General Comments in Response to Reviewers:

[Please see pg. 3 for point-by-point responses]

We very much thank the editor and all three reviewers for their thoughtful comments on our Stage 1 submission. In response to this feedback, we have made several changes to our preregistered report, which we agree have strengthened our submission. Specifically, we have made the following major changes:

1) As noted by Reviewers 1 and 2, in our first submission, our Introduction was not entirely clear. Specifically, our critical research question (does curiosity *directly* benefit memory?) was obscured by our discussion of our effort to replicate Cen et al. (2024), as well by as our discussion of the effects of active vs. passive learning. In our resubmission, we have reframed our Introduction and hypotheses to better emphasize the novel aspects of our planned study.

2) Reviewers 1 and 3 both raised concerns surrounding our power analysis. In particular, Reviewer 1 suggested that we determine our sample size based on the interaction between experimental condition and curiosity (the critical effect for our novel research question), rather than determining our sample size based on our effort to replicate Cen et al. (2024). We very much agree with this suggestion in principle; however, determining the appropriate sample size with which to target the [experimental condition x curiosity] interaction is complicated by the fact that our predictions concerning the direction and magnitude of this interaction are nonspecific. Thus, we have modified our sampling strategy to incorporate Bayesian stopping criteria ("Sequential Bayes Factor with maximal *n*"; Schönbrodt & Wagenmakers, 2018). We plan to recruit *either* to a maximum of 120 participants, *or* until we obtain compelling evidence for or against the critical two-way interaction [which we define as Bayes Factor (BF) > 10 or BF < .10, respectively]. This sampling strategy is more fully detailed in response to Reviewer 1, as well as in our revised Method.

3) Reviewer 3 asked us to clarify our use of the term 'environmental memory', and we agree that this term was not appropriately defined in our original submission. In our revision, and for consistency with Cen et al. (2024), we now refer to 'cognitive map formation' rather than 'environmental memory'. Specifically, we are assessing the levels of accuracy and detail present in participants' cognitive maps of explored environments (as indexed by their room drawings).

4) This point was not raised by the previous round of reviews, but, in our original submission, we planned to conduct three memory tests (Room Drawing, Free Recall, and Introspective Recall). However, like Cen et al. (2024), we only planned to analyse the data from the Room Drawing test.

   In our revised submission, we have entirely removed the remaining two tests (Free Recall and Introspective Recall) from our methodology. We made this decision because it substantially reduces the amount of time required to run the full procedure, which in turn reduces the demand placed on participants, and also allows us to implement some additional questionnaires.

> Specifically, at the end of the experiment, we will ask all participants to complete the Five-Dimensional Curiosity Scale Revised (5DCR; Kashdan, Disabato, Goodman, & McKnight, 2020). This scale is not central to our hypotheses, but may pave the way for exploratory analyses and future research.
>
> Additionally, we plan to incorporate a small set of supplementary questionnaires that index topics such as ADHD symptomology and mood. These supplementary questionnaires will be analysed by undergraduate students at Cardiff University as part of their final year research projects, but those analyses are not central to the research question described here, and they will *not* be reported in our Stage 2 submission. Because these questionnaires will only be administered after our task has otherwise concluded, their inclusion will not meaningfully influence our participants' behaviour along critical measures (such as exploration patterns and trial-by-trial curiosity).

In addition to the above, we have made several smaller changes in response to reviewers' specific comments, which are detailed below in our point-by-point responses. We again thank the reviewers and the editor for their thoughtful comments on our manuscript, and we hope these changes strengthen our resubmission.

Reviewer 1

The review is organized according to PCI RR reviewer guidelines. Some of the comments may not fully fit the section-headers/questions, but I added them just to try to aid the authors with the next version of the manuscript, should they and the editor decide to resubmit.

I) Does the research question (RQ) make sense in light of the theory or applications? Is it clearly defined? Where the proposal includes hypotheses, are the hypotheses capable of answering the research question?

Yes, the RQ makes perfect sense. The title, the abstract, and the introduction suggest that the study aims at elucidating the causal pathway between curiosity and memory. The authors set out to clarify whether the state of curiosity affects memory above and beyond its motivational effect active exploration. This would suggest that curiosity reflects a physiological state in which mnemonic encoding is more powerful or efficient. On the other hand, curiosity might enhance learning — not through priming the memory system for incoming information — but by energizing more thorough exploration. In this case, curiosity would only improve memory to the extent that the learner is able to act on it.

This seems to me like a perfectly valid research question, as I cannot convincingly argue against any of the two alternatives. The novel experimental treatment (passive group) is appropriately introduced to address the research question. However, hypotheses and the corresponding analysis approaches need to be developed a little more (see section III below).

One minor comment is to adjust the prioritization of study goals in the pre-registration (as well as the subsequent report). The manuscript often shifts the spotlight from the novel and interesting research question to replication of old results. It would be easier to assess the proposed hypotheses and analyses if they clearly and saliently addressed the main research question.

