Research Article

Psychological Science
2022, Vol. 33(9) 1463–1476

DOI: 10.1177/09567976221091838
www.psychologicalscience.org/PS

Auxiliary Scene-Context Information Provided by Anchor Objects Guides Attention and Locomotion in Natural Search Behavior

Jason Helbing[1], Dejan Draschkow[2,3], and Melissa L.-H. Võ[1]

[1]Scene Grammar Lab, Department of Psychology, Goethe University Frankfurt; [2]Brain and Cognition Laboratory, Department of Experimental Psychology, University of Oxford; and [3]Oxford Centre for Human Brain Activity, Wellcome Centre for Integrative Neuroimaging, Department of Psychiatry, University of Oxford

Abstract

Successful adaptive behavior requires efficient attentional and locomotive systems. Previous research has thoroughly investigated how we achieve this efficiency during natural behavior by exploiting prior knowledge related to targets of our actions (e.g., attending to metallic targets when looking for a pot) and to the environmental context (e.g., looking for the pot in the kitchen). Less is known about whether and how individual nontarget components of the environment support natural behavior. In our immersive virtual reality task, 24 adult participants searched for objects in naturalistic scenes in which we manipulated the presence and arrangement of large, static objects that anchor predictions about targets (e.g., the sink provides a prediction for the location of the soap). Our results show that gaze and body movements in this naturalistic setting are strongly guided by these anchors. These findings demonstrate that objects auxiliary to the target are incorporated into the representations guiding attention and locomotion.

Keywords

attention, locomotion, natural behavior, contextual guidance, anchor objects, scene grammar, virtual reality, open data

Received 8/15/21; Revision accepted 2/21/22

Investigating what guides attention and action in real-world settings is essential to understanding natural human behaviors (Ballard et al., 1995; Draschkow et al., 2021; Foulsham et al., 2011; Hayhoe & Ballard, 2005; Á. Kristjánsson & Draschkow, 2021; Tatler et al., 2011). Past research has extensively studied how behavior is guided by (a) the properties of the targets of our goals (Wolfe, 2020, 2021; Wolfe & Horowitz, 2017) and (b) the global environmental context (Hutchinson & Turk-Browne, 2012; Neider & Zelinsky, 2006; Torralba et al., 2006; Wolfe, Võ, et al., 2011). For example, the color yellow is a key target feature in the process of searching for both bananas and tennis balls, yet we are much more likely to identify a yellow object in a kitchen as a banana (Bar, 2004; Davenport & Potter, 2004; Lauer et al., 2018, 2021).

It remains unclear whether and how aspects of our environment that are neither properties of the target itself nor low-level global contextual cues (such as summary statistics; Brady et al., 2017; Greene & Oliva, 2009) influence behavioral guidance. After all, our surroundings are not random compositions of arbitrary parts but comprise a multitude of stand-alone objects that are connected by high-level environmental regularities (Greene, 2013; Mack & Eckstein, 2011), making our environment both comprehensible and functional (Võ, 2021; Võ et al., 2019). Are individual objects from the environment that are not the target of our actions incorporated into the representations we use to guide attention?

Correction (January 2023): This article has been updated with missing details in the funding statement.

Corresponding Author:
Jason Helbing, Goethe University Frankfurt, Department of Psychology, Scene Grammar Lab
Email: jason.helbing@stud.uni-frankfurt.de

A promising candidate category of objects that might be used for behavioral guidance is *anchor objects* (Boettcher et al., 2018; Draschkow & Võ, 2017; Võ, 2021; Võ et al., 2019). These objects are hypothesized to structure the spatial predictions in our surroundings by providing a hierarchy of object information that supports priors (i.e., predictions) about the presence and location of other nearby local objects. For example, a sink predicts not only that the soap is nearby but also specifically that it will be somewhere on top of it; a reading lamp is often next to rather than on top of the bed. In this way, anchors can act as a bridge between target objects and their global scene context (Võ, 2021; Võ et al., 2019).

A commonly used approach to demonstrate how global contextual information affects target-related processes, such as object recognition, visual search, memorization, or action, is to violate regularities within scenes (e.g., by placing the tennis ball in the refrigerator). This subversion of our scene-related expectations can lead to changes in behavior, gaze dynamics, and electrophysiological correlates (Biederman et al., 1982; Davenport & Potter, 2004; Draschkow & Võ, 2017; Ganis & Kutas, 2003; Henderson et al., 1999; Hollingworth & Henderson, 1998; Lauer & Võ, 2022; Võ & Henderson, 2011; Võ & Wolfe, 2013a, 2013b). This approach has been used to investigate how the relationship between targets and global scene context influences cognition, but it can also be utilized to investigate how other objects in the environment guide behavior (Mack & Eckstein, 2011).

In the present study, we observed search in realistic 3D virtual reality environments and independently manipulated (a) the local availability of anchor objects (by replacing them by size-matching gray cuboids) as well as (b) the consistency of the high-level global scene context (by rearranging all objects against expectations, essentially shuffling object locations). This allowed us to investigate whether search behavior is guided by the anchor objects' semantic identity (what specific anchor object it is) and by the spatial arrangement of anchor objects (how they provide a rough spatial layout for local objects, i.e., a syntax of sorts; Võ et al., 2019). To increase ecological validity, we used a repeated-search design in which participants completed a large number of searches in one scene multiple times (Hout & Goldinger, 2010; Võ & Wolfe, 2012, 2013b; Wolfe, Alvarez, et al., 2011). Furthermore, these repetitions in our design allowed us to control for a variety of design-related variables that are known to contribute to learning in repeated search (Li et al., 2016; Võ & Wolfe, 2015). Combining virtual reality with eye and motion tracking allowed us to capture eye movements and body locomotion simultaneously. Given how indicative eye movements are of top-down control processes in everyday tasks (Land & Hayhoe, 2001), our study provided us with optimal measures to investigate how auxiliary anchor-object information guides attention and locomotion in natural behavior.

Statement of Relevance

Everyday tasks, such as finding a teakettle, often appear effortless despite requiring us to move our entire body through space. We waste little attentional and locomotive effort in this search because we can use knowledge about what we are looking for (the teakettle is blue) and its likely surroundings (the teakettle is in the kitchen). It is less clear whether objects that are not the target (e.g., the stove) are also incorporated in the representations that guide our behavior. Using realistic but highly controlled virtual reality environments in combination with eye and motion tracking, we demonstrated that meaningful nontarget information facilitates attentional allocation, speeds object recognition, and minimizes costly body movements. These findings highlight the important realization that the representations we use to make us efficient actors in natural search behavior can contain entire bound objects that are not the target of our actions.

We hypothesized that when people search for objects in scenes, both the semantic identity of anchor objects and their spatial arrangement guide search behavior and, thus, facilitate the localization and recognition of objects. This guidance should be apparent in eye-tracking measures related to (a) how efficiently targets are located (time to first target fixation, number of fixations per trial, scan-path length) and (b) how quickly objects are recognized (the time between first target fixation and the participants' response; decision time) as well as (c) motion-tracking parameters capturing how much participants move (length and spatial extent of movement before finding the target). Specifically, we hypothesized that the semantic identity of anchor objects and their spatial arrangement interact in their guidance: Finding the target should be most efficient in consistent scenes with intact anchors; removing anchor information in spatially consistent scenes or scrambling object locations in scenes where anchor information is available should interfere with the representations guiding search behavior, making it harder to find the target. However, search in inconsistently arranged scenes without anchor information should result in more efficient object localization than search in inconsistently arranged scenes with anchor information, because the anchors' semantic identity cannot be used to guide attention meaningfully in the absence of regular spatial relations between objects. Therefore, in spatially inconsistent scenes, we expected anchors to interfere with search guidance.

Method

Participants

We recruited 24 participants (a convenience sample acquired through on-campus and social media advertising in the summer of 2019; age: $M = 23.5$ years, range = 18–37 years; 18 women and 6 men; 22 right-handed and 2 left-handed; height: $M = 170.1$ cm, range = 155–183 cm) at Goethe University Frankfurt. Sample size was set to be larger (Brysbaert, 2019) than in a similar study (Boettcher et al., 2018) in which three experiments revealed robust results with 12 participants. Here, we set the sample size to 24 to enable counterbalancing. Participants were fluent German speakers, had normal or corrected-to-normal visual acuity (at least 20/25 vision) and normal color vision as assessed by the Ishihara test, and reported no neurological diseases. All participants were volunteers, gave informed consent, and were compensated with either course credit or €24. Participants were naive to the purpose of the experiment.

