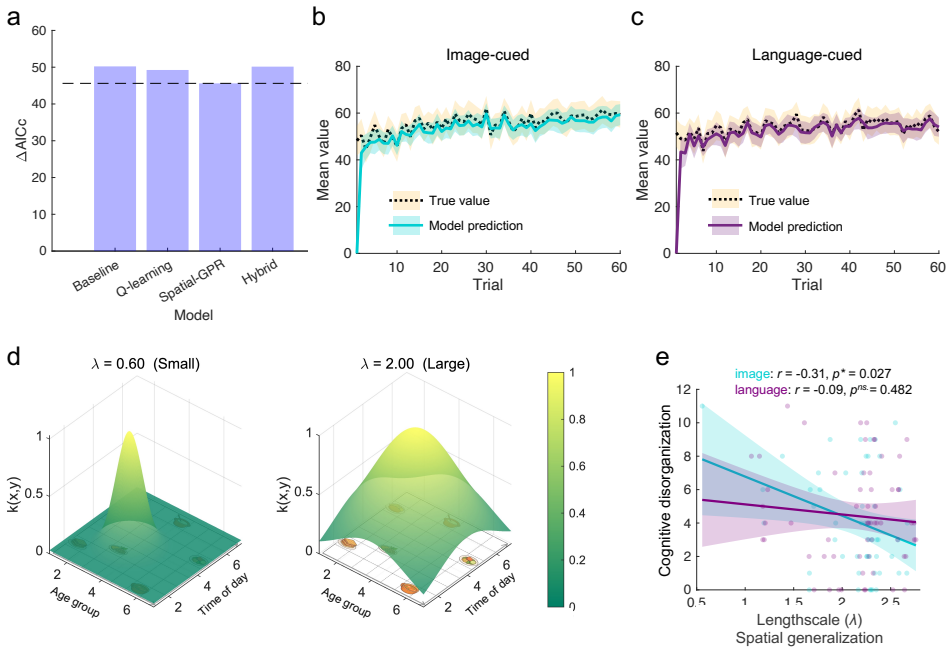

Fig. 4 Model-based analysis of choice behavior. **a.** Model comparison based on the corrected Akaike information criterion (ΔAICc; lower values indicate better model fit). **b-c.** Comparisons of the true value (dashed line) for the chosen option with the estimated value (solid line) from the best-fitting Spatial-GRP model for the image-cued and language-cued conditions, respectively. The shaded areas represent 95% confidence intervals. **d.** Kernel surfaces with different length-scale parameters λ over a semi-transparent background image. The left panel shows a sharper surface with a small length scale (λ=0.60), whereas the right panel shows a broader surface with a large length scale (λ=2.00). Higher λ indicates greater spatial generalization in subjective value computations during decision-making. **e.** Correlation between the length-scale parameter λ from the Spatial-GRP model and cognitive disorganization score (O-LIFE subscale). Lines show separate linear fits for the image-cued (cyan) and language-cued (purple) conditions, with shaded areas representing 95% confidence intervals. Significance levels: $^*p < 0.05$; $ns.$ = not significant.

Discussion

This study found that participants formed and used an internal 2D conceptual structure inferred from fragmented experiences, as reflected in both similarity judgments and reward-based decisions. Participants with higher cognitive disorganization showed reduced reliance on spatial representations learned from visual stimuli, whereas performance was largely preserved when the structure was provided through language. This pattern is more consistent with a difficulty in deriving coherent structure from visual perceptual input than with a general impairment in relational reasoning or reward learning. Such percept-to-structure transformations have been proposed to depend on hippocampal–entorhinal processes [21], which may help explain why visually based mapping was more sensitive to individual differences in disorganization.

Selective impairment for visually derived structure: links to the shallow cognitive map hypothesis

The finding that higher cognitive disorganization was accompanied by reduced reliance on spatial representations is consistent with theoretical proposals that schizophrenia-spectrum disturbances involve instability in cognitive map representations [13]. The shallow cognitive map hypothesis proposes that psychosis involves unstable hippocampal attractor dynamics that yield weaker and less coherent relational structure. Empirical studies have reported attenuated grid-like coding during physical spatial navigation [15], reduced global organization of semantic networks during movie watching [22], and reduced reliance on model-based planning during reward-guided decision making [23, 24]. Together, these findings suggest that schizophrenia-spectrum disorganization is associated with shallow and unstable cognitive map representations, particularly when relational structure must be inferred from visual perceptual input.

Consistent with this interpretation, in the current study, reward-based decisions relied on the remembered spatial layout when the internal map was built from visual inputs. This pattern suggests a greater reliance on abstract spatial structure during visual-to-structure transformations; a capacity that appears to be attenuated in schizophrenia-spectrum disorganization.

Gaussian-process models have been applied to explain reward generalization in physical space [8] and in visual feature spaces [25]. Here, we showed that reward generalization within a 2D conceptual space constructed from both image-based and language-based inputs is likewise well captured by a Gaussian-process model. We further showed that individuals with higher cognitive disorganization showed narrower generalization kernels (smaller length scales) in the image-cued condition, providing a computational signature of shallow or noisy spatial representations. Together, these findings provide a computationally grounded account that spatial sensitivity and reward-based spatial generalization were selectively weakened in individuals with higher schizophrenia-spectrum disorganization when structure had to be extracted from visual inputs.

Why linguistic structure remains preserved: evidence from language-based processing in psychosis

A striking feature of our results is the absence of schizotypy-related cognitive disorganization impairments in the language-cued condition. This dissociation provides important theoretical insight. Unlike visual stimuli, where relational structure must be actively inferred from variable sensory input, linguistic descriptions provide an explicit, symbolically structured scaffold for representing relations [26], thereby reducing the demands on active relational integration.

Several findings from schizophrenia-spectrum research align with this interpretation. Although incoherence in language output is typically considered to reflect disorganized thinking in psychopathology [27, 28], this association is not always observed [29, 30]. Recent large-sample computational work highlights that language-output anomalies can arise from multiple mechanisms, including differences in underlying associative structure, executive regulation over expression, or motivational

factors, rather than providing a direct readout of thought organization [31]. Accordingly, the absence of overt linguistic incoherence does not imply intact underlying cognitive structure.

In addition, the neural system that supports conceptual and linguistic structure appears to provide a relatively stable relational scaffold. Word meaning is represented using grid-like and distance codes in entorhinal and medial temporal regions [32], and neural representations of words and pictures show substantial overlap [33]. These findings indicate that linguistic input arrives in a format that is already organized into a relational structure. In line with this view, spatial sensitivity and generalization under language input were not related to individual differences in mental spatial ability, whereas both measures covaried with mental spatial ability when inputs were image-based. This dissociation suggests that linguistic cues engage relational structures that are already organized or abstracted. These findings raise the possibility that linguistic input engages a qualitatively different mapping process from that required to infer spatial structure from perceptual experience.

Taken together, this evidence suggests that language reduces the need to derive structure from noisy perceptual input, effectively bypassing the computational step most vulnerable to cognitive disorganization. According to this view, disorganization impairs the formation of spatially coherent maps from visual cues, but not the use of structure when it is linguistically provided. This account also explains why performance in the language-cued condition did not correlate with disorganization and why domain-general abilities such as mental imagery or spatial working memory did not predict performance: linguistic descriptions do not require the construction of an internal metric space in the first place.