Response: Thank you very much for your feedback. We agree that our original framing disproportionately over-emphasized the replication relative to our novel research question, and we have reframed our Introduction accordingly. In particular, we now stress that our main theoretical interest is in the possible two-way interaction between experimental condition (Active vs. Passive) and curiosity.

In our revision, we have reframed our hypotheses in the following way:

> *"Among the Active Group, we expect to replicate Cen et al.'s (2024) findings (1) that curiosity predicts higher Path RE and (2) that curiosity and Path RE independently predict more accurate cognitive map formation. Additionally, and based on prior research surrounding the benefits of active learning (e.g., Craddock et al., 2011; Gaunet et al., 2001; Schomaker & Wittmann, 2021), we predict that participants in the Active Group will display better memory than participants in the Passive Group.*
>
> *However, our critical analysis concerns the influence of curiosity within the Passive Group, compared to the influence of curiosity within the Active Group. If curiosity reflects a general state of enhanced encoding, then we expect Passive Group participants to display better memory for rooms that they are more curious about. Conversely, if curiosity exerts its influence on memory because it facilitates specific patterns of exploration, then we expect that Passive Group participants' memory will not*

differ as a function of curiosity."

II) Is the protocol sufficiently detailed to enable replication by an expert in the field, and to close off sources of undisclosed procedural or analytic flexibility?

The protocol seems sufficiently detailed to enable replication by an expert in the field, but this job could be further simplified. The authors could specify their models in more detail to aid the reviewers' assessment. What priors are planned to be used and why? What are the likelihood functions? Are Likert-scale variables treated as continuous (if so, why?) or ordinal-categorical variables? What group would be considered the reference group? It would be great to see conventional formal model specifications (and perhaps some DAGs).

Response: Thank you for raising these points. We agree that our analyses could be more fully specified, and have updated our revised submission accordingly. Specifically:

- We plan to use weakly informative priors (centred on 0) because our predictions surrounding the magnitude and direction of the critical interaction (between experimental condition (Active/Passive Group) and curiosity) are nonspecific

- We plan to treat our Likert-scale variables as continuous. In Cen et al. (2024), we had originally wondered whether it might be preferable to treat these variables as ordinal, but found that doing so increased model run times without meaningfully changing the results.

III) Is there an exact mapping between the theory, hypotheses, sampling plan (e.g. power analysis, where applicable), preregistered statistical tests, and possible interpretations given different outcomes?

I would like to comment on several potential issues concerning the correspondence between the RQ, proposed hypotheses and methods of analysis.

Hypothesis 2 is simple and straightforward. However, the model addressing the hypothesis includes several variables serving the replication purpose rather than directly testing the hypothesis. I don't think it is necessarily wrong, but it would be helpful if the authors provided theoretical or logical justifications of including these variables in the model (that addresses hypothesis 2).

Response: This is certainly a valid consideration. When considering the main effect of experimental condition (Active vs. Passive) on memory [formerly 'Hypothesis 2'; now listed as Preliminary Research Question 3 (PRQ3) in our revised Study Design Template], we opted to include other variables (specifically: curiosity, interest, Path RE, and Head-Direction RE) as control variables so that we can more clearly interpret the resulting data. For instance, it is plausible that Active Group participants might report higher levels of curiosity and/or interest than Passive Group participants (e.g., because they might feel more engaged in the task), which could in turn manifest as a main effect of group in a simplified analysis (considering group as the sole predictor). Additionally, Head-Direction RE could potentially influence memory because it determines the variability in viewing angle that participants experience. By mean-centring all four of these variables and including them in our analysis, we hope to make stronger claims surrounding our active/passive manipulation. Specifically, we will assess whether participants with average levels of curiosity, interest, Path RE, and head-direction RE would still be expected to display differences in memory depending on whether they engaged in active vs.

passive exploration.

Paragraph 1B from section 2.1 of the PCI RR Guide for Reviewers explains that "The inclusion of hypotheses is not required– a Stage 1 RR can instead propose estimation or measurement of phenomena without expecting a specific observation or relationship between variables.". I believe that Hypothesis 3 of the manuscript, which the authors label as "nonspecific predictions", should be treated not as a planned hypothesis-testing procedure, but as an exploratory analysis, because there is no hypothesis (i.e., a theoretically derived prediction). The authors do provide possible interpretations of different outcomes, but these seem more like ad hoc rather than theoretically inspired explanations.

An alternative way of treating this problem is to list 3 different hypotheses and explain which modeling outcomes would support / refute each.

Should the authors and the editor find my assessment mistaken, it might help (to avoid this kind of issue in the future) to unambiguously list the hypotheses in one place in the manuscript and connect the modeling / analyses part more directly to that section by stating which results (e.g., a significantly positive coefficient of X on Y) would support or refute a given hypothesis. It would also be helpful to provide an interpretation of all coefficients of variables included in the models.