The research protocol was approved by the local ethics committee of the Faculty of Psychology and Sport Sciences at Goethe University Frankfurt.

Apparatus

To implement our virtual reality eye-tracking paradigm, we used a Tobii Pro VR Integration unit (Tobii Pro, Danderyd, Sweden), which is a retrofitted version of the HTC Vive head-mounted display (HTC Corporation, Taoyuan City, Taiwan). The Tobii Pro VR Integration unit has a built-in binocular dark-pupil eye tracker that streams eye movements at a sampling rate of 90 Hz (the refresh rate of the head-mounted display) with a declared spatial accuracy of approximately 0.5° and a 100° (horizontally) × 110° (vertically) trackable field of view (full field of the head-mounted display). Past assessments of the eye tracker's practically achievable accuracy have yielded a precision below 1.1° within a 20° window centered in the view ports and a worst-case maximal latency below 30 ms (David et al., 2020, 2021). The head-mounted display uses two organic light-emitting diode (OLED) screens with a resolution of 1,080 × 1,200 pixels. Two base stations (Lighthouse tracking system) emit 60 infrared pulses per second, which are detected by 37 infrared sensors in the head-mounted display; this enables location tracking to a fraction of a millimeter. Tracking is further optimized by an accelerometer and a gyroscope in the head-mounted display. Participants held an HTC Vive controller in their writing hand. The trigger at the back of this wireless controller, which participants were instructed to pull with their index finger, was used for response collection.

The experiment was programmed and run in *Unity* (Version 2017.3.0; Unity Technologies, 2017) using *SteamVR* (Version 1.6.10; Valve Corporation, 2019) on a computer equipped with Microsoft Windows 10.

Environments

Sixteen virtual indoor scenes were created (three living rooms, three bedrooms, three bathrooms, three kitchens, and four offices; Fig. 1a). They were all of equal size, approximately 380 cm (length) × 350 cm (width) × 260 cm (height). Textures for wall coverings, flooring materials, and ceilings were tailored to the room category (e.g., tiles in the bathrooms). In every scene, there were 36 category-appropriate objects. All of them were singletons, meaning that no object (or a different exemplar from the same object category) was present more than once in the same scene. In every scene, one object was the door of the room. Of the remaining objects, there were seven that we considered the anchors of the scene and 28 local objects. Anchors were large, static objects (e.g., couch, stove, shower, desk), whereas local objects were smaller and movable items (e.g., pillow, frying pan, shampoo bottle, pencil) that people typically interact with when performing actions in a scene. In addition to these experimental scenes, there was a practice room with objects that would not be expected in any of the other presented scene categories (e.g., traffic light, diving helmet, triceratops) to avoid any memory interference with the experimental scenes. The 3D models used for the scenes were a mixture of purchased assets from CGAxis and free resources taken from several online repositories (Archive 3D, CGTrader, Free3D, TurboSquid, and the Unity Asset Store).

Using a 2 × 2 design (Scene Consistency × Anchor Presence), we created four different versions of every scene (Fig. 1b). In the syntactically consistent version with intact anchors, the scene was entirely in keeping with expectations about its components and their arrangement. Manipulating *scene consistency* entailed repositioning all objects (anchors and local objects independently) to locations in which they would not be expected, hence creating an inconsistent scene in which the spatial link between anchors and their local objects was broken (Draschkow & Võ, 2017; Võ & Wolfe, 2013a, 2013b). In inconsistently arranged scenes, objects did, however, adhere to the laws of physics (e.g., did not float or intersect with one another) and were not placed in a way that occluded them significantly compared to their location in consistent scenes. The inconsistent object arrangement was prepared by the experimenters beforehand and was the same for all participants (i.e., if the inconsistent location of a coffee mug in a bedroom was chosen to be on a pillow, all participants visiting