Conceptual and methodological contributions

This study has made several conceptual advances. First, it refines the concept of cognitive maps by demonstrating that visually derived maps and language-supported relational representations are not equivalent: Spatial sensitivity, memory, and generalization depend on individual differences in spatial ability and cognitive disorganization only when relational structure must be inferred from perceptual (image-based) experience, but not when it is accessed through linguistic cues. Second, the findings show that impairments associated with cognitive disorganization emerge selectively when cognitive map structure must be inferred and organized from visual inputs. Third, the results suggest a conceptual role for language and conceptual systems as a relatively stable relational scaffold. Linguistic input appears to provide access to abstracted and familiar relational structure that is less dependent on online spatial transformation processes, potentially buffering against vulnerabilities in map formation. This perspective offers a new angle on how language may play a protective or compensatory role within the schizophrenia spectrum.

Methodologically, the paradigm grounds abstract structure learning in meaningful, non-arbitrary associations between age, time of day, and food items, allowing relational structure to be acquired through coherent experiences. All participants remained engaged throughout the study and completed the experiment within a single session. Note that our sample was not limited to university students: 20% of participants were

over 30 years of age, including a participant aged 64 years. In addition, we extend abstract knowledge maps to reward learning and decision making, and use computational modeling to show that reward generalization relies on inferred spatial structure rather than simple associative strategies.

Moreover, the current findings suggest that treatment approaches may benefit from incorporating environmental modifications that increase linguistic structure for individuals with high cognitive disorganization. In particular, explicit verbal labeling and structured psychoeducation may help make relational information more accessible, rather than requiring individuals to infer it solely from perceptual cues.

Finally, although the present study does not include neural measures, the behavioral and computational signatures it reveals offer a powerful assay for probing distinct neural contributions to cognitive map formation and map use. This framework lays the groundwork for integrating model-based behavioral quantification with neural measures (e.g., grid-like codes, representational geometry, structure-dependent value generalization) to examine how perceptual and linguistic structure interact in psychosis-spectrum cognition.

Conclusion

Together, our findings indicate that cognitive disorganization selectively affects the construction and use of cognitive maps derived from visual inputs, while leaving maps built from linguistic descriptions largely intact. This dissociation suggests that the core vulnerability in psychosis-related disorganization lies in transforming perceptual experiences into a stable relational structure. By contrast, symbolic formats such as language may provide external scaffolding that preserves reasoning even when internally generated structure is less stable. Understanding how perceptual and linguistic routes to structure interact, and how they fail, offers a promising path toward more mechanistic accounts of cognition in psychosis spectrum disorders. In sum, individuals with higher cognitive disorganization struggle when structure must be inferred from what they see, yet perform unaffected by cognitive disorganization when that structure is already supplied through language.

Acknowledgements. We thank Sebastijan Veseli and Todd Hare's group for insightful discussions, and all participants for their contributions to this study.

Funding. This project has received funding from the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (grant agreement No 101118756). It is part of the ERC Synergy project DELTA-LANG.

Declaration of competing interest. PH has received grants and honoraria from Novartis, Lundbeck, Mepha, Janssen, Boehringer Ingelheim, Neurolite, and OM Pharma outside of this work. No other disclosures were reported

Data and code availability. All data, materials, and codes are publicly available at https://github.com/xiaoyanwu2024/cogmap-ssd.git.

Author Contributions. The study was conceptualized by X. Wu, F. Rabe, and P. Homan. Methodology was developed by X. Wu and P. Homan. Software development, formal analysis, data curation, visualization, and project administration were carried out by X. Wu. Data collection was conducted by X. Wu and P. Wu. The original draft was written by X. Wu, and the manuscript was reviewed and edited by X. Wu, V. Edkins, F. Rabe, P. Homan, W. Hinzen and I. E. Sommer. Supervision and funding acquisition were provided by P. Homan.

Methods

This study is a pre-registered study (https://aspredicted.org/, #242157), and all analyses followed the pre-specified plan unless otherwise noted. Fisher's z-tests for between-group comparisons and Bayes factor analyses for non-significant results were conducted as additional analyses to quantify evidence and were not included in the original pre-registration.

Participants

Our target sample size was 80 participants. Given that the study involved demanding training tasks, we recruited more volunteers to compensate for potential dropouts or incomplete sessions. The final sample comprised 111 individuals, with 51 in the image group and 60 in the language group. Participants were right-handed adults ($\geq$18 years) recruited via social media, with normal or corrected-to-normal vision, C1-level or higher English proficiency, and no prior education or training in psychology. The final sample ranged in age from 18 to 64 years. Written informed consent was obtained from all participants prior to the experiment. The study procedures adhered to the principles of the Declaration of Helsinki and were approved by the Ethics Committee of the Psychiatrische Universitätsklinik Zürich.

Experiment procedures

Participants were assigned to one of two between-group input conditions. The grading of each of the two variables (i.e., age group and time of day) were presented as images with symbolic icons in the image-cued condition and as English text in the language-cued condition. The whole experiment included several parts and lasted 2–3 hours (mean = 2 h 10 min; Fig. 1a).

Pre-experiment questionnaires

Participants provided demographic information and completed standardized personality measures, including the short form of the $O-LIFE$ for schizotypy[20], the Big Five Inventory [34], and the Positive and Negative Affect Schedule ($PANAS$)[35]. To account for potential appetite-related confounds arising from the food images, participants also reported their hunger level and the time since their last meal. No significant group differences were observed across these measures (Supplementary Table A1).

Learning phase

In the learning phase, participants memorized each food item together with its corresponding "features," namely who ate it (age group) and when it was eaten (time of day). Participants were required to learn the grading of each of the two variables, presented either as image-based or language-based stimuli, through sequential fragmented exposure. They then learned the associations between the graded variables and the food items. No terms referring to "maps" or "two-dimensional spaces" were ever presented. Previous research has shown that such two-dimensional association learning can spontaneously form an internal two-dimensional representational space, i.e., a cognitive map.

Ordering task. Participants learned the grading of two variables, age group and time of day, through sequential fragmented exposure. Each variable consisted of seven graded levels (age group: toddler, child, adolescent, adult, middle-aged, elderly, old; time of day: dawn, morning, noon, afternoon, nightfall, evening, night). In each trial, participants judged the relative order of two randomly selected levels from the same variable (e.g., whether "toddler" is younger or older than "adolescent"). Training continued until participants reached at least 85% accuracy over the most recent 20 trials, ensuring that they understood the ordinal structure of each variable. The learning order of the two variables was randomized across participants. After reaching the accuracy criterion, participants were asked to arrange the seven levels of each variable from smallest to largest. Note that the inputs for the two variables were presented using different types of stimuli in the image-cued and language-cued groups. In the image-cued group, the categories were represented by cartoon-style faces (age) and symbolic icons (time of day, e.g., a sun or moon), whereas in the language-cued group they were presented as English words (Fig. 1b).

Association learning. Participants then learned the associations between the graded variables and the food items. The task involved six food items paired with specific combinations of age group and time of day. In each trial, a target age–time combination was shown at the top of the screen along with six candidate food items below. Participants selected the correct food item using the mouse and received immediate feedback on accuracy. Training continued until participants reached at least 85% accuracy across the most recent 20 trials. Again, the inputs for the age–time combinations were presented using different types of stimuli in the image-cued and language-cued groups (Fig. 1c).