Response: Thank you for this suggestion. We agree that our key research question (whether curiosity directly influences memory, or whether that influence instead depends on active exploration) could potentially be framed as an exploratory analysis, given that our predictions are non-specific. However, through this analysis, we hope to distinguish between two competing theoretical perspectives (which we now detail more fully in our Introduction). Specifically, our Introduction now reads as follows:

> *"Interestingly, the benefits of curiosity appear to be wide-ranging. In at least some contexts (c.f., Hollins, Seabrooke, Inkster, Wills, & Mitchell, 2023; Keller, Salvi, Leiker, Gruber, & Dunsmoor, 2024), participants show enhanced memory not just for information that they are specifically curious about (e.g., trivia items), but also for incidental information encountered during high-curiosity states (e.g., faces presented during states of curiosity; Gruber et al., 2014; Murphy, Dehmelt, Yonelinas, Ranganath, & Gruber, 2021; see also Chen, Twomey, & Westermann, 2022). These benefits appear to be driven by curiosity-induced modulations within the dopaminergic circuit that bolster hippocampus-dependent memory (Gruber & Ranganath, 2019), suggesting that curiosity could index a general state of arousal in which encoding is more efficient (van Schijndel, Jansen, & Raijmakers, 2018).*
>
> *Alternatively, curiosity might be beneficial not because it directly promotes encoding (e.g., through dopaminergic modulation), but instead because it energizes specific patterns of exploration that are conducive to learning about particular contexts. For example, in his seminal work on cognitive map formation, Tolman (1948) argued that the latent maze learning he observed among rats was motivated by curiosity. That is, the animal's curiosity triggers exploration of a new environment, and that exploration promotes memory for the environment (see also Berlyne, 1966; O'Keefe & Nadel, 1978; Wang & Hayden, 2021)."*

Additionally, in our revised 'Statistical Analyses' section, we now more fully specify which results would support each theory:

> *"If (1) we obtain a significant main effect of pre-room curiosity [meaning that, among participants in the (dummy coded) Active Group, higher curiosity predicts better memory], and (2) we do not obtain any significant higher order interactions involving experimental condition, then we will conclude that the effect of pre-room curiosity does not meaningfully differ across experimental conditions: in other words, curiosity directly benefits memory even without the ability to actively engage in curiosity-guided exploration*.
>
> *Alternatively, if (1) any significant main effect of pre-room curiosity is qualified by a higher order term involving experimental condition, and (2) the nature of the interaction is such that pre-room curiosity has no effect (or a significantly weaker effect) within the Passive Group, relative to the Active Group, then we will conclude that the benefits of curiosity are (at least partially) dependent on the ability to actively engage in curiosity-guided exploration."*

Taking into account the above revisions, we have opted to frame our research question as a planned hypothesis testing procedure (but we are happy to reframe this if the editor and reviewers disagree with our assessment).

Power analysis

This is a minor comment. While the research question is clear, valid, and appealing, a several planning decisions seem to subserve the replication of a previous study (Cen et al., 2024), rather than addressing the answering the main question.

One example is that it seems like the reported power analysis was conducted to determine the sample size for a test that would replicate test results from Cen et al. (2024), but given the novel research questions that aim to expand on this work, focusing the power analysis on test replications seems a bit odd. Furthermore, it would be useful to read more details about how and why the specific power analysis technique works, especially given the plans to perform Bayesian parameter estimation (as opposed to null-hypothesis significance testing). The information provided about the power analysis is very limited, but from what is given, I am wondering why did the authors opt to use normal distributions rather than the posteriors (which are not guaranteed to be normal) from the previous study?

Response: Thank you for raising this point. Because our predictions surrounding the interaction between experimental condition and curiosity are nonspecific, we were unable to inform our analysis plan through a potential effect size. However, we agree that our sampling plan would be strengthened if it were focused on obtaining that critical effect, rather than replicating Cen et al. (2024). With this consideration in mind, we have adapted our sampling strategy to a modified sequential Bayes Factor design ("Sequential Bayes Factor with maximal *n*"; Schönbrodt & Wagenmakers, 2018). We plan to continue data collection *either* until we obtain compelling evidence for or against the critical interaction, *or* until our prespecified maximum sample size (*N* = 120*)* is met. Our priors, analysis points, and stopping criteria are now fully detailed in our revised Method and 'Statistical Analyses' sections, as below:

"Sample Size Determination. Our main theoretical interest is the possible two-way interaction between experimental condition (Active vs. Passive Group) and curiosity. However, because our predictions concerning the direction and magnitude of this interaction are nonspecific, we plan to determine our sample size in accordance with a modified Sequential Bayes Factor design ("Sequential Bayes Factor with maximal n"; Schönbrodt & Wagenmakers, 2018). Specifically, we will collect data either until we obtain compelling evidence for or against a two-way [experimental condition x curiosity] interaction [defined here as Bayes Factor (BF) > 10 or BF < .10, respectively], or until we reach a prespecified maximal sample size of 120 participants (N = 60 in each of the Active and Passive groups). If we reach this maximum without meeting our prespecified evidential thresholds, then the final BFs will still be interpreted, but our Stage 2 report will specify that our criteria for compelling evidence were not met.