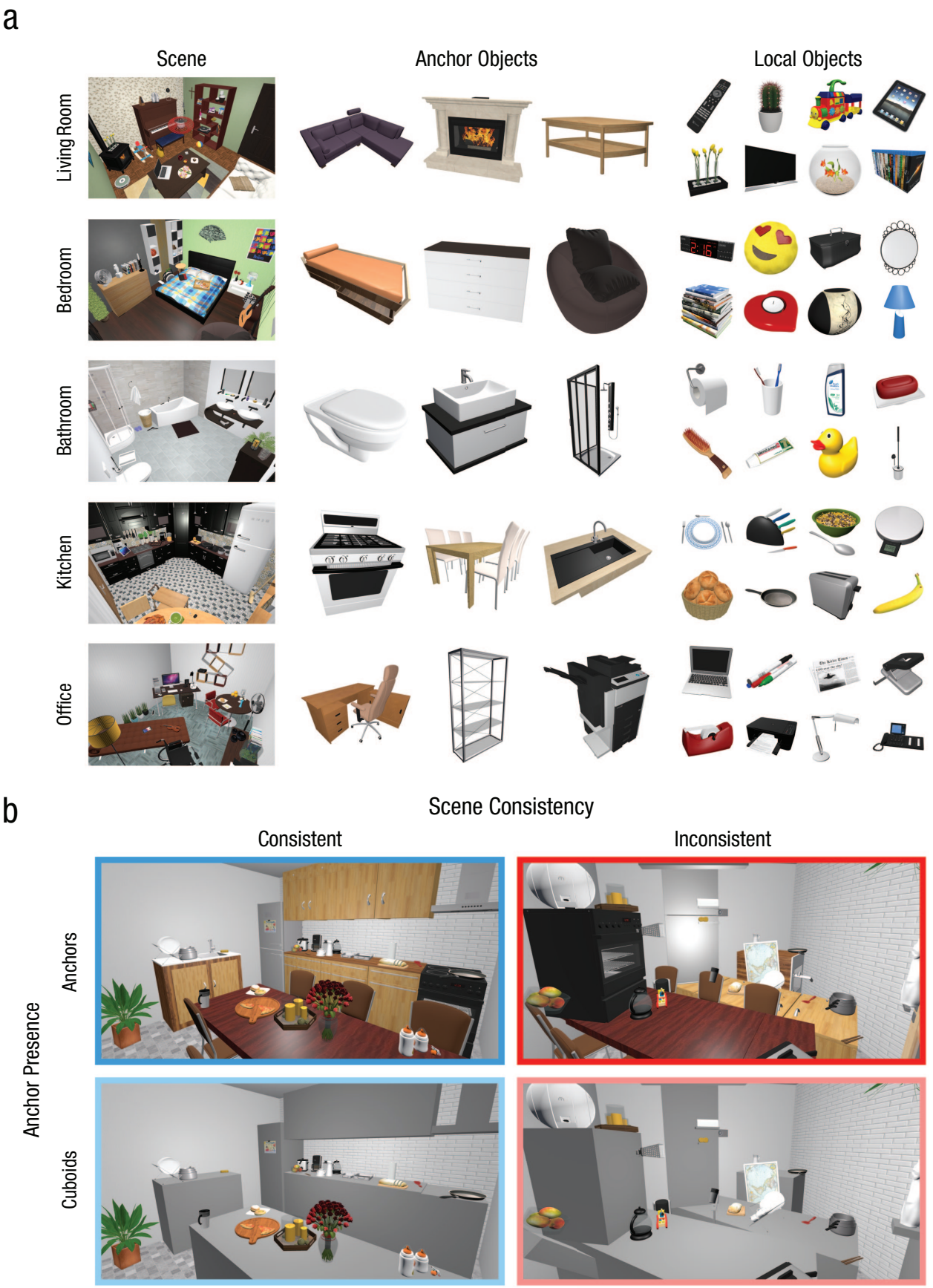

Fig. 1. *(continued on next page)*

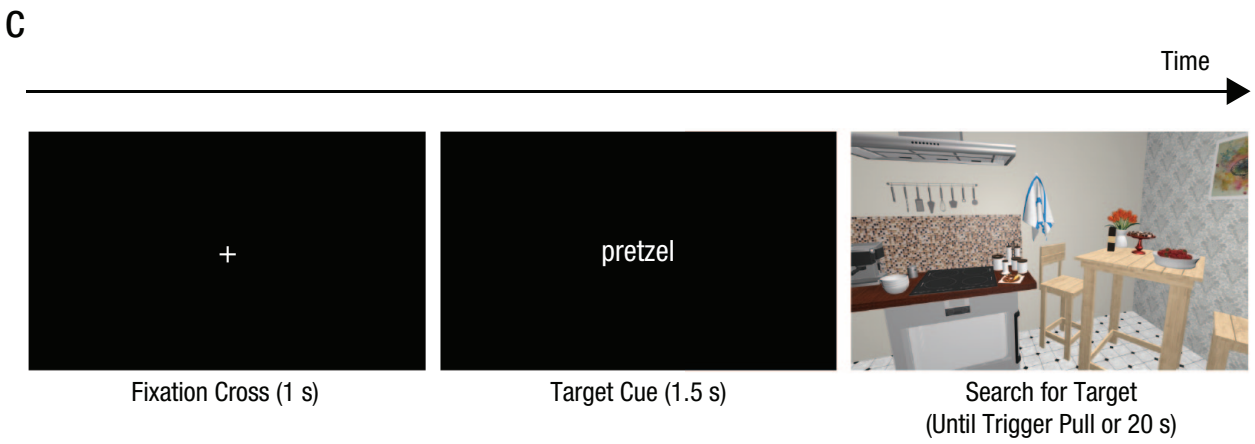

Fig. 1. Experimental stimuli, conditions, and trial sequence. Example scenes, anchor objects, and local objects from each of the five room categories are shown in (a). The four scene-manipulation conditions are shown in (b): These consisted of consistently or inconsistently arranged scenes in which anchor objects were either intact or replaced by cuboids. The procedure of a single search trial is shown in (c). Note that the target cue is presented here in English for display purposes (it was presented in German in the experiment).

this scene in the inconsistent condition would find the mug in this location). The manipulation of *anchor presence* consisted of replacing anchor objects (and the door) by formations of gray cuboids, the sizes of which matched those of the anchors. Therefore, besides (a) the regular scenes (consistently arranged with intact anchors), there were also (b) consistently arranged scenes with cuboids for anchors, (c) inconsistently arranged scenes with intact anchors, and (d) inconsistently arranged scenes with cuboids for anchors.

Images showing overviews of all scenes in all conditions of the experiment are provided in the Supplemental Material available online.

Procedure

After arriving at the lab, participants were familiarized with the virtual reality apparatus and lab space as well as the calibration procedure of the eye tracker. Once equipped with the head-mounted display and controller, they were instructed to search for the cued objects in the scene on every trial and to pull the trigger on the controller while looking at the target once they had located it. They were informed that they could move freely within the virtual rooms, that the targets were always present exactly once in the scene, and that there was a time-out after 20 s. There were 10 practice trials in the practice room before the actual experimental trials started.

A video demonstration of example trials is available at https://osf.io/5xhet/. In every trial, participants were first presented with a fixation cross for 1 s. Then, a verbal cue in German was presented for 1.5 s, indicating the search target of the trial. Both the cue and a plus sign that was used as the fixation cross were presented in white 64-point sans-serif font at a viewing distance of about 80 cm in the center of the display (and would move along with participants' movements to remain there). The visual surroundings were completely black during the fixation cross and the presentation of the target cue. Once the target cue disappeared, the scene became visible, and participants could search in it until they either pulled the trigger or the search time-out of 20 s was reached (Fig. 1c).

There were 25 consecutive trials in each scene and 16 different scenes per participant (four in each condition). Between scenes, the environment changed into an empty room with gray walls in which participants had to move to a small blue square on the floor and could then initiate the next scene's search trials. This was done to ensure that (a) when starting search trials in a new scene, participants would not stand inside of objects and (b) all participants started from roughly the same point with all objects equally visible. A 5-point calibration of the eye tracker was carried out after every fourth scene. Once search trials in all scenes were completed, participants revisited every scene and performed the same search task again with the same trials (second episode) and then one more time (third episode). There were 10-min breaks between episodes. The entire experimental session, including instructions and breaks, took between 2.5 hr and 3 hr.

The assignment of scenes to conditions (scene consistency, anchor presence) was different for every participant: Scenes were randomly assigned to the four conditions with the constraint that there could not be more than one scene of the same category in any condition. Given the number of scenes in each room category (see the Environments section), this meant that

there was one office in each of the four conditions, whereas each of the other room categories (living room, bedroom, kitchen, bathroom) was missing from one condition, as there were just three exemplars of each. The order of scenes was also balanced with respect to the conditions: Every second scene had cuboids in place of anchor objects (the state of the first scene alternated with every participant), and consistency was varied in an ABBA–BAAB–ABBA–BAAB pattern. For each scene, there was a fixed set of 25 targets (out of the 28 local objects). The experimenters selected the targets on the basis of the objects' nameability (i.e., objects for which it was hard to find a conventional official name were avoided as targets because the cuing procedure was achieved by means of verbal labels). The order of the 25 trials in a scene was random in every episode.

Data analysis

Data exclusion. Analyses were performed only on trials in which participants responded accurately, that is, trials in which the target was found (*hits*; 97.1%). A trial was considered accurate when gaze was detected on the bounding box of the target object (the smallest possible cuboid around the convex hull of the 3D object mesh) at the moment the trigger was pulled. Additionally, all non–time-out trials in which this was not the case were rewatched after data collection to check whether the participant had actually misidentified the object or whether gaze was just not on the target because of imprecisions of eye tracking or because the participant prematurely pulled the trigger a moment before their gaze would have hit the target. Trials in which the participant had most likely been right about the target were coded as accurate. Of all hits, gaze was on the target at the trigger pull on 92.9%. About half of the inaccurate trials were time-outs (47.1%).

Eye-tracking measures. Eye-movement samples (gaze points) were recorded at 90 Hz. For fixation filtering, we used a velocity-based algorithm (Salvucci & Goldberg, 2000; velocity-threshold identification [I-VT]) with a velocity threshold of 100° per second (Tobii Pro, 2018) and an additional minimum fixation duration of 100 ms. To account for small, bridgeable tracking interruptions, we allowed for gaps of up to 75 ms between two consecutive gaze points for both to be considered part of the same fixation (Komogortsev et al., 2010). *Time to first fixation* was calculated as the time that elapsed between search onset and the beginning of the first fixation on the target object of the trial. This measure was computed only on trials in which the target was fixated at least once and the first target fixation did not start at search onset (84.5% of hits). *Decision time* was obtained by subtracting the time to first fixation from the trial's response time (i.e., elapsed time between search onset and the point in time at which the trigger was pulled). The *number of fixations* is a simple trial-based fixation count (on all trials with a target fixation; 92.8% of hits). *Scan-path length* was computed as the sum of euclidian distances of consecutive fixations' centroids. Naturally, this measure was obtainable only on trials with more than one fixation (80.6% of hits). We used the time to first fixation, number of fixations, and scan-path length as measures of how efficiently overt attention was guided in a search trial. We interpreted decision time as a measure of how quickly targets were identified once fixated (object recognition).

Locomotion data. The position of the head-mounted display in 3D space was sampled at 90 Hz as well. From this, we calculated two measures of how much participants had moved on a trial. The *length of movement* was computed as the sum of euclidian distances of the horizontal-plane coordinates of consecutive position samples. The *spatial extent of movement* was approximated by calculating the surface area of the convex hull of all position samples' horizontal-plane coordinates. We considered both of these trial-based measures of how efficiently participants moved in a search trial.