Nevigation task. Participants were presented with a food item and asked to identify the corresponding time of day and age group. In each trial, a target food item was shown at the top of the screen. Participants selected the correct age group and time of day from all graded levels. Feedback was provided immediately after each selection, and a trial was recorded as correct only if both the time and age choices were correct on the first attempt. Training continued until participants reached 12 consecutive correct trials. Again, all possible options covering the full grading range were presented using different stimulus formats in the image-cued and language-cued conditions (Fig. 1d). Performance accuracy in the above tasks did not differ significantly between the two conditions (see Supplementary Fig. A1).

Testing phase

Similarity rating task. Participants judged the similarity of food-item pairs based solely on the learned graded variables (Fig. 1e). In each trial, two food images appeared side by side, and participants rated their similarity from 0 (not similar) to 10 (highly similar) using a mouse-controlled slider with a random starting position. The task comprised 60 trials. Participants were explicitly instructed to judge the similarity between pairs of food items based solely on who ate them and when they were eaten, and not on any other attributes such as taste, color, or prior pairings.

Choice task. Participants then completed a choice task in which unknown rewards depended on each item's location within an unshown 2D map defined by the learned graded variables (Fig. 1f). Participants were given a cover story stating that the price of each food item depended on both the eater's age group and the time of day at which it was eaten, and that the way in which these two factors affect price differs between the two restaurants, Tree and River. Unbeknownst to participants, the two restaurants followed different location–value mappings (Fig. 1g), and the specific mapping rules were not disclosed. The exact pricing computation is provided in the caption of (Fig. 1). Note that each price included a ± 2 noise term. The assignment between restaurant names (Tree, River) and the corresponding value maps was randomized across participants.

The task consisted of blocks associated with each of the two restaurants. At the beginning of each block, participants were informed which restaurant they were in, and a Tree or River symbol remained at the top of the screen throughout that block. In each trial, two food items appeared side by side, and participants selected one using the left or right arrow key on a keyboard. Only the value of the chosen item was revealed after making a choice. Participants were instructed to maximize their rewards. The task consisted of four blocks (two for each restaurant) with 30 trials per block, resulting in 120 trials in total.

At the end of the choice task, participants were asked to estimate the values of all food items for each restaurant based on what they had experienced during the task. Unbeknownst to participants, the estimation phase included one additional food item per restaurant (i.e., the question mark in Fig. 1g) that had never been presented during the choice task. This manipulation allowed us to assess the model's ability to generalize value estimates to untrained items.

Placement task. After the choice task, participants were asked to place the food items onto a 2D map defined by age group and time of day. The grid was structured by age (youngest to oldest; vertical axis) and time of day (earliest to latest; horizontal axis), as shown in Fig. 1h. All food items initially appeared outside the grid and participants used the mouse to drag and drop each item onto the location corresponding to its previously learned age–time association.

Post-experiment assessment

After the experiment, participants completed a brief questionnaire assessing whether they experienced any uncertainties during the task and whether such uncertainties

influenced their choices. They also performed two additional cognitive tasks: an N-back task [36] to index working memory capacity and a mental rotation task [37] to assess spatial transformation ability, each lasting approximately 7–8 minutes. In the final session, participants completed an audio recording task; these data were not included in the present analyses.

Behavioral analysis

As specified in the pre-registration, similarity ratings were analyzed using a linear mixed-effects model (LMM) with similarity as the dependent variable. Fixed effects were group (image vs. language), 2D distance, and their interaction, with random intercepts for participants. The model was fitted using maximum likelihood estimation (default) in MATLAB (`fitlme`). All participants were included in this analysis.

Computational modeling

As specified in the pre-registration, we compared four computational models with different hypotheses about decision-making during the choice task to examine how participants learned and generalized rewards across contexts. The data from all participants were included in this analysis.

Model 1: Baseline model

The baseline model assumes that participants make decisions with a fixed bias toward the left option, independent of previous rewards or option values. Choices were modeled with a single free parameter, p, denoting the probability of choosing the left option:

$$P(a_t = \text{left}) = p, \quad P(a_t = \text{right}) = 1 - p \tag{1}$$

Model 2: Q-learning model

This model assumes that participants make decisions based on option values that are updated using a standard model-free reinforcement learning algorithm. Option values $Q(o)$ are updated after each trial t based on the received reward r_t:

$$Q(o) \leftarrow Q(o) + \alpha(r_t - Q(o)) \tag{2}$$

where $\alpha \in (0, 1)$ is the learning rate. Choices follow a softmax rule:

$$P(a_t = o_1) = \frac{1}{1 + \exp[\beta(Q(o_2) - Q(o_1))]} \tag{3}$$

where β is the inverse temperature controlling choice stochasticity. Two free parameters were estimated: α and β.

Model 3: Spatial gaussian process regression(GPR) model

This model assumes that participants generalize experienced rewards across the 2D spatial layout of the map using Gaussian process regression (GPR). Each map

context is modeled independently, maintaining its own history of chosen locations $X = \{x_1, \ldots, x_n\}$ and corresponding rewards $Y = (r_1, \ldots, r_n)^\top$. The covariance between any two past choices is computed via a radial basis function (RBF) kernel:

$$k(x, x') = \exp\Big(- \frac{\|x - x'\|^2}{2\lambda^2} \Big) \tag{4}$$

where λ is the lengthscale parameter governing the extent to which input similarity decreases with distance, reflecting the weight assigned to spatial information in the generalization process. Larger λ values produce broader generalization across space. For each candidate option x, the predicted mean reward is

$$\mu(x) = k(x, X)[K + 10^{-3} I]^{-1} Y \tag{5}$$

where $K = k(X, X)$ denotes the kernel matrix whose (i, j)-th entry is given by $K_{ij} = k(x_i, x_j)$. A small jitter term $10^{-3} I$ is added to the kernel matrix to ensure numerical stability when computing its inverse. Choices follow a softmax over the predicted means:

$$P(a_t = o_1) = \frac{1}{1 + \exp[\beta(\mu(o_2) - \mu(o_1))]} \tag{6}$$

Free parameters: λ (spatial lengthscale) and β (inverse temperature).

Model 4: Hybrid model

The hybrid model assumes that participants combine model-free Q-learning and model-based spatial generalization. For each option o, the hybrid value is a weighted combination of Q-learning and GPR-based predictions:

$$V(o) = wQ(o) + (1 - w)\mu(o) \tag{7}$$

where $w \in [0, 1]$ determines the relative reliance on Q-values versus GP predictions. Choices are generated via a softmax:

$$P(a_t = o_1) = \frac{1}{1 + \exp[\beta(V(o_2) - V(o_1))]} \tag{8}$$

The Q-values are updated after each choice using the same rule as in Model 2, and the GP memory is updated with the chosen location and reward. The free parameters are α (learning rate), β (inverse temperature), λ (spatial lengthscale), and w (Q/GP weight).

Model fitting and comparison

Model parameters were estimated for each participant by maximizing the log-likelihood of observed choices given the model. For each participant and model, the log-likelihood is computed as

$$\mathcal{L} = \sum_{t=1}^{T} \log P(a_t \,|\, \theta) \tag{9}$$

where $P(a_t \mid \theta)$ is the model-predicted probability of the observed choice on trial t, given parameters θ. To reduce the risk of local minima, parameter estimation was performed in two independent runs, each using 1000 random starting points within a global optimization routine (`GlobalSearch` in MATLAB). For each subject, the best-fitting parameters were defined as those yielding the maximum log-likelihood across all runs and starting points.