To reduce the likelihood of obtaining false positive evidence for or against the target interaction (which could result from variability within small sample sizes), we will only begin analysing our data once we have obtained a minimum sample size of 30 participants (N = 15 per group). If our evidential criteria for the two-way interaction between experimental condition and curiosity are not met, then we will continue collecting data, stopping to reanalyse after every 6 participants (N = 3 per group) until one of our stopping criteria is reached. Importantly, when Bayesian stopping criteria are properly specified and adhered to, they do not carry the risks commonly associated with optional stopping (i.e., inflation of Type I error rates; Sanborn & Hills, 2014; Schönbrodt & Wagenmakers, 2018; Wagenmakers, Gronau, & Vandekerckhove, 2019)."

"All statistical analyses will be conducted in R (Version 4.2.2; R Core Team, 2023), with Bayesian multi-level models fit via the brms package (Bürkner, 2017, 2018). For each individual model (see below), we will use Bayesian estimation methods for parameter inference, with multiple chains (N = 4) and iterations (N = 4800) to promote convergence. Results will be interpreted according to the Bayes Factors (BFs) calculated from marginal likelihoods in accordance with our prespecified criteria for compelling evidence for or against a given effect (BF > 10 or BF < .10, respectively; see Method). To ensure robustness, we will conduct the same posterior predictive checks and sensitivity analyses as in Cen et al. (2024), including comparing model predictions to observed data as well as assessing the influence of alternate priors and model specifications. Across models, and because our predictions surrounding the magnitude and direction of critical effects are nonspecific, we will employ conservative priors centred on 0 for all main effects and interactions."

For proposals that test hypotheses, have the authors explained precisely which outcomes will confirm or disconfirm their predictions?

Yes.

Is the sample size sufficient to provide informative results?

To be determined by further planning.

Response: Please see above for details of our revised sampling plan.

Where the proposal involves statistical hypothesis testing, does the sampling plan for each hypothesis propose a realistic and well justified estimate of the effect size?

Probably. However, only a single power analysis has been conducted to determine the sample size for 80% power of a previously observed effect size (see section III-3 above).

Response: Please see above for details of our revised sampling plan.

Have the authors avoided the common pitfall of relying on conventional null hypothesis significance testing to conclude evidence of absence from null results? Where the authors intend to interpret a negative result as evidence that an effect is absent, have authors proposed an inferential method that is capable of drawing such a conclusion, such as Bayesian hypothesis testing or frequentist equivalence testing?

The main hypothesis (that curiosity enhances memory regardless of the ability to actively explore) can be supported by demonstrating the absence of an interaction between group (active vs passive) and curiosity. This would be tricky to demonstrate with a simple HDI-based test.

Response: Please see above for details of our revised sampling plan.

Have the authors minimised all discussion of post hoc exploratory analyses, apart from those that must be explained to justify specific design features?

Yes.

Have the authors clearly distinguished work that has already been done (e.g. preliminary studies and data analyses) from work yet to be done?

Yes.

Have the authors prespecified positive controls, manipulation checks or other data quality checks?

Yes, the authors plan to closely replicate the active-exploration group from a recent study by Cen et al. (2024) as a positive control for this study.

However, the authors do not seem to plan any manipulation checks for the novel treatment (passive viewing). They do acknowledge a potential limitation of this manipulation: participants can still exert active control (e.g., eye movements) in the passive-viewing condition, but no plans to address this limitations are provided. Nevertheless, this seems like a very minor issue since participants in the active group are also allowed to control their eye movements.

Response: Thank you for raising this point. As you note, participants in both the Active and Passive Groups have the opportunity to direct their visual exploration via eye movements. We consider this to be a promising direction for future research: for example, follow-up studies could measure participants' eye movements directly so as to assess whether patterns of visual attention differ across groups (we opted not to include eye-tracking here because, given the dynamic nature of the task, we would ideally want to purchase specialized headsets that would allow us to define dynamic regions-of-interest and streamline our analyses). However, before pursuing the more nuanced aspects of visual attention, we first hope to characterize the basic effect of curiosity: that is, whether Passive Group participants still benefit from their curiosity in

absence of the ability to direct their (physical) exploration through a given environment. If they do, then we could next ask whether visual attention (among other factors) might help explain the general effect of curiosity.

When proposing positive controls or other data quality checks that rely on inferential testing, have the authors included a statistical sampling plan that is sufficient in terms of statistical power or evidential strength?

See section III-3.

Response: Please see above for details of our revised sampling plan.

Does the proposed research fall within established ethical norms for its field? Regardless of whether the study has received ethical approval, have the authors adequately considered any ethical risks of the research?

The proposed research does not seem to involve any violations of ethical norms. However, I am no expert in ethics in cogsci research and the authors do not explicitly discuss any ethical considerations in the planned study.