Statistical model and software. Data preprocessing and analyses were carried out in the *R* statistical programming language (Version 3.6.2; R Core Team, 2019) using *RStudio* (Version 1.2.5033; RStudio Team, 2019). Linear mixed-effects models (LMMs) and generalized linear mixed-effects models (GLMMs), run with the *lme4* package (Version 1.1-21; Bates et al., 2015), were used to analyze the effects in our data. We chose to use LMMs and GLMMs because they allowed us to control for between-subject and between-stimulus variance simultaneously and, thus, yielded advantages over traditional general-linear-model approaches, such as F1/F2 analyses of variance (Baayen et al., 2008; Kliegl et al., 2011). The lmer_alt() wrapper from the *afex* package (Version 1.0-1; Singmann et al., 2021) was used to correctly remove correlations between random effects. The final models' architecture is specified as follows for all dependent variables:

$$\begin{aligned}\log\left(Y_{ijk}\right) = {} & \beta_0 + S_{0j} + I_{0k} + \left(\beta_1 + S_{1j}\right)X_{1i} + \left(\beta_{2a} + S_{2aj}\right)X_{2ai} \\ & + \left(\beta_{2b} + S_{2bj}\right)X_{2bi} + \beta_3 X_{3i} + \beta_{4a} X_{4ai} + \beta_{4b} X_{4bi} \\ & + \beta_5 X_{5i} + \beta_6 X_{6i} + \beta_{12a} X_{1i} X_{2ai} + \beta_{12b} X_{1i} X_{2bi} \\ & + \beta_{13} X_{1i} X_{3i} + \ \ldots + \beta_{56} X_{5i} X_{6i} + \varepsilon_{ijk}.\end{aligned}$$

In this equation, Y_{ijk} represents the dependent variable outcome i of subject j with search target (item) k, β_0 is the fixed intercept, S_{0j} is the random intercept of subject j, I_{0k} is the random intercept of item k, β_l is the fixed-effect parameter of X_l (double-index β_{lm} indicates two-way interactions X_lX_m), X_{li} is the predictor l of outcome i (l: 1 = scene consistency, 2 = anchor presence, 3 = trial number, 4 = episode number, 5 = incidental gaze duration, 6 = target angle), S_{lj} is the random X_l slope of subject j, and ε_{ijk} represents the residual of outcome i (subject j, item k). Note that (a) for predictors and their fixed-effects parameters, when one factor is coded into two variables for contrasts (Scene Consistency × Anchor Presence, episode transitions), this is indicated by subscript letters a and b behind the variable index l; (b) in case of the number of fixations, we did not log-transform the fixation count but instead used a Poisson link function (GLMM); and (c) that for the scan-path-length model only, the random by-participant slopes for anchor presence, $S_{2,j}$, were restricted to zero.

All models were fitted with the restricted-maximum-likelihood criterion. For each model, we report unstandardized regression coefficients with the t statistic (or z statistic in case of the fixation-count GLMM) and the results of a two-tailed test corresponding to a 5% error criterion for significance. To obtain p values for LMMs, we used an implementation of Satterthwaite's degrees-of-freedom method from the *lmerTest* package (Version 3.1-1; Kuznetsova et al., 2017); GLMM p values were based on asymptotic Wald tests from *lme4*. Further details about the model structure and the model-selection procedure are outlined in the Supplemental Material.

Dependent variables. To investigate the impact of our scene manipulations on the search process, we used the time to first fixation, decision time, number of fixations, scan-path length, length of movement, and spatial extent of movement as dependent measures. Of these, we interpreted the time to first fixation, number of fixations, and scan-path length as measures of how efficiently objects were localized, and we used decision time as indicative of how rapidly the objects' identity was verified (object recognition/identification). With the two movement measures, we aimed to identify differences in how much participants moved through the scenes in the different conditions. After inspecting all dependent variables' distributions, linear model residuals, and power coefficients (λ) of the Box–Cox procedure (Box & Cox, 1964), which was run with the *MASS* package (Version 7.3-51.5; Venables & Ripley, 2002), we log-transformed these values to approximate a normal distribution more closely and meet LMM assumptions. The only exception to this was the fixation count, which was not log-transformed (O'Hara & Kotze, 2010); instead, we used a Poisson GLMM to predict the number of fixations.

Results

We found that overt attention (as assessed by eye movements indicative of efficient target localization) and locomotion were supported by auxiliary anchor-object information across all dependent variables. Below, we break these effects down in more detail. Effects related to the interaction of scene consistency and anchor semantics, which are central to our research question, are described in the following three sections sorted by topic (overt attention, object recognition, body locomotion). All other significant effects are outlined in the Supplemental Material: They largely replicate well-known effects from the visual-search and scene-perception literature (Draschkow & Võ, 2017; Lauer & Võ, 2022; Võ & Wolfe, 2013b, 2015; Wolfe, 2020). All eye- and motion-tracking measures' LMM or GLMM parameter estimates, with their t/z statistic and corresponding p values, are given in Table S1 in the Supplemental Material.

Auxiliary scene information guides overt attention

In consistent scenes with intact anchors, the target was fixated numerically more quickly than in consistent scenes in which cuboids replaced those anchors (Fig. 2a); however, this effect was not significant, $b = 0.04$, $t = 2.03$, $p = .05$. The time to the first target fixation was faster for cuboids than for intact anchors in inconsistent scenes, $b = -0.09$, $t = -3.39$, $p = .002$. Further, there were fewer fixations on trials in consistent scenes with intact anchors than in consistent scenes with cuboids, $b = 0.05$, $t = 3.56$, $p < .001$ (Fig. 2b). In inconsistent scenes, this effect was again reversed: More fixations were made when anchors were present than when cuboids were present, $b = -0.08$, $t = -6.66$, $p < .001$. Finally, in consistent scenes, scan paths were longer when anchors were replaced by cuboids, $b = 0.08$, $t = 3.17$, $p = .002$ (Fig. 2c). For inconsistent scenes, scan-path length was shorter in scenes with cuboids than in those with anchors, $b = -0.05$, $t = -2.26$, $p = .02$. In short, in consistent scenes, the presence of anchors facilitated search, whereas it disrupted attentional guidance in inconsistent scenes (causing less efficient search).

The successful attentional guidance by the anchor objects is further illustrated in the example spatial distribution of fixations in Figure 2e. The auxiliary anchor objects provided useful guidance in consistently arranged scenes but became distracting visual clutter in inconsistent