Model comparison was performed using the corrected Akaike information criterion (AICc, [38]), which evaluates relative model fit while penalizing model complexity and adjusting for finite sample sizes. For a model with K free parameters, sample size n, and maximized log-likelihood $\log L$, AICc is given by

$$\mathrm{AICc} = -2 \log L + \frac{2K(K+1)}{n-K-1} \tag{10}$$

Lower AICc values indicate better trade-offs between model fit and complexity.

Model validation

Parameter recovery

To assess parameter identifiability, we conducted a parameter recovery analysis. Synthetic choice data were generated for a subset of participants by simulating behavior from the model using their individually fitted parameters. The model was then refitted to these simulated data using the same optimization procedure as in the original analysis. For each parameter, recovered values were strongly correlated with the true generating values across participants (Supplementary Fig. A2), indicating reliable recoverability.

Model prediction

To evaluate the model's predictive adequacy, we examined whether the fitted model could reproduce the trajectory of subjective value estimates observed in participants' behavior. Using each participant's fitted parameters, we computed trial-by-trial model-predicted expected values for the chosen option and compared these predictions with the corresponding true outcome values experienced by participants. Predicted and observed values were averaged across participants at each trial, separately for each map and cue condition, and uncertainty was quantified using 95% confidence intervals reflecting between-subject variability. Across both image-cued and language-cued conditions, the model closely tracked the evolution of experienced values over learning (Fig. 4b-c), indicating that it captured the dynamics of value updating and choice-relevant expectations.

Value generalization

To examine whether the model captured value generalization beyond directly experienced outcomes, we compared participants' explicit value ratings for individual items with the model-predicted values derived from the inferred spatial map. For each participant, we used the individually estimated parameters of the spatial-GPR model to

compute predicted expected values for all items in each map based on each participant's history of experienced rewards. These model-predicted values were then aligned with participants' rated values and compared at the item level.

We next examined value generalization at the individual level. For each participant, explicit item ratings were compared with model-predicted values derived from the fitted model, yielding participant-specific correspondence patterns. The strength of this correspondence depended on which model best accounted for the participant's choice behavior: participants best fit by the spatial-GPR model (M3) or the hybrid model (M4) showed close alignment between rated values and model predictions, whereas this relationship was weaker for participants best fit by baseline or non-spatial learning models (M1–M2) (see Supplementary Fig. A5). Quantifying prediction error specifically for untrained items revealed that participants whose behavior was best explained by the spatial-GPR model exhibited the smallest deviations between rated and predicted values (Supplementary Fig. A5), providing direct evidence that value generalization relied on the inferred spatial structure.

References

[1] Tolman, E.C.: Cognitive maps in rats and men. Psychological review **55**(4), 189 (1948) https://doi.org/10.1037/h0061626

[2] Doeller, C.F., Barry, C., Burgess, N.: Evidence for grid cells in a human memory network. Nature **463**(7281), 657–661 (2010) https://doi.org/10.1038/nature08704

[3] Garvert, M.M., Dolan, R.J., Behrens, T.E.: A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. elife **6**, 17086 (2017) https://doi.org/10.7554/eLife.17086

[4] Bellmund, J.L., Gärdenfors, P., Moser, E.I., Doeller, C.F.: Navigating cognition: Spatial codes for human thinking. Science **362**(6415), 6766 (2018) https://doi.org/10.1126/science.aat6766

[5] Theves, S., Schäfer, T.A., Reisner, V., Cothi, W., Barry, C.: Category boundaries modulate memory in a place-cell-like manner. Current Biology **34**(23), 5546–5553 (2024) https://doi.org/10.1016/j.cub.2024.09.083

[6] Schuck, N.W., Cai, M.B., Wilson, R.C., Niv, Y.: Human orbitofrontal cortex represents a cognitive map of state space. Neuron **91**(6), 1402–1412 (2016) https://doi.org/10.1016/j.neuron.2016.08.019

[7] Baram, A.B., Muller, T.H., Nili, H., Garvert, M.M., Behrens, T.E.J.: Entorhinal and ventromedial prefrontal cortices abstract and generalize the structure of reinforcement learning problems. Neuron **109**(4), 713–723 (2021) https://doi.org/10.1016/j.neuron.2020.11.024

[8] Garvert, M.M., Saanum, T., Schulz, E., Schuck, N.W., Doeller, C.F.: Hippocampal spatio-predictive cognitive maps adaptively guide reward generalization. Nature Neuroscience **26**(4), 615–626 (2023) https://doi.org/10.1038/s41593-023-01283-x

[9] Veselic, S., Muller, T.H., Gutierrez, E., Behrens, T.E., Hunt, L.T., Butler, J.L., Kennerley, S.W.: A cognitive map for value-guided choice in the ventromedial prefrontal cortex. Cell **188**(12), 3259–3273 (2025) https://doi.org/10.1016/j.cell.2025.03.038

[10] Mason, O., Claridge, G.: The oxford-liverpool inventory of feelings and experiences (o-life): further description and extended norms. Schizophrenia research **82**(2-3), 203–211 (2006) https://doi.org/10.1016/j.schres.2005.12.845

[11] Kuperberg, G.R.: Language in schizophrenia part 1: An introduction. Language and Linguistics Compass **4**(8), 576–589 (2010) https://doi.org/10.1111/j.1749-818X.2010.00216.x

[12] Nour, M.M., Liu, Y., El-Gaby, M., McCutcheon, R.A., Dolan, R.J.: Cognitive maps and schizophrenia. Trends in Cognitive Sciences **29**(2), 184–200 (2025) https://doi.org/10.1016/j.tics.2024.09.011

[13] Musa, A., Khan, S., Mujahid, M., El-Gaby, M.: The shallow cognitive map hypothesis: A hippocampal framework for thought disorder in schizophrenia. Schizophrenia **8**(1), 34 (2022) https://doi.org/10.1038/s41537-022-00247-7

[14] Nour, M.M., McNamee, D.C., Liu, Y., Dolan, R.J.: Trajectories through semantic spaces in schizophrenia and the relationship to ripple bursts. Proceedings of the National Academy of Sciences **120**(42), 2305290120 (2023) https://doi.org/10.1073/pnas.2305290120

[15] Convertino, L., Bush, D., Zheng, F., Adams, R.A., Burgess, N.: Reduced grid-like theta modulation in schizophrenia. Brain **146**(5), 2191–2198 (2023) https://doi.org/10.1093/brain/awac416

[16] Constantinescu, A.O., O'Reilly, J.X., Behrens, T.E.: Organizing conceptual knowledge in humans with a gridlike code. Science **352**(6292), 1464–1468 (2016) https://doi.org/10.1126/science.aaf0941

[17] Theves, S., Fernandez, G., Doeller, C.F.: The hippocampus encodes distances in multidimensional feature space. Current Biology **29**(7), 1226–1231 (2019) https://doi.org/10.1016/j.cub.2019.02.035

[18] Viganò, S., Fornari, E., Sulpizio, S., al.: Grid-like and distance codes for representing word meaning in the human brain. NeuroImage **236**, 118085 (2021) https://doi.org/10.1016/j.neuroimage.2021.118085