Thank you for raising this point. We can confirm that our planned procedures have already been approved by the Cardiff University School of Psychology Research Ethics Committee, and we have clarified this point in our revised Methods section.

Reviewer 2

Review: O'Donoghue et al. – Disentangling the influences of curiosity and active exploration on environmental memory

This preregistered report is proposing to test the relationship between curiosity, active exploration, and spatial memory. In a replication and expansion of a study that is currently in a preprint, the authors propose to compare an active exploration group with a group that sees a video of the same exploration path. All participants give assessments of their curiosity before entering each room to explore. They are tested with a drawing task.

This preregistration is based on a preprint from the same lab. This makes assessment rather difficult because the preprint has not yet been vetted by peer review, and so some of the things I will point out here might come out in the review of the other manuscript, but might not. Some of my concerns are about the general methodology, the logic/rationale of the predicted hypotheses, and the theoretical basis of the study. Finally, there is a literature on active learning, on both spatial and non-spatial information, that would be useful for the authors to consider.

Response: Thank you for raising this point. Our lab's existing preprint (Cen et al., 2024) is currently under the first round of revision at *Communications Psychology* (first round of reviews received and responses submitted), and we will update our references to it upon publication. In the initial stages of planning the present study, we also considered the first round of reviewer comments that we received in response to Cen et al. (2024).

Introduction/Rationale:

1. From the abstract and the basic task description in the introduction, it sounds like it is just a straight active-passive manipulation without anything related to curiosity. It is not stated clearly in either place that people will also do the curiosity assessment. Being passive does not prevent someone from being curious, so it is important to clearly delineate how these factors are going to be tested.

Response: Thank you for raising this point. On reflection, we realize that our framing and task descriptions were not entirely clear (please see also our responses to Reviewer 1). In our revision, we have modified our Introduction and Methods sections to better clarify our novel research question. Critically, participants in both the Active and Passive Groups will self-report their curiosity prior to exploring each virtual room: as such, curiosity levels will vary within participants (across both groups), whereas the exploration condition will be manipulated between groups. This manipulation will allow us to disentangle the effects of curiosity and exploration, so that we can determine whether self-reported curiosity predicts memory within the Passive Group even in absence of the ability to engage in curiosity-driven exploration. Alternatively, if the benefits of curiosity depend on the exploration patterns that are energized by high-curiosity states, then we expect that higher Path RE will predict better memory regardless of group (Active vs. Passive), whereas, among the Passive Group participants, curiosity will not predict memory.

2. The theoretical basis of this study is not clearly laid out in the introduction. Are the authors hypothesizing that curiosity is a larger factor than active learning? Or vice versa? The

hypotheses for the passive group are non-specific, but I think more thinking through all the alternatives regarding the relative contributions of active learning and curiosity would help bring clarity to the possible outcomes.

Response: Our research question is not specifically focused on the relative influences of curiosity and active learning (please see also our above response). We agree that our framing of this matter was previously unclear, and we have modified our Introduction accordingly. Our active/passive manipulation is a tool to assess whether self-reported curiosity has a direct influence on memory – or instead, whether the benefits of curiosity depend on the ability to engage in active, curiosity-driven exploration. In other words: although past research supports that curiosity generally facilitates cognitive map formation, we do not yet know whether these benefits persist when learners do not have the ability to actively explore the environments that they are curious about. Using our yoked design, we will be able to disentangle the effects of curiosity itself (varying trial-by-trial across all participants, regardless of group) from the effects of exploration style (manipulated between the Active and Passive Groups).

3. Relatedly, in the design table, an alternative outcome is that there will be no group difference, and thus supporting the idea of curiosity. But this seems to go against the rationale for there being a group difference favoring active learning. The predicted group difference then seems to indicate that active learning is more important than curiosity. The logic here seems to be a bit off from the authors' overarching claims about curiosity.

Response: As above, we do not intend to directly investigate the relative benefits of active learning vs. curiosity; rather, our critical analysis concerns the (potential) interaction between curiosity and experimental condition (Active vs. Passive Group).

Based on past research, we do expect to observe a main effect of group – such that overall, Active Group (compared to Passive Group) participants display better memory for the novel rooms (this is now listed as 'PRQ3' in our revised Study Design Table). Among the Active Group, we also expect to observe a main effect of curiosity, such that higher curiosity facilitates better environmental memory ('PRQ2' in our revised Study Design Table). However, for the Passive Group, our predictions surrounding the effect of curiosity are nondirectional ('Critical Research Question' in our revised Study Design Table). If curiosity benefits memory regardless of the ability to engage in curiosity-driven exploration, then we expect to observe a main effect of curiosity (along with the hypothesized main effect of group), but we do not expect to observe a group x curiosity interaction. Conversely, if the benefits of curiosity depend on the ability to engage in active, curiosity-driven exploration, then we expect to observe an [experimental condition x curiosity] interaction such that curiosity does predict memory in the Active Group, but does not predict memory in the Passive Group.