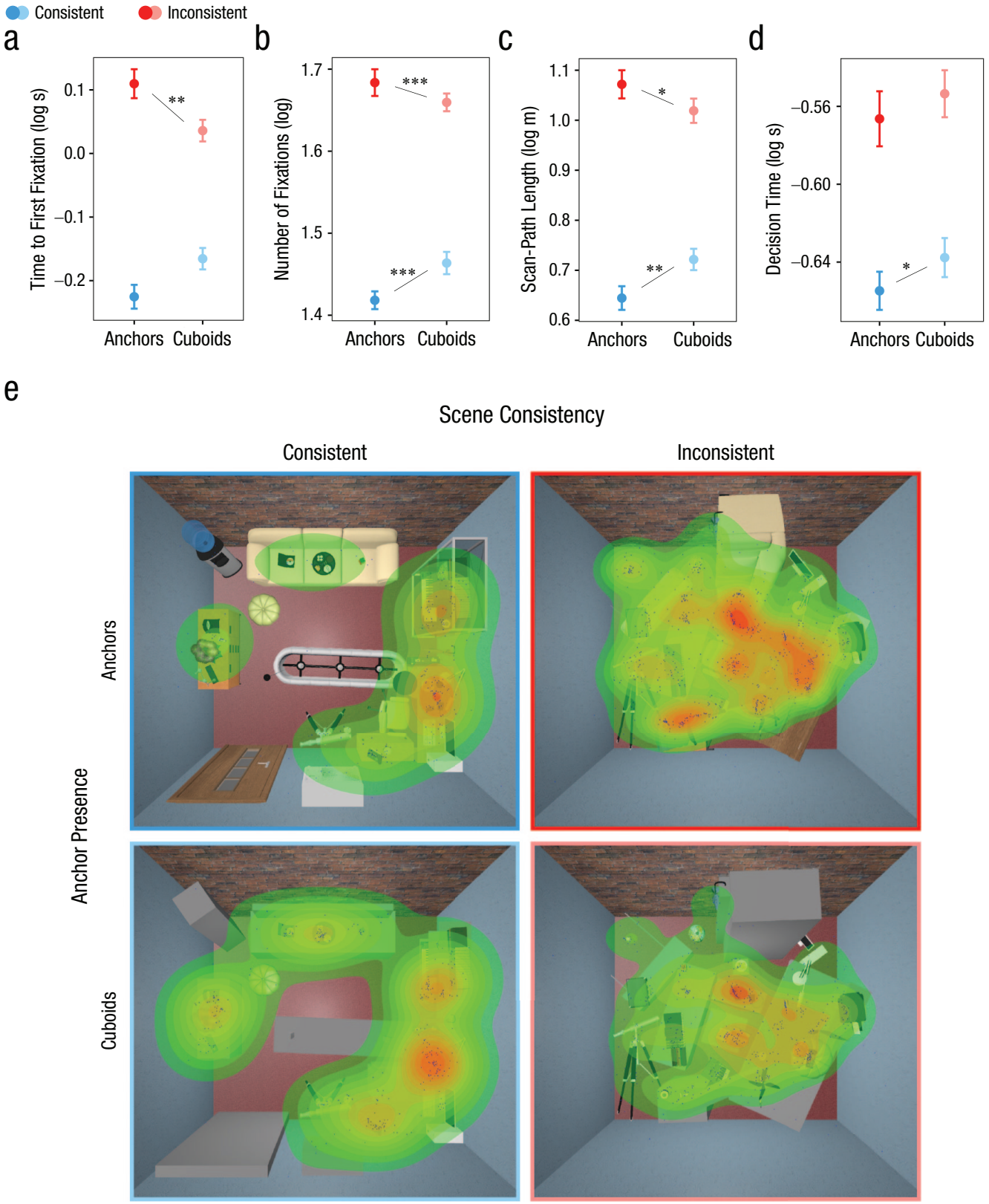

Fig. 2. Eye-movement results. The graphs show the effect of anchor presence (anchors vs. cuboids) and scene-consistency manipulation (consistent vs. inconsistent) on time to first fixation (a), number of fixations (b), scan-path length (c), and decision time (d). Asterisks indicate significant differences between anchor-presence conditions (*$p < .05$, **$p < .01$, ***$p < .001$). Error bars represent standard errors of the mean. The distribution of fixations in space (e) is shown for the first five search trials of all participants in the four different conditions of an office scene. Each blue dot represents a fixation, and the color gradient reflects the density of fixations with the length of individual fixations taken into account.

scenes, highlighting the interplay of the anchors' identity and arrangement in guiding attention.

Auxiliary scene information aids object recognition

Decision time was calculated as the time between the participants' first target fixation and their response. It is indicative of how rapidly the target identity is verified and functions as a proxy for object recognition/identification. In consistent scenes with intact anchors, decision time was significantly faster than in consistent scenes with cuboids, $b = 0.03$, $t = 2.59$, $p = .01$ (Fig. 2d). For inconsistent scenes, there was no significant difference in decision time between the anchor and cuboid conditions, $b = 0.01$, $t = 0.37$, $p = .72$. These patterns indicate that anchor objects facilitate the identification of nearby local objects in intact scenes, which is in line with classic consistency effects in object recognition (Bar, 2004; Biederman et al., 1982; Davenport & Potter, 2004; Lauer et al., 2018; Sauvé et al., 2017) and recent evidence that scene context helps us to disambiguate bottom-up object information (Wischnewski & Peelen, 2021).

Auxiliary scene information supports efficient locomotion

The pattern of locomotion results resembled that of the eye-tracking measures. In consistent scenes, the length of movement was shorter when anchors were present than when replaced by cuboids, $b = 0.07$, $t = 3.09$, $p = .005$, whereas in inconsistent scenes, it was shorter for cuboids than for anchors, $b = -0.08$, $t = -2.97$, $p = .007$ (Fig. 3a). Likewise, movement in space was more limited in consistent scenes with anchors than in consistent scenes with cuboids, $b = 0.15$, $t = 3.31$, $p = .001$, but was again more extensive in inconsistent scenes with anchors than in inconsistent scenes with cuboids, $b = -0.14$, $t = -2.45$, $p = .02$ (Fig. 3b). These patterns demonstrate that auxiliary scene information not only shapes attentional allocation but also guides body movements in realistic interactions within immersive virtual reality. These effects are also evident in the example movement paths depicted in Figure 3c.

Discussion

Our results show that efficiently locating objects in immersive environments, with respect to both eye and body movements, relies on auxiliary nontarget information provided by a class of stand-alone objects known as anchor objects (Boettcher et al., 2018; Draschkow & Võ, 2017; Võ, 2021; Võ et al., 2019). Efficient attentional guidance and locomotion rely on a combination of (a) the consistent composition of the environments' building blocks and—once this intact spatial layout is provided—(b) the semantic identity of anchor objects. These findings reveal that individual objects from the environment that are not the target of our actions can be incorporated into the representations we use to guide attention and locomotion.

In our study, we showed that auxiliary anchor objects can play an important part in guiding behavior. These objects have been proposed to structure the spatial predictions in natural surroundings by providing a hierarchy of object information that supports priors about the presence and location of nearby potential target objects (Boettcher et al., 2018; Draschkow & Võ, 2017; Võ, 2021; Võ et al., 2019). The conceptualization of these objects stems from approaches designed to describe similarities between the structure of language and the structure of scenes (Biederman, 1972; Biederman et al., 1973, 1982; Võ et al., 2019). In these approaches, scenes can be regarded as "grammatical" compositions of sub-scenelike *phrases* (e.g., a sink phrase), each of which is arranged around a central anchor object (sink) that supports predictions of the presence and location of the nearby local objects (toothbrush, soap, etc.). The efficiency of searching for objects in real-world environments stems from the ability to exclude whole phrases (e.g., the toilet or shower phrase) from the search area when looking for a toothbrush. Our results highlight the behavioral relevance of this phrasal structure within scenes: On an intraphrase level (i.e., when the object arrangement within a phrase is intact), the identity of the anchor object is necessary auxiliary information to improve performance. On an interphrase level (i.e., spatially consistent arrangement vs. inconsistent arrangement), we found that attentional guidance relies on intact phrase-like clusters of objects, as breaking these up decreased search performance (or, in other words, increased search effort).

More global expectations related to what belongs in a scene (*scene semantics*; object identities; e.g., the pot goes in the kitchen) are typically distinguished from rules about where objects are located (*scene syntax*; the pot often rests on a stovetop; Draschkow & Võ, 2017; Võ, 2021; Võ et al., 2019; Võ & Wolfe, 2013a). In addition to this approach being a useful metaphor for describing scene regularities and their violations, there is evidence for commonalities between the processing of language and scenes, as they share similarities in their organization (Draschkow & Võ, 2017; Võ, 2021; Võ et al., 2019; Võ & Wolfe, 2013a) and development (Maffongelli et al., 2020; Öhlschläger & Võ, 2020). In the context of our study, scene semantics and syntax can also be applied to describe our two manipulations. Replacing anchors by cuboids can be considered a manipulation that