[19] Viganò, S., Piazza, M.: Distance and direction codes underlie navigation of a novel semantic space in the human brain. Journal of Neuroscience **40**(13), 2727–2736 (2020) https://doi.org/10.1523/JNEUROSCI.1849-19.2020

[20] Mason, O., Linney, Y., Claridge, G.: Short scales for measuring schizotypy. Schizophrenia Research **17**(1), 65–72 (1995) https://doi.org/10.1016/0920-9964(95)00031-3

[21] Zheng, X.Y., Hebart, M.N., Grill, F., Dolan, R.J., Doeller, C.F., Cools, R., Garvert, M.M.: Parallel cognitive maps for multiple knowledge structures in the hippocampal formation. Cerebral Cortex **34**(2), 485 (2024) https://doi.org/10.1093/cercor/bhad485

[22] Matsumoto, Y., Nishida, S., Hayashi, R., Son, S., Murakami, A., Yoshikawa, N., Ito, H., Oishi, N., Masuda, N., Murai, T., *et al.*: Disorganization of semantic brain networks in schizophrenia revealed by fmri. Schizophrenia Bulletin **49**(2), 498–506 (2023) https://doi.org/10.1093/schbul/sbac157

[23] Karagoz, A.B., Moran, E.K., Barch, D.M., Kool, W., Reagh, Z.M.: Evidence for shallow cognitive maps in schizophrenia. Cognitive, Affective, & Behavioral Neuroscience, 1–18 (2025) https://doi.org/10.3758/s13415-025-01283-3

[24] Omlor, W., Cecere, G., Misra, A.R., Huang, G.-Y., Wu, X., Edkins, V., Pergola, G., Quednow, B., Tobler, P., Homan, P.: Goal-directed control in schizophrenia: Loss-biased engagement of the anterior cingulate relates to negative-symptom outcomes. medRxiv, 2025–07 (2025) https://doi.org/10.1101/2025.07.02.25330758

[25] Wu, C.M., Schulz, E., Garvert, M.M., Meder, B., Schuck, N.W.: Similarities and differences in spatial and non-spatial cognitive maps. PLoS computational biology **16**(9), 1008149 (2020) https://doi.org/10.1371/journal.pcbi.1008149

[26] Friederici, A.D.: The brain basis of language processing: from structure to function. Physiological reviews **91**(4), 1357–1392 (2011) https://doi.org/10.1152/physrev.00006.2011

[27] Palominos, C., He, R., Fröhlich, K., Mülfarth, R.R., Seuffert, S., Sommer, I.E., Homan, P., Kircher, T., Stein, F., Hinzen, W.: Approximating the semantic space: word embedding techniques in psychiatric speech analysis. Schizophrenia **10**(1), 1–10 (2024) https://doi.org/10.1038/s41537-024-00524-7

[28] Palominos, C., Kirdun, M., Nikzad, A.H., Spilka, M.J., Homan, P., Sommer, I.E., Tang, S.X., Hinzen, W.: A single composite index of semantic behavior tracks symptoms of psychosis over time. Schizophrenia Research **279**, 116–127 (2025) https://doi.org/10.1016/j.schres.2025.03.038

[29] Kircher, T., Krug, A., Stratmann, M., Ghazi, S., Schales, C., Frauenheim, M.,

Turner, L., Fährmann, P., Hornig, T., Katzev, M., *et al.*: A rating scale for the assessment of objective and subjective formal thought and language disorder (tald). Schizophrenia research **160**(1-3), 216–221 (2014) https://doi.org/10.1016/j.schres.2014.10.024

[30] Bozhilova, N.S., Michelini, G., Kuntsi, J., Asherson, P.: Mind wandering perspective on attention-deficit/hyperactivity disorder. Neuroscience & Biobehavioral Reviews **92**, 464–476 (2018) https://doi.org/10.1016/j.neubiorev.2018.07.010

[31] Fradkin, I., Adams, R.A., Siegelman, N., Moran, R., Dolan, R.J.: Latent mechanisms of language disorganization relate to specific dimensions of psychopathology. Nature Mental Health **2**(12), 1486–1497 (2024) https://doi.org/10.1038/s44220-024-00351-w

[32] Viganò, S., Rubino, V., Di Soccio, A., Buiatti, M., Piazza, M.: Grid-like and distance codes for representing word meaning in the human brain. NeuroImage **232**, 117876 (2021) https://doi.org/10.1016/j.neuroimage.2021.117876

[33] Shinkareva, S.V., Malave, V.L., Mason, R.A., Mitchell, T.M., Just, M.A.: Commonality of neural representations of words and pictures. Neuroimage **54**(3), 2418–2425 (2011) https://doi.org/10.1016/j.neuroimage.2010.10.042

[34] McCrae, R.R., Costa, P.T.J.: A five-factor theory of personality. In: Pervin, L.A., John, O.P. (eds.) Handbook of Personality: Theory and Research, pp. 139–153. Guilford Press, New York (1999)

[35] Watson, D., Clark, L.A., Tellegen, A.: Development and validation of brief measures of positive and negative affect: The panas scales. Journal of Personality and Social Psychology **54**(6), 1063–1070 (1988) https://doi.org/10.1037/0022-3514.54.6.1063

[36] Colwell, M., Tagomori, H., Martens, M., Murphy, S., Harmer, C.: N-Back (Oxford) – fMRI and Non-scanner Release (Python). https://doi.org/10.5281/zenodo.8003407 . https://doi.org10.5281/zenodo.8003407

[37] Shepard, R.N., Metzler, J.: Mental rotation of three-dimensional objects. Science **171**(3972), 701–703 (1971) https://doi.org/10.1126/science.171.3972.701

[38] Hurvich, C.M., Tsai, C.-L.: Regression and time series model selection in small samples. Biometrika **76**(2), 297–307 (1989) https://doi.org/10.1093/biomet/76.2.297

[39] John, O.P., Srivastava, S.: In: Pervin, L.A., John, O.P. (eds.) The Big Five Trait taxonomy: History, measurement, and theoretical perspectives, pp. 102–138 (1999)

Appendix A Supplementary Information

A.1 Supplementary Tables

Table A1 Demographic and psychometric measures for the image and language groups

Measurements	Image-cued ($N = 51$)	Language-cued ($N = 60$)	Test Statistic	p value
Age (years)	26.54 ± 4.66	27.63 ± 6.81	$t = -0.96$	$p = 0.340$
Gender (M/F)	18/31	17/40	$\chi^2 = 3.46$	$p = 0.629$
Years of Education	16.39 ± 1.79	16.68 ± 2.47	$t = -0.70$	$p = 0.763$
Hungry Level	3.09 ± 1.60	3.56 ± 1.65	$t = -1.49$	$p = 0.139$
Minutes Since Last Meal (min)	234.86 ± 170.26	292.36 ± 279.90	$t = -1.28$	$p = 0.204$
O-LIFE total	11.75 ± 6.21	13.33 ± 6.85	$t = -1.27$	$p = 0.207$
O-LIFE Unusual Experiences	3.14 ± 2.48	3.43 ± 2.87	$t = -0.58$	$p = 0.566$
O-LIFE Cognitive Disorganization	4.14 ± 3.09	4.43 ± 3.04	$t = -0.51$	$p = 0.613$
O-LIFE Introvertive Anhedonia	2.67 ± 2.02	2.63 ± 2.09	$t = 0.09$	$p = 0.932$
O-LIFE Impulsive Nonconformity	1.80 ± 1.36	2.83 ± 1.91	$t = -3.22$	$\mathbf{p = 0.002}$
BFI-Neuroticism	23.02 ± 5.64	23.70 ± 6.48	$t = -0.58$	$p = 0.560$
BFI-Extraversion	25.53 ± 4.95	25.80 ± 5.70	$t = -0.26$	$p = 0.792$
BFI-Openness	35.94 ± 5.96	38.43 ± 5.53	$t = -2.28$	$\mathbf{p = 0.024}$
BFI-Agreeableness	31.08 ± 3.47	30.55 ± 4.15	$t = 0.72$	$p = 0.473$
BFI-Conscientiousness	30.10 ± 5.67	29.47 ± 5.55	$t = 0.59$	$p = 0.556$
Positive Affect	32.10 ± 5.84	31.53 ± 6.42	$t = 0.48$	$p = 0.631$
Negative Affect	21.94 ± 6.97	20.58 ± 6.91	$t = 1.03$	$p = 0.306$
Working memory (N-back task)	0.89 ± 0.08	0.90 ± 0.06	$t = -0.32$	$p = 0.746$
Mental rotation task	23.82 ± 7.88	25.72 ± 7.84	$t = -1.26$	$p = 0.209$