4. There is already a literature on active learning and several studies that have examined active learning (especially making decisions about how to explore), both in a spatial context and in more general principles. Some of the authors questions may have been addressed already. Here are several references that are important and could aid adding to the theoretical basis of this study:

> Chrastil & Warren 2012 Active and passive contributions to spatial learning, Psychomomic Bulletin & Review

Chrastil & Warren 2015 Active and passive spatial learning in human navigation: Acquisition of graph knowledge

Gureckis & Markant 2012 Self-directed learning: A cognitive and computational perspective, Perspectives on Psychological Science

Markant & Gureckis 2014 Is it better to select or to receive? Learning via active and passive hypothesis testing, JEP: General

Voss et al. 2011 Hippocampal brain-network coordination during volitional exploratory behavior enhances learning, Nature Neuroscience

Response: Thank you for providing these references. In our revision, we have deemphasized our previous discussion of Active vs. Passive learning in effort to better focus on our critical research question surrounding the interaction between curiosity and experimental condition (please see comments above). However, we have made a note of these references and will be sure to incorporate them accordingly during the Stage 2 submission (when we will interpret the strength of the evidence for a main effect of experimental condition).

Methods:

1. For the power analysis, what is the effect/model that the authors are trying to test? What was the actual value of this effect? For example, is the simulation based on the correlation between path roaming entropy and curiosity? What about the between-subjects manipulation of active-passive groups? The previous study was within-subjects, so there needs to be a sample size justification for the between-subjects comparisons.

Response: Thank you for raising these concerns. We agree that our original power analysis was underspecified. In our revision, we have modified our sampling strategy so that, rather than powering according to our replication effort, we will determine our stopping criteria according to the strength of the evidence for or against the critical two-way interaction between experimental condition and curiosity (please see revised Method and responses to Reviewer 1).

2. Ratings of the drawings are the primary outcome measure of this experiment, but they are quite subjective. The ratings could also interact with drawing, even though participants were just told to put a box around an object – those instructions could actually make some of the spatial distortions and proportions less accurate (e.g., someone might just put the same sized box in the same orientation for everything if they did not understand of the instruction. There is not a lot of nuance in "draw simple boxes with labels in them").

Response: Thank you for raising this point. We have done our best to ensure that the drawing test instructions specify the need to accurate reflect object sizes and proportions, and that our raters' judgements depend on accuracy, rather than drawing ability. However, we agree that this is a very important consideration. Importantly, one of our evaluation criteria – Object Presence – is entirely independent of participants' drawing ability, because high scores along this dimension depend solely on whether layout-defining objects (e.g., windows, large furniture items) are appropriately labelled in the drawing. In Cen et al. (2024), we found that scores along Object Presence were highly correlated with scores along the remaining three dimensions, suggesting that participants' drawing abilities were not an important factor in raters'

evaluations.

3. Has the previous research examined whether any of the 16 rooms tend to have higher curiosity scores on their own? Are the rooms matched for basic information like the number of objects in them and how much space is possible to move in? This is not as relevant to the main question of active learning, but interesting on its own.

Response: This is certainly an important consideration. The rooms themselves are not matched for numbers of objects, but they are (generally) matched for object density (which in turn influences ease of movement). Additionally, Figure S4 of Cen et al., 2024 (available in the Supplemental Materials of that paper) supports that, in general, each room elicits a wide variety of curiosity ratings.

Importantly, participants always self-report their curiosity for a room based solely on its label (e.g., 'Lounge), and before seeing the room itself. Thus, and although some aspects of participants' exploration may by influenced by the particular characteristics of a given room (e.g., numbers of objects), the critical finding in Cen et al. (2024) -- that curiosity predicts both Path RE and cognitive map formation – cannot be explained by physical differences between the virtual environments themselves. This conclusion likewise applies to any effect of curiosity that we obtain in the planned study.

4. Is there eye tracking being done here? It doesn't sound like it, but the head is being tracked and gaze direction seems like it would be very helpful here.

Response: We apologize for the confusion here – our description of Head-Direction RE was not clear in our original submission, and in our revision, we have modified it accordingly. Here and in Cen et al. (2024), we did not track participants' physical (real-world) head movements: rather, by moving the mouse, participants will be able to direct the viewing angle (i.e., the 'head' orientation) of their onscreen virtual avatar, and these viewing angles will be used to calculate Head-Direction RE. Nevertheless, we agree that tracking real-world head and eye movements could prove a very promising direction for future research.

5. Since head direction is being tracked, list the equipment information for that tracking.

Response: Please see response above.

6. The passive group does not need to work with the controls, but please elaborate on the active group's training and familiarization procedures with the controls to ensure that using the controls does not distract the active group during learning.