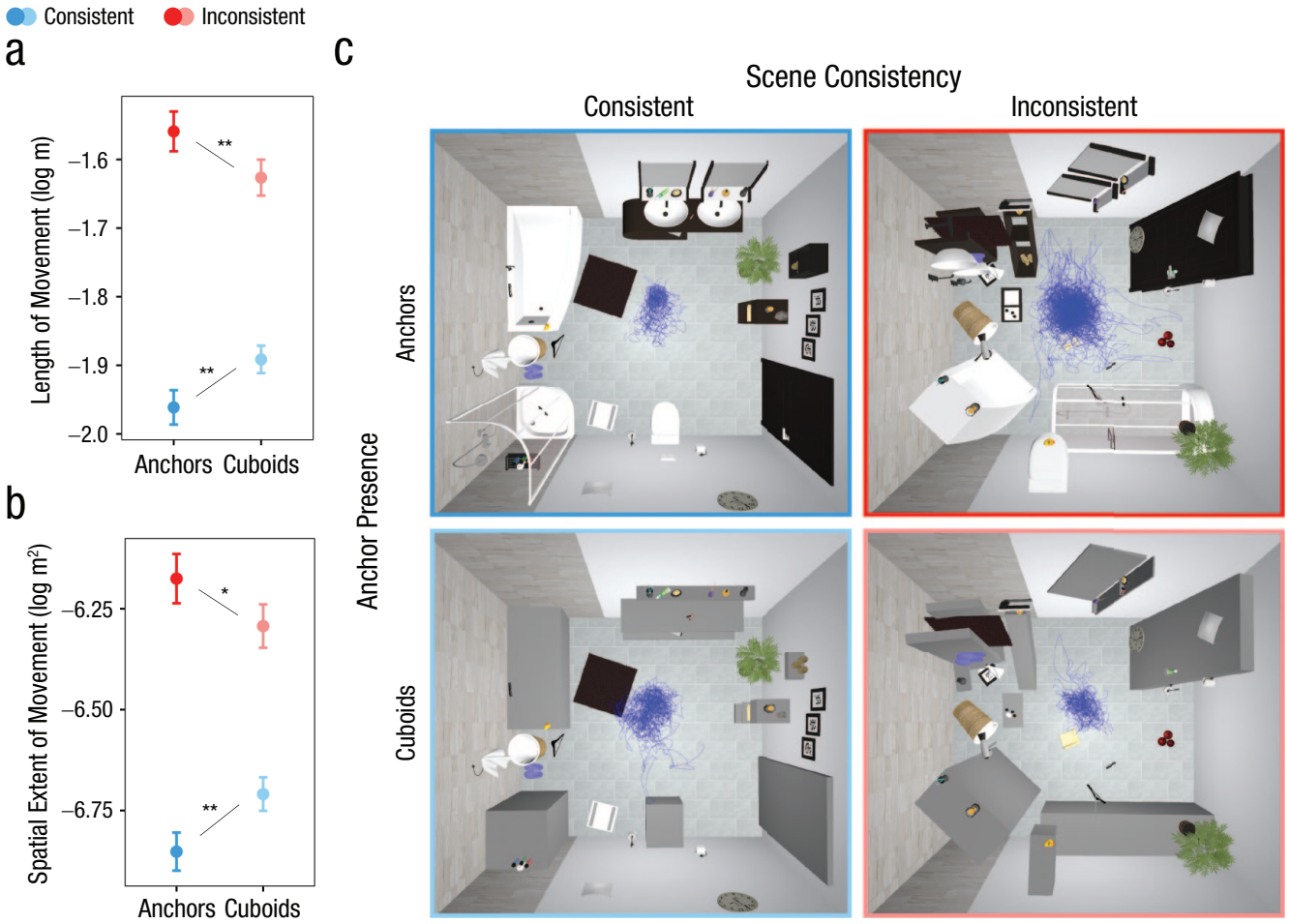

Fig. 3. Body-locomotion results. The graphs show the effect of anchor presence (anchors vs. cuboids) and scene-consistency manipulation (consistent vs. inconsistent) on length of movement (a) and spatial extent of movement (b). Asterisks indicate significant differences between anchor-presence conditions (*$p < .05$, **$p < .01$). Error bars represent standard errors of the mean. Movement paths from all trials of all participants are shown in the four different conditions of a bathroom scene (c). Paths are represented by blue lines.

primarily operates on a semantic level, because the spatial layout (syntax) of other objects in the scene remains intact. The consistency manipulation, on the other hand, can be described as a violation of scene syntax, because the natural spatial layout is distorted. Thus, our results highlight how the interplay of semantic and syntactic scene information can increase the efficiency of attention, locomotion, and object recognition. We want to stress the universality and flexibility of this efficiency because it is not limited to well-known environments—hence the term "grammar." That is, just as we can understand sentences we have never heard before because we know the meaning of the words and the rules of how they need to be arranged to form meaning, we can understand new scenes by knowing the identities of objects and the rules that govern their spatial layout (*scene grammar*; Võ et al., 2019).

This study and previous work have identified anchor objects as building blocks of a hierarchical scene organization, which is of unique importance to how we form predictions of object locations (Boettcher et al., 2018; Draschkow & Võ, 2017; Võ et al., 2019). In future studies, it will be important to investigate these predictions in a more fine-grained manner. Here, we selected larger static objects as anchors and observed how they shaped predictions for the remaining objects as targets. In reality, it is likely that the hierarchy of objects predicting each other in space is more profound than that. For example, many of the objects we selected as local objects are probably anchoring predictions themselves: A large computer monitor on a desk likely predicts the keyboard and mouse resting below. In many cases, these predictions could be multidirectional (e.g., a glass of milk and a plate of cookies side by side, predicting each other). Therefore, more complex object networks, in which weighted links between objects indicate the extent to which they predict each other, will most likely provide us with better models of how spatial priors are

formed during natural behavior. Analyzing large databases of scenes to extract regularities of objects' frequencies, co-occurrences, and spatial relations to each other will be key in this endeavor (Boettcher et al., 2018; Greene, 2013; Võ et al., 2019; Yang et al., 2019). Furthermore, it will be important to look more closely at eye movements during the search process when anchors guide attention: Although we have shown that these are indicative of increased efficiency of the search process when anchor objects and the scene's structure are intact, more research in even more standardized environments is needed to understand precisely how fixations are related to anchor guidance. What role do fixations on anchors play in guiding search? How are saccades between anchors and local objects guided by scene grammar? When do we not fixate (i.e., skip) the anchor before fixating the target?

It is worth noting that we included repetitions in our trial-by-trial design because we believe that repeatedly searching through the same, unchanging environment reflects what we experience daily (rather than jumping from one scene to another, we tend to look for several items within the same scene, e.g., when preparing dinner in a kitchen; Hout & Goldinger, 2010; Võ & Wolfe, 2012; Wolfe, Alvarez, et al., 2011). We accounted for these repetitions in our statistical models, but nevertheless, using different research designs with altered trial structures (e.g., comparing repeated search in changing and unchanging scenes or looking only at initial search trials in a larger number of scenes) will be important when aiming to more precisely disentangle the differential roles of semantic knowledge (general assumptions about scenes, such as those provided by anchor objects) and episodic memory (knowing specific scenes and their unique regularities; Võ & Wolfe, 2013b).

Methodologically, our study joins the rapidly growing list of efforts to investigate search in realistic virtual reality scenes (Beitner et al., 2021; Bennett et al., 2021; David et al., 2020, 2021; Draschkow & Võ, 2017; Enders et al., 2021; Figueroa et al., 2017; Hadnett-Hunter et al., 2019; Helbing et al., 2020; Kit et al., 2014; T. Kristjánsson et al., 2022; Li et al., 2016, 2018; Lukashova-Sanz & Wahl, 2021; Olk et al., 2018). Studies such as these enable us to probe search flexibly while ensuring both unprecedented ecological validity (realistic environments, navigable space, and behaviorally relevant task settings) and a high degree of experimental control (precise timing, eye and motion tracking, and full control over the field of view). We believe that this approach is essential in order to replicate, scrutinize, and extend findings from decades of screen-based experimentation on scene perception and visual search. Only when behavior is studied in these naturalistic settings can we get a functional perspective of underlying cognitive processes (Foulsham et al., 2011; Á. Kristjánsson & Draschkow, 2021; Malcolm et al., 2016; Tatler et al., 2011, 2013). To increase the generalizability of our findings to other settings (Yarkoni, 2022), it will be relevant to investigate search in large-scale virtual environments with multiple connected scenes (e.g., apartments, office spaces, train stations), because our representations of these complex multiscene spaces may carry with them unexplored possibilities for auxiliary guidance by contextual information. Further, to increase the generalizability of our findings beyond groups conveniently proximate to the research site (often undergraduate students who might not represent the target population; Henrich et al., 2010), it will be important to sample larger and more representative populations. This large-scale and more diverse sampling can be enabled by remote online experimentation using virtual reality, as the market for consumer virtual reality systems is growing (Draschkow, 2022).

The unparalleled efficiency of natural adaptive behavior in real-world environments is an impressive property of human cognition. Broadly, our findings demonstrate that this efficiency is supported by spatial priors generated by auxiliary information that is not a direct property of the targets of our actions. More precisely, our findings reveal that target representations used for guiding natural behavior can include standalone objects that anchor people's hierarchical representations of scenes and the objects within them.