Note: Values are presented as mean $\pm$ standard deviation. BFI refers to the Big Five Inventory [39]; O-LIFE refers to the Oxford–Liverpool Inventory of Feelings and Experiences [20]. Positive Affect and Negative Affect were assessed using the PANAS scales [35]. Working memory was assessed using an N-back task [36], and spatial transformation ability was assessed using a mental rotation task [37]. Bolded p values indicate statistical significance ($p < 0.05$). All t tests were two-tailed independent-samples tests.

A.2 Supplementary Figures

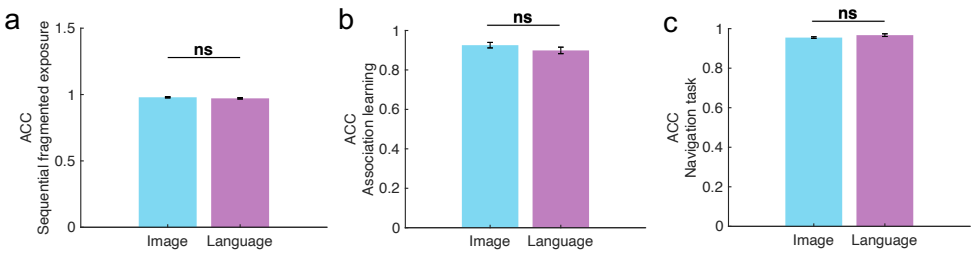

Fig. A1 Group comparison of training accuracy. a. Accuracy in the sequential fragmented exposure of the ordering task. **b.** Accuracy in the association learning task. **c.** Accuracy in the navigation task. Bars show group means $\pm$ SEM. Independent-samples t-tests revealed no significant group differences (sequential fragmented exposure: $t(109) = 1.25$, $p = 0.215$; association learning: $t(109) = 1.21$, $p = 0.228$; navigation task: $t(109) = -1.50$, $p = 0.137$). *ns*, not significant.

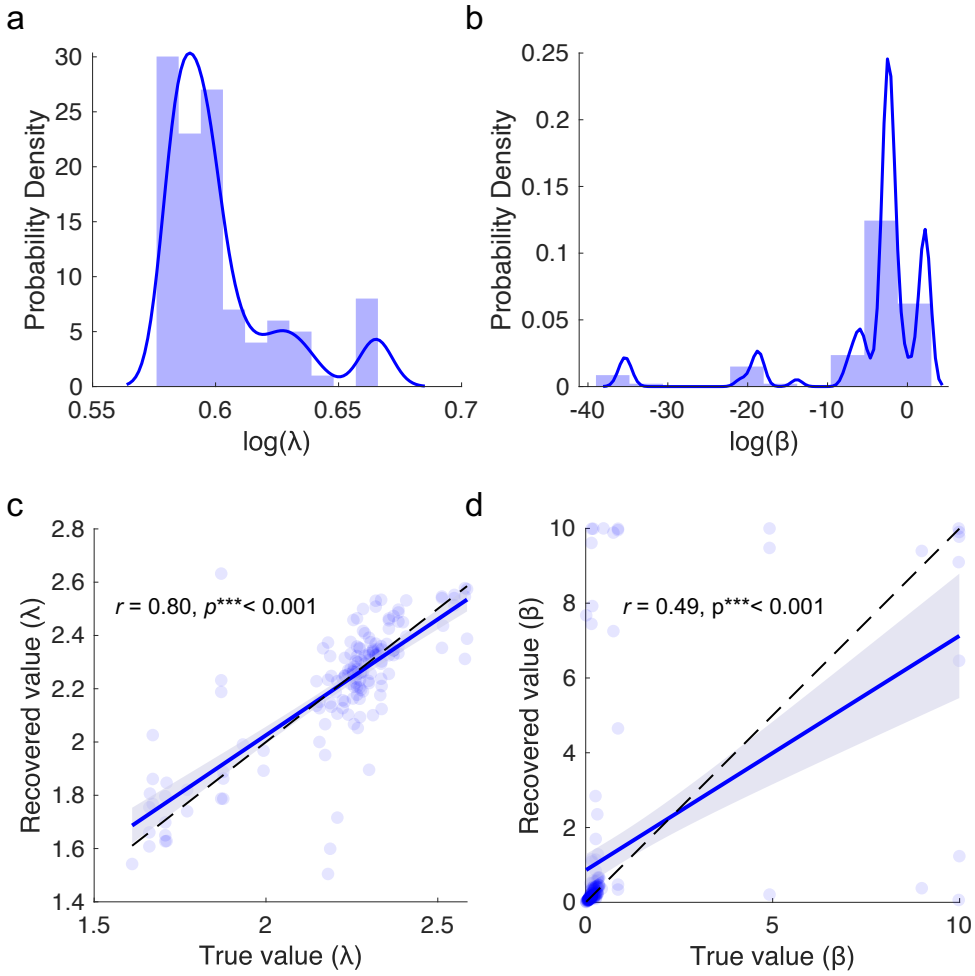

Fig. A2 **a–b.** Distributions of parameter estimates from the best-fitting model (Model 3: Spatial-GPR) across all participants. Panels display histograms with kernel density estimates for the log-transformed parameters: (a) $\log(\lambda)$ (length-scale) and (b) $\log(\beta)$ (inverse temperature). **c–d.** Parameter recovery results for the best-fitting model (Model 3: Spatial-GPR). Panels show the relationship between true and recovered parameter values across simulated datasets for (c) the length-scale parameter (λ) and (d) the inverse temperature parameter (β). Blue lines indicate linear fits with 95% confidence intervals, and dashed diagonal lines represent perfect recovery. Reported r values correspond to the Pearson correlations between true and recovered parameter estimates. Significance levels: $^{***}p < 0.001$.