Response: Thank you for raising this point. The Passive Group will also learn to operate the controls because – and although they will passively watch videos of each room – they will still actively navigate down the walkway to reach the target room on every trial (note that the walkway is a fixed path, thus severely restricting any potential for exploration, and that we will not analyse walkway data). We incorporated this active walkway movement to help retain task engagement within the Passive Group. Once Passive Group participants reach the end of the

walkway and press 'E' to enter the room, the video of a prior Active Group participant's exploration begins, and the Passive Group participant no longer needs to interface with the movement controls.

In our revision, we have elaborated upon the Active Group's training procedure (which is identical to Cen et al., 2024). On each of three training trials, Active Group participants will have the opportunity to familiarize themselves with the movement controls by exploring three virtual rooms (an empty room on Trial 1; fully-decorated rooms – a cinema and a bridal shop -- on Trials 2 and 3). Participants are encouraged to spend as much time in each room as they like, and in particular, to ensure that they are comfortable with the controls (or ask any questions they may have) before proceeding with the actual experiment. Additionally, the experimenter will check in with the participant to ensure comfort with the navigation controls at the end of the training phase.

Because of this training procedure, and because our controls are relatively simple (and standard across many video games, which young adults often have prior experience with – arrow keys to control physical movement, and mouse to control viewing angle), we expect any effect of using the controls themselves to be minimal.

7. Provide more rationale for why the curiosity score is mean-centered for each individual.

Response: Thank you for raising this consideration. We are collecting curiosity (and interest) measures using Likert-type scales, which are known to have both individual differences and cultural biases (e.g., Lee et al., 2002). Some individuals tend to cluster their responses around scale midpoints, while others are more likely to respond at the extremes (and these differences need not reflect 'true' individual or cultural differences along the variables being measured). Mean-centring participants' curiosity and interest ratings helps us account for these potential response biases by reframing each participant's ratings relative to their own baseline. We have now clarified this decision in our revised Analysis Plan.

8. For the second step of the analysis plan, Path RE and head direction RE are entered as predictors, but that should be the same across groups since the passive group is matched to the active. Say more on how this works.

Response: Thank you for noting this point; we have clarified how these predictors will be coded and interpreted in our revised Analysis Plan. Although both forms of RE are matched across group, we are interested in whether their effects on memory differ across the Active and Passive Groups (meaning that we will interpret any main effects of RE, as well as any interactions between group and RE). The potential interactions between group and RE are of particular theoretical interest. Cen et al. (2024) found that higher RE predicted better memory, and we expect to replicate this effect within the Active Group. Within the Passive Group, one could also predict that higher RE facilitates better memory (because more thorough exploration – even just passively watching – benefits memory), but one could alternatively predict that higher RE either has no effect on memory (perhaps because, in absence of high curiosity, more thorough exploration doesn't necessarily benefit memory), or even that higher RE might impair memory (perhaps because, if the presence of low curiosity, higher RE facilitates boredom and/or inattention).

In our analysis, group (Active vs. Passive) will be dummy-coded, such that the Active Group serves as our reference point. Thus, any main effects of Path and/or Head-Direction RE will be specific to the Active Group.

9. In the design table, for the second research question second column, it is not clear what "nondirectional" means in this context.

Response: Thank you for noting this issue. We have modified our wording throughout the revised pre-registration. By 'nondirectional', we mean that we do not have specific predictions concerning the magnitude and direction of the potential [experimental condition x curiosity] interaction. As detailed in our comments above, and although our predictions for the Active Group are fully specified (based on Cen et al., 2024), curiosity and Path RE could plausibly influence Passive Group participants' memory in several ways – either benefitting memory, having no effect, or even impairing memory.

Reviewer 3

The authors designed a study to understand the relationship between curiosity and environmental memory, and whether active exploration mediates it. Previous studies (Cen et al., 2024) found heightened curiosity about a virtual room to correlate with how widely active participants chose to explore it and their subsequent memory of elements from the virtual room. However, existing studies cannot confirm whether the observed memory benefits are due to curiosity itself or curiosity-induced patterns in exploration, operationalized as path roaming entropy (Path RE). The authors propose to address this question by replicating Cen et al. (2024) and extending the original design with a passive group. They expect to replicate findings from Cen et al. (2024) and record better memory in the Active, compared to the Passive group. The authors do not have specific predictions on the main effects of Path RE and curiosity on memory. Overall, the proposal presents a valid and interesting research plan. However, some clarifications are needed to ensure the soundness of the analyses. I also have concerns (see section 1E) about the conclusions it will be possible to draw from a yoked design.

1A. Scientific validity

While several studies have examined the relationship between free choice/exploration and memory, often with a yoked design, they did not examine the direct relationship between curiosity and memory benefits. Therefore, the research question addressed in the present paper is interesting, and the authors' attempt to address it is commendable.

Response: Thank you very much for this comment. We are happy to hear that the reviewer thinks that our research question is of interest.

1B. Logic, rationale, and plausibility of the proposed hypotheses (where a submission proposes hypotheses)

Overall, the authors' explanation of the research gap and hypotheses is clear, but the authors do not mention whether they have considered any possible effects of interactions between Path RE and curiosity on memory.