Transparency

Action Editor: Sachiko Kinoshita
Editor: Patricia Bauer
Author Contributions

D. Draschkow and M. L.-H. Võ contributed equally to this work. All authors conceptualized the experimental design and methodology. J. Helbing programmed the experiment and collected the data. J. Helbing and D. Draschkow analyzed the data and created visualizations. J. Helbing wrote the original draft of the manuscript. D. Draschkow and M. L.-H. Võ reviewed and edited the manuscript. D. Draschkow and M. L.-H. Võ supervised the project. All the authors approved the final version of the manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was supported by a grant from the German Research Foundation (Grant No. SFB/TRR 135, Project No. 222641018, Subproject C7) to M. L.-H. Võ, by a grant from the Hessian Ministry of Science and Art (Project "The Adaptive Mind") to M. L.-H. Võ, and by the Deutschlandstipendium scholarship from the Federal Ministry of Education and Research and Goethe University Frankfurt to J. Helbing.

Open Practices

All data and analysis scripts have been made publicly available via OSF and can be accessed at https://osf.io/jxczq/. The design and analysis plans for this study were not preregistered. This article has received the badge for Open Data. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iD

Jason Helbing https://orcid.org/0000-0003-1158-4534

Acknowledgments

We thank Jenny Helbing and Rieke Löffler for their valuable help with the stimulus material and data collection as well as Julia Beitner and Erwan David for helpful conversations about the work presented here.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/09567976221091838

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Ballard, D. H., Hayhoe, M. M., & Pelz, J. B. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, *7*(1), 66–80. https://doi.org/10.1162/jocn.1995.7.1.66

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*(8), 617–629. https://doi.org/10.1038/nrn1476

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Beitner, J., Helbing, J., Draschkow, D., & Võ, M. L.-H. (2021). Get your guidance going: Investigating the activation of spatial priors for efficient search in virtual reality. *Brain Sciences*, *11*(1), Article 44. https://doi.org/10.3390/brainsci11010044

Bennett, C. R., Bex, P. J., & Merabet, L. B. (2021). Assessing visual search performance using a novel dynamic naturalistic scene. *Journal of Vision*, *21*(1), Article 5. https://doi.org/10.1167/jov.21.1.5

Biederman, I. (1972). Perceiving real-world scenes. *Science*, *177*(4043), 77–80. https://doi.org/10.1126/science.177.4043.77

Biederman, I., Glass, A. L., & Stacy, E. W. (1973). Searching for objects in real-world scenes. *Journal of Experimental Psychology*, *97*(1), 22–27. https://doi.org/10.1037/h0033776

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*(2), 143–177. https://doi.org/10.1016/0010-0285(82)90007-X

Boettcher, S. E. P., Draschkow, D., Dienhart, E., & Võ, M. L.-H. (2018). Anchoring visual search in scenes: Assessing the role of anchor objects on eye movements during visual search. *Journal of Vision*, *18*(13), Article 11. https://doi.org/10.1167/18.13.11

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *26*(2), 211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Brady, T. F., Shafer-Skelton, A., & Alvarez, G. A. (2017). Global ensemble texture representations are critical to rapid scene perception. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(6), 1160–1176. https://doi.org/10.1037/xhp0000399

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*(1), Article 16. https://doi.org/10.5334/joc.72

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science*, *15*(8), 559–564. https://doi.org/10.1111/j.0956-7976.2004.00719.x

David, E., Beitner, J., & Võ, M. L.-H. (2020). Effects of transient loss of vision on head and eye movements during visual search in a virtual environment. *Brain Sciences*, *10*(11), Article 841. https://doi.org/10.3390/brainsci10110841

David, E. J., Beitner, J., & Võ, M. L.-H. (2021). The importance of peripheral vision when searching 3D real-world scenes: A gaze-contingent study in virtual reality. *Journal of Vision*, *21*(7), Article 3. https://doi.org/10.1167/jov.21.7.3

Draschkow, D. (2022). Remote virtual reality as a tool for increasing external validity. *Nature Reviews Psychology*. Advance online publication. https://doi.org/10.1038/s44159-022-00082-8

Draschkow, D., Kallmayer, M., & Nobre, A. C. (2021). When natural behavior engages working memory. *Current Biology*, *31*(4), 869–874.e5. https://doi.org/10.1016/j.cub.2020.11.013

Draschkow, D., & Võ, M. L.-H. (2017). Scene grammar shapes the way we interact with objects, strengthens memories, and speeds search. *Scientific Reports*, *7*, Article 16471. https://doi.org/10.1038/s41598-017-16739-x

Enders, L. R., Smith, R. J., Gordon, S. M., Ries, A. J., & Touryan, J. (2021). Gaze behavior during navigation and visual search of an open-world virtual environment. *Frontiers in Psychology*, *12*, Article 681042. https://doi.org/10.3389/fpsyg.2021.681042

Figueroa, J. C. M., Arellano, R. A. B., & Calinisan, J. M. E. (2017). A comparative study of virtual reality and 2D display methods in visual search in real scenes. In D. N. Cassenti (Ed.), *Advances in human factors in simulation and modeling: AHFE 2017: Advances in intelligent systems and computing* (Vol. 591, pp. 366–377). Springer. https://doi.org/10.1007/978-3-319-60591-3_33

Foulsham, T., Walker, E., & Kingstone, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vision Research*, *51*(17), 1920–1931. https://doi.org/10.1016/j.visres.2011.07.002

Ganis, G., & Kutas, M. (2003). An electrophysiological study of scene effects on object identification. *Cognitive Brain Research*, *16*(2), 123–144. https://doi.org/10.1016/S0926-6410(02)00244-6

Greene, M. R. (2013). Statistics of high-level scene context. *Frontiers in Psychology*, *4*, Article 777. https://doi.org/10.3389/fpsyg.2013.00777

Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, *58*(2), 137–176. https://doi.org/10.1016/j.cogpsych.2008.06.001

Hadnett-Hunter, J., Nicolaou, G., O'Neill, E., & Proulx, M. (2019). The effect of task on visual attention in interactive virtual environments. *ACM Transactions on Applied Perception*, *16*(3), Article 17. https://doi.org/10.1145/3352763

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*(4), 188–194. https://doi.org/10.1016/j.tics.2005.02.009

Helbing, J., Draschkow, D., & Võ, M. L.-H. (2020). Search superiority: Goal-directed attentional allocation creates more reliable incidental identity and location memory than explicit encoding in naturalistic virtual environments. *Cognition*, *196*, Article 104147. https://doi.org/10.1016/j.cognition.2019.104147

Henderson, J. M., Weeks, P. A., Jr., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(1), 210–228. https://doi.org/10.1037/0096-1523.25.1.210

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*, Article 29. https://doi.org/10.1038/466029a

Hollingworth, A., & Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General*, *127*(4), 398–415. https://doi.org/10.1037//0096-3445.127.4.398

Hout, M. C., & Goldinger, S. D. (2010). Learning in repeated visual search. *Attention, Perception, & Psychophysics*, *72*(5), 1267–1282. https://doi.org/10.3758/APP.72.5.1267

Hutchinson, J. B., & Turk-Browne, N. B. (2012). Memory-guided attention: Control from multiple memory systems. *Trends in Cognitive Sciences*, *16*(12), 576–579. https://doi.org/10.1016/j.tics.2012.10.003

Kit, D., Katz, L., Sullivan, B., Snyder, K., Ballard, D., & Hayhoe, M. (2014). Eye movements, visual search and scene memory, in an immersive virtual environment. *PLOS ONE*, *9*(4), Article e94362. https://doi.org/10.1371/journal.pone.0094362

Kliegl, R., Wei, P., Dambacher, M., Yan, M., & Zhou, X. (2011). Experimental effects and individual differences in linear mixed models: Estimating the relationship between spatial, object, and attraction effects in visual attention. *Frontiers in Psychology*, *1*, Article 238. https://doi.org/10.3389/fpsyg.2010.00238

Komogortsev, O. V., Gobert, D. V., Jayarathna, S., Koh, D. H., & Gowda, S. M. (2010). Standardization of automated analyses of oculomotor fixation and saccadic behaviors. *IEEE Transactions on Biomedical Engineering*, *57*(11), 2635–2645. https://doi.org/10.1109/TBME.2010.2057429

Kristjánsson, Á., & Draschkow, D. (2021). Keeping it real: Looking beyond capacity limits in visual cognition. *Attention, Perception, & Psychophysics*, *83*(4), 1375–1390. https://doi.org/10.3758/s13414-021-02256-7