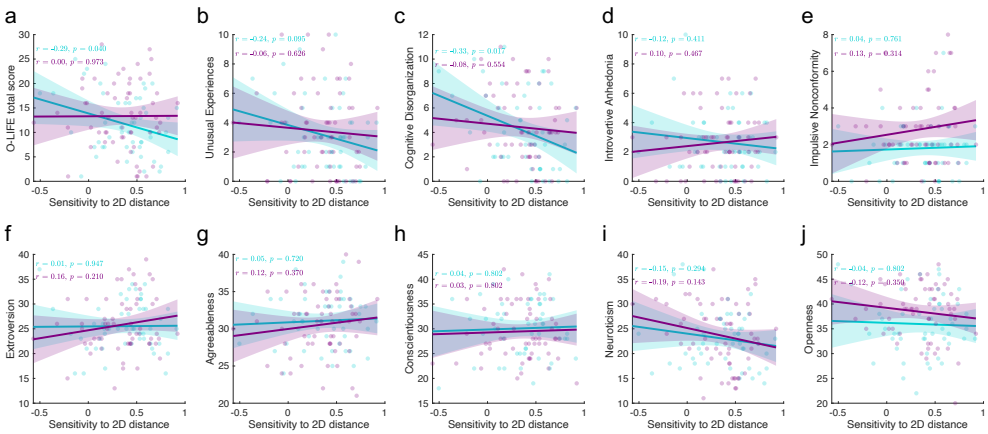

Fig. A3 a–j. Correlations between spatial sensitivity and individual difference measures. Spatial sensitivity (x-axis) reflects the Fisher Z-transformed correlation between pairwise 2D distances and dissimilarity scores, with larger values indicating stronger sensitivity to spatial structure. The first row (a–e) shows correlations with schizotypy traits measured by the O-LIFE measurement, and the second row (f–j) shows correlations with the Big Five personality traits (control analysis). Significant associations were observed only for schizotypy measures, indicating that spatial sensitivity was selectively related to schizotypy rather than general personality dimensions. Cyan lines denote linear fits for the image-cued group and purple lines for the language-cued group, with shaded areas representing 95% confidence intervals. Reported r values correspond to Pearson correlations.

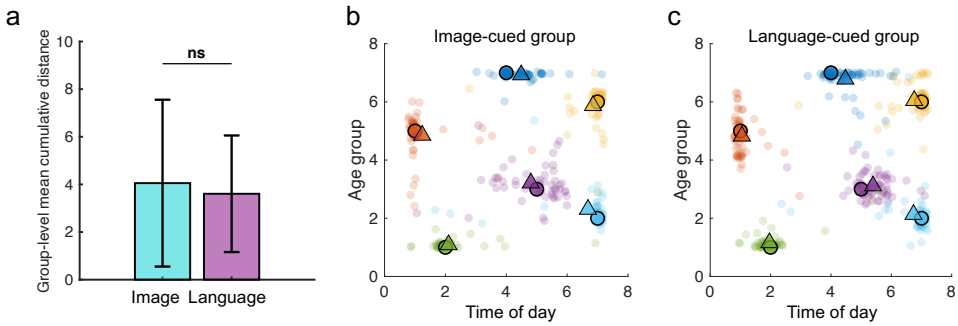

Fig. A4 Results of the placement task. a. Group-level mean cumulative distance between participants' estimated item locations and the true locations in the 2D space, shown separately for the image-cued and language-cued groups. Bars depict group means ± SEM. No significant difference was observed between the two groups. **b.** Comparison of the true locations and participants' estimated locations for the food items within the 2D map in the image-cued group. Circles indicate the true item positions, and triangles indicate the participants' average estimated positions. Each transparent colored dot represents one participant's estimated location, with different colors denoting different food items. The solid arrow illustrates the 2D placement error for one participant in one example food item, defined as the Euclidean distance between the true and estimated item locations (lower values indicate greater accuracy). **c.** Same format as panel b, but for the language-cued group. *ns*, not significant.

subj 1, M1 baseline
subj 2, M1 baseline
subj 3, M1 baseline
subj 4, M1 baseline
subj 5, M1 baseline
subj 6, M1 baseline
subj 7, M1 baseline
subj 8, M1 baseline
subj 9, M1 baseline
subj 10, M1 baseline
subj 11, M1 baseline
subj 12, M1 baseline
subj 13, M1 baseline
subj 14, M1 baseline
subj 15, M1 baseline
subj 16, M1 baseline
subj 17, M1 baseline
subj 18, M1 baseline
subj 19, M1 baseline
subj 20, M1 baseline
subj 21, M1 baseline
subj 22, M1 baseline
subj 23, M1 baseline
subj 24, M1 baseline
subj 25, M1 baseline
subj 26, M1 baseline
subj 27, M1 baseline
subj 28, M1 baseline
subj 29, M1 baseline
subj 30, M1 baseline
subj 31, M1 baseline
subj 32, M1 baseline
subj 33, M1 baseline
subj 34, M1 baseline
subj 35, M1 baseline
subj 36, M1 baseline
subj 37, M1 baseline
subj 38, M1 baseline
subj 39, M1 baseline
subj 40, M1 baseline
subj 41, M1 baseline
subj 42, M2 Q-leaning
subj 43, M2 Q-leaning
subj 44, M2 Q-leaning
subj 45, M2 Q-leaning
subj 46, M2 Q-leaning
subj 47, M2 Q-leaning
subj 48, M2 Q-leaning
subj 49, M2 Q-leaning
subj 50, M2 Q-leaning
subj 51, M2 Q-leaning
subj 52, M2 Q-leaning
subj 53, M3 spatial-GPR
subj 54, M3 spatial-GPR
subj 55, M3 spatial-GPR
subj 56, M3 spatial-GPR
subj 57, M3 spatial-GPR
subj 58, M3 spatial-GPR
subj 59, M3 spatial-GPR
subj 60, M3 spatial-GPR
subj 61, M3 spatial-GPR
subj 62, M3 spatial-GPR
subj 63, M3 spatial-GPR
subj 64, M3 spatial-GPR
subj 65, M3 spatial-GPR
subj 66, M3 spatial-GPR
subj 67, M3 spatial-GPR
subj 68, M3 spatial-GPR
subj 69, M3 spatial-GPR
subj 70, M3 spatial-GPR
subj 71, M3 spatial-GPR
subj 72, M3 spatial-GPR
subj 73, M3 spatial-GPR
subj 74, M3 spatial-GPR
subj 75, M3 spatial-GPR
subj 76, M3 spatial-GPR
subj 77, M3 spatial-GPR
subj 78, M3 spatial-GPR
subj 79, M3 spatial-GPR
subj 80, M3 spatial-GPR
subj 81, M3 spatial-GPR
subj 82, M3 spatial-GPR
subj 83, M3 spatial-GPR
subj 84, M3 spatial-GPR
subj 85, M3 spatial-GPR
subj 86, M3 spatial-GPR
subj 87, M3 spatial-GPR
subj 88, M3 spatial-GPR
subj 89, M3 spatial-GPR
subj 90, M3 spatial-GPR
subj 91, M3 spatial-GPR
subj 92, M3 spatial-GPR
subj 93, M3 spatial-GPR
subj 94, M3 spatial-GPR
subj 95, M3 spatial-GPR
subj 96, M3 spatial-GPR
subj 97, M4 hybrid
subj 98, M4 hybrid
subj 99, M4 hybrid
subj 100, M4 hybrid
subj 101, M4 hybrid
subj 102, M4 hybrid
subj 103, M4 hybrid
subj 104, M4 hybrid
subj 105, M4 hybrid
subj 106, M4 hybrid
subj 107, M4 hybrid
subj 108, M4 hybrid
subj 109, M4 hybrid
subj 110, M4 hybrid
subj 111, M4 hybrid
model prediction
rated value
Prediction error for unseen subjects
M1 baseline
M2 Q-learning
M3 spatial-GPR
M4 hybrid
Individual best model

Fig. A5 Participant-level model predictions of item values. Each panel shows the relationship between participants' rated item values (x-axis) and model-predicted values (y-axis), computed from the group level best-fitting model (Model 3: Spatial-GRP) using participant-specific parameters as inputs. Panel titles indicate participant ID and their best-fitting model. Blue dots represent items learned during the choice task, and red dots indicate unobserved items that required value generalization (corresponding to question-mark positions in Fig.1g). Green lines denote linear fits with shaded 95% confidence intervals, and the statistics report Pearson correlation coefficients (r, p). Across participants, the Spatial-GPR (Model 3) and Hybrid (Model 4) models showed stronger correspondence between predicted and rated values compared with the Baseline and Q-learning models, capturing both learned values and generalized value inference for unobserved items.