Moreover, hypothesis #3 could be split into two (e.g., #3a and #3b) to ease references to the effects of Path RE vs. curiosity since these are addressed separately in the study.

Response: Thank you for raising these points. Although not central to our main research question (which focuses on the two-way interaction between curiosity and experimental condition), we certainly agree that it would also be valuable to consider the potential two-way interaction between curiosity and Path RE. We have now included that interaction term in our Analysis Plan, and we do intend to interpret all terms included in our model (even those not central to our hypotheses). Relatedly, we also now plan to consider potential three-way interaction between group, curiosity, and path RE (the presence of which would inform our interpretation of any component two-way interactions). It is very possible that any two-way interaction between curiosity and Path RE might depend on group (passive vs. active), which would certainly be relevant to our research aims.

In our revision, we have also modified our Study Design Table to split up each of our component research questions.

1C. Soundness and feasibility of the methodology and analysis pipeline (including statistical power analysis or alternative sampling plans where applicable)

Overall, the methodology seems sound and feasible. However, the following points should be addressed.

The authors mention participants may be compensated with either course credits or monetary compensation. Does this mean some participants will be compensated in one way, and others in another? If so, different compensation schemas might affect participants' behavior in the experiment, interacting with the effects of interest. For example, monetary compensation, compared to course credits, might reduce the effects of curiosity. Do the authors plan to test/control for this? If not, why not?

Response: Thank you for raising this point; this is certainly a valid consideration. We plan to collect the vast majority of our data during teaching semesters in Cardiff University's academic year, in which case all participants will most likely be compensated with course credit. However, if we are unable to finish data collection during teaching semesters, we will also consider recruiting participants in exchange for monetary compensation (so that we do not need to pause data collection until the next academic year begins). Taking this plan into account, we expect the proportion of participants who receive monetary compensation to be very small, and thus, unlikely to impact our results. (although we agree that form of compensation is an important theoretical consideration). Regardless of compensation type, all participants will also be informed that their performance in the task has no bearing on the amount of compensation that they receive.

The authors state that condition assignment to either the active or passive group will be nonrandom. However, the authors do not mention any condition assignment matching based on specific demographics, compensation strategies, and whether the experiment is taking place in a room with other people or not (e.g., could 5 participants recruited individually be matched with 5 participants who did the experiment in a single large room), nor whether any such information about participants will be recorded.

Response: We agree that matched condition assignment would be a strength; as such, we have modified our assignment plan (see Method) to ensure that Active and Passive group participants are matched along two variables: gender (which past research supports can influence spatial memory) and compensation type (credit vs. monetary compensation). We do not plan to match participant pairs according to testing condition (alone vs. in a group) due to practical considerations [for example, if one participant in a group were to cancel on short notice, then another participant in that slot would require condition reassignment, which complicates the yoking protocol and introduces more room for experimenter error]. From Cen et al. (2024), we do not have a priori reasons to believe that these testing conditions meaningfully impact participants' behaviour.

1D. Clarity and degree of methodological detail

Some details are missing from the current methodology sections.

If the authors used a specific software to obtain the reported sample size, this should be appropriately cited.

Response: We have updated our sampling plan, which now depends on prespecified stopping criteria rather than a power analysis to determine sample size (please see responses to Reviewer 1).

Participants will be allowed to take a 5-minute break before taking the memory test. Differences in how people decide to use this time (e.g., by passively mind-wandering or scrolling through social media) might affect the results. The authors should clarify whether participants will be allowed to use their own devices during this time.

Response: Thank you for raising this important consideration. We agree that this is a complicating factor, and we updated our Method accordingly. Specifically, we now ask participants to complete a short distractor task (a series of math problems) during the 5-minute break.

Information about the specific instructions participants received should be reported, particularly regarding whether participants were aware of the memory test from the beginning of the experiment, and the exact phrase used to instruct participants in the Active vs. Passive group regarding the room exploration/viewing section of the task.

Response: We agree that these are important considerations and have clarified them in our revision. The specific instructions given to participants at the start of each phase are now available in the Supplemental Materials (please see the 'Supplemental Materials' folder at https://osf.io/8vpjg/). Participants will not be informed about the memory tests until they reach the Memory Test phase.

Information about curiosity and interest scales is inconsistent: the authors describe a 10-point scale on page 8 as well as the Figure 2 caption, but a 6-point scale on page 10.

Response: Thank you for catching this typo. Curiosity will be measured along a 10-point scale, and we have fixed this typo in our revision.

The https://map-scoring.vercel.app/link to view the coding system by Cen et al. (2024) requires credentials. Could the authors please provide instructions on how to access this?

Response: We have now created a dummy account that reviewers can use to access our scoring system: to sign in with it, please input 'scoring.system.reviewer@gmail.com' in the email field, and 'ReviewerAccount2024' in the password field.

1E. Sufficient outcome-neutral conditions (e.g. absence of floor or ceiling effects; positive controls; other quality checks) for ensuring