Kristjánsson, T., Draschkow, D., Pálsson, Á., Haraldsson, D., Jónsson, P. Ö., & Kristjánsson, Á. (2022). Moving foraging into three dimensions: Feature- versus conjunction-based foraging in virtual reality. *Quarterly Journal of Experimental Psychology*, *75*(2), 313–327. https://doi.org/10.1177/1747021820937020

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Land, M. F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research*, *41*(25–26), 3559–3565. https://doi.org/10.1016/S0042-6989(01)00102-X

Lauer, T., Cornelissen, T. H. W., Draschkow, D., Willenbockel, V., & Võ, M. L.-H. (2018). The role of scene summary statistics in object recognition. *Scientific Reports*, *8*, Article 14666. https://doi.org/10.1038/s41598-018-32991-1

Lauer, T., Schmidt, F., & Võ, M. L.-H. (2021). The role of contextual materials in object recognition. *Scientific Reports*, *11*, Article 21988. https://doi.org/10.1038/s41598-021-01406-z

Lauer, T., & Võ, M. L.-H. (2022). The ingredients of scenes that affect object search and perception. In B. Ionescu, W. A. Bainbridge, & N. Murray (Eds.), *Human perception of visual information: Psychological and computational perspectives* (pp. 1–32). Springer. https://doi.org/10.1007/978-3-030-81465-6_1

Li, C.-L., Aivar, M. P., Kit, D. M., Tong, M. H., & Hayhoe, M. M. (2016). Memory and visual search in naturalistic 2D and 3D environments. *Journal of Vision*, *16*(8), Article 9. https://doi.org/10.1167/16.8.9

Li, C.-L., Aivar, M. P., Tong, M. H., & Hayhoe, M. M. (2018). Memory shapes visual search strategies in large-scale environments. *Scientific Reports*, *8*, Article 4324. https://doi.org/10.1038/s41598-018-22731-w

Lukashova-Sanz, O., & Wahl, S. (2021). Saliency-aware subtle augmentation improves human visual search performance in VR. *Brain Sciences*, *11*(3), Article 283. https://doi.org/10.3390/brainsci11030283

Mack, S. C., & Eckstein, M. P. (2011). Object co-occurrence serves as a contextual cue to guide and facilitate visual search in a natural viewing environment. *Journal of Vision*, *11*(9), Article 9. https://doi.org/10.1167/11.9.9

Maffongelli, L., Öhlschläger, S., & Võ, M. L.-H. (2020). The development of scene semantics: First ERP indications for the processing of semantic object-scene inconsistencies in 24-month-olds. *Collabra: Psychology*, *6*(1), Article 17707. https://doi.org/10.1525/collabra.17707

Malcolm, G. L., Groen, I. I. A., & Baker, C. I. (2016). Making sense of real-world scenes. *Trends in Cognitive Sciences*, *20*(11), 843–856. https://doi.org/10.1016/j.tics.2016.09.003

Neider, M. B., & Zelinsky, G. J. (2006). Scene context guides eye movements during visual search. *Vision Research*, *46*(5), 614–621. https://doi.org/10.1016/j.visres.2005.08.025

O'Hara, R. B., & Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, *1*(2), 118–122. https://doi.org/10.1111/j.2041-210X.2010.00021.x

Öhlschläger, S., & Võ, M. L.-H. (2020). Development of scene knowledge: Evidence from explicit and implicit scene knowledge measures. *Journal of Experimental Child Psychology*, *194*, Article 104782. https://doi.org/10.1016/j.jecp.2019.104782

Olk, B., Dinu, A., Zielinski, D. J., & Kopper, R. (2018). Measuring visual search and distraction in immersive virtual reality. *Royal Society Open Science*, *5*(5), Article 172331. https://doi.org/10.1098/rsos.172331

R Core Team. (2019). *R: A language and environment for statistical computing* (Version 3.6.2) [Computer software]. https://www.R-project.org/

RStudio Team. (2019). *RStudio: Integrated development for R* (Version 1.2.5033) [Computer software]. http://www.rstudio.com/

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In A. T. Duchowski (Chair), *Proceedings of the 2000 Symposium on Eye Tracking Research and Applications (ETRA '00)* (pp. 71–78). Association for Computing Machinery. https://doi.org/10.1145/355017.355028

Sauvé, G., Harmand, M., Vanni, L., & Brodeur, M. B. (2017). The probability of object–scene co-occurrence influences object identification processes. *Experimental Brain Research*, *235*(7), 2167–2179. https://doi.org/10.1007/s00221-017-4955-y

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2021). *afex: Analysis of factorial experiments* (Version 1.0-1) [Computer software]. https://CRAN.R-project.org/package=afex

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, *11*(5), Article 5. https://doi.org/10.1167/11.5.5

Tatler, B. W., Hirose, Y., Finnegan, S. K., Pievilainen, R., Kirtley, C., & Kennedy, A. (2013). Priorities for selection and representation in natural tasks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1628), Article 20130066. https://doi.org/10.1098/rstb.2013.0066

Tobii Pro. (2018, April 24). *When do I use the I-VT attention filter?* https://connect.tobiipro.com/s/article/When-do-I-use-the-I-VT-Attention-filter

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*(4), 766–786. https://doi.org/10.1037/0033-295X.113.4.766

Unity Technologies. (2017). *Unity* (Version 2017.3.0) [Computer software]. https://unity.com/

Valve Corporation. (2019). *SteamVR* (Version 1.6.10) [Computer software]. https://store.steampowered.com/steamvr

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.

Võ, M. L.-H. (2021). The meaning and structure of scenes. *Vision Research*, *181*, 10–20. https://doi.org/10.1016/j.visres.2020.11.003

Võ, M. L.-H., Boettcher, S. E. P., & Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology*, *29*, 205–210. https://doi.org/10.1016/j.copsyc.2019.03.009

Võ, M. L.-H., & Henderson, J. M. (2011). Object–scene inconsistencies do not capture gaze: Evidence from the flash-preview moving-window paradigm. *Attention, Perception, & Psychophysics*, *73*(6), 1742–1753. https://doi.org/10.3758/s13414-011-0150-6

Võ, M. L.-H., & Wolfe, J. M. (2012). When does repeated search in scenes involve memory? Looking at versus looking for objects in scenes. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 23–41. https://doi.org/10.1037/a0024147

Võ, M. L.-H., & Wolfe, J. M. (2013a). Differential electrophysiological signatures of semantic and syntactic scene processing. *Psychological Science*, *24*(9), 1816–1823. https://doi.org/10.1177/0956797613476955

Võ, M. L.-H., & Wolfe, J. M. (2013b). The interplay of episodic and semantic memory in guiding repeated search in scenes. *Cognition*, *126*(2), 198–212. https://doi.org/10.1016/j.cognition.2012.09.017

Võ, M. L.-H., & Wolfe, J. M. (2015). The role of memory for visual search in scenes. *Annals of the New York Academy of Sciences*, *1339*(1), 72–81. https://doi.org/10.1111/nyas.12667

Wischnewski, M., & Peelen, M. V. (2021). Causal neural mechanisms of context-based object recognition. *eLife*, *10*, Article e69736. https://doi.org/10.7554/eLife.69736

Wolfe, J. M. (2020). Visual search: How do we find what we are looking for? *Annual Review of Vision Science*, *6*, 539–562. https://doi.org/10.1146/annurev-vision-091718-015048

Wolfe, J. M. (2021). Guided Search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review*, *28*(4), 1060–1092. https://doi.org/10.3758/s13423-020-01859-9

Wolfe, J. M., Alvarez, G. A., Rosenholtz, R., Kuzmova, Y. I., & Sherman, A. M. (2011). Visual search for arbitrary objects in real scenes. *Attention, Perception, & Psychophysics*, *73*(6), 1650–1671. https://doi.org/10.3758/s13414-011-0153-3

Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour*, *1*(3), Article 0058. https://doi.org/10.1038/s41562-017-0058

Wolfe, J. M., Võ, M. L.-H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences*, *15*(2), 77–84. https://doi.org/10.1016/j.tics.2010.12.001

Yang, W., Wang, X., Farhadi, A., Gupta, A., & Mottaghi, R. (2019, May 7). *Visual semantic navigation using scene priors* [Conference session]. Seventh International Conference on Learning Representations, New Orleans, LA, United States. https://arxiv.org/abs/1810.06543

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, Article E1. https://doi.org/10.1017/S0140525X20001685