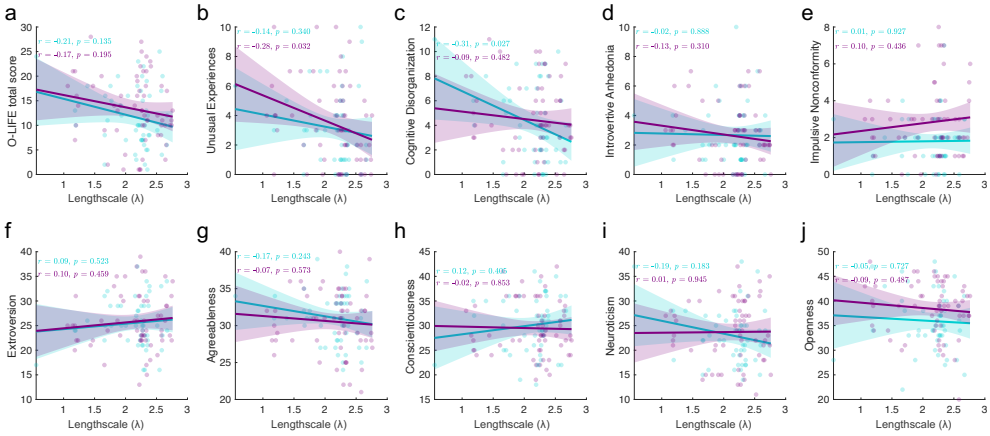

Fig. A6 a–j. Correlations between the lengthscale parameter (λ) from the best-fitting model and individual difference measures. The lengthscale (λ; x-axis) reflects the degree of spatial generalization in the model (larger values indicate broader generalization across space). The first row (a–e) shows associations with schizotypy traits measured by the O-LIFE questionnaire, and the second row (f–j) shows associations with the Big Five personality traits (control analysis). Significant correlations were observed only for schizotypy subscales, suggesting that spatial generalization in the model was selectively related to schizotypy rather than general personality dimensions. Cyan and purple lines indicate the image-cued and language-cued groups, respectively, with shaded areas representing 95% confidence intervals. Reported r values denote Pearson correlations.

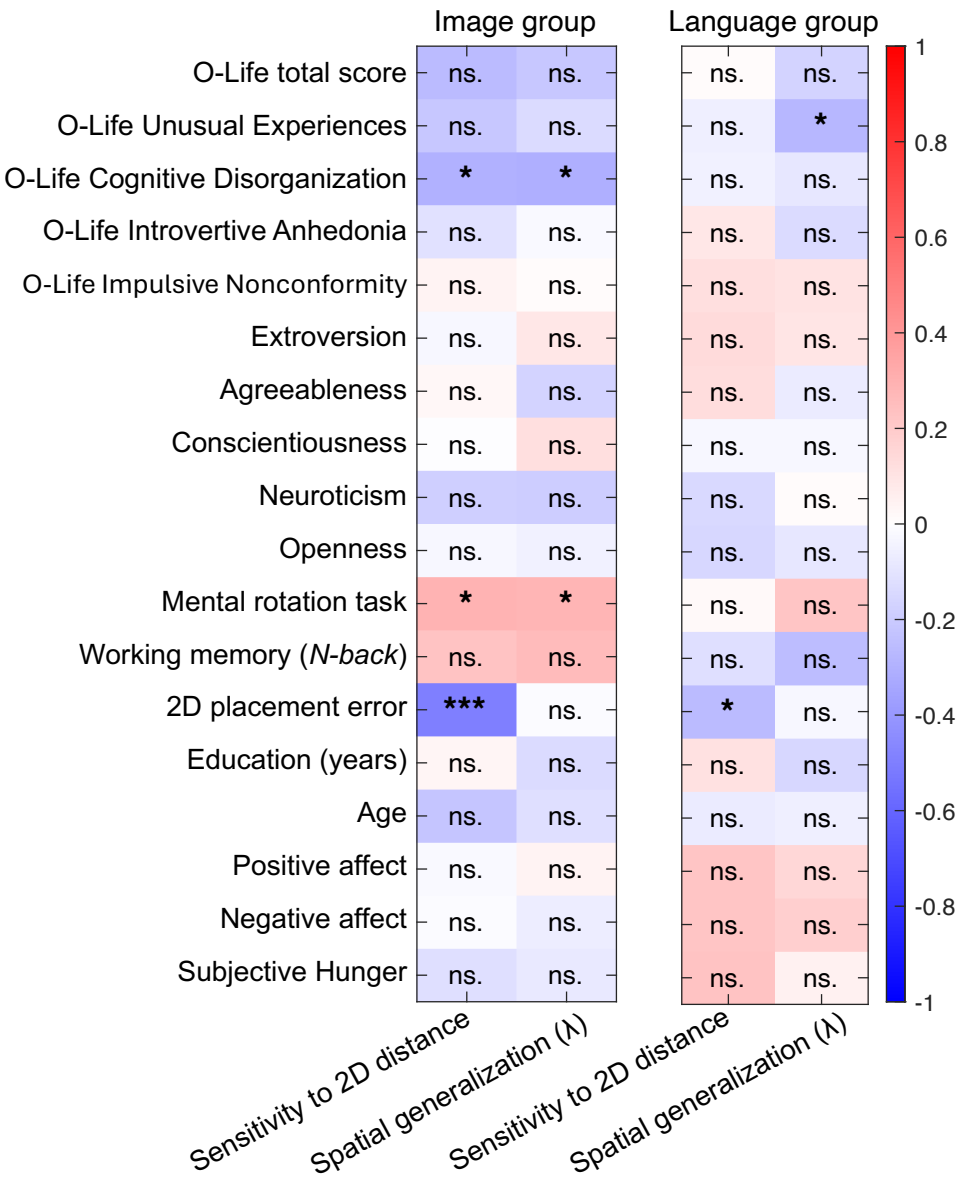

Fig. A7 Correlation matrix of psychological and behavioral measures with spatial sensitivity and generalization. The heatmap displays Pearson correlations between individual differences (including O-LIFE subscales, Big Five personality traits, cognitive performance measures, demographic variables, and affective ratings) and two key behavioral indices: spatial sensitivity and generalization strength (γ). Each cell shows the correlation coefficient, color-coded from negative (blue) to positive (red) values, as indicated by the color bar on the right. Significance levels are denoted within each cell: ns = not significant; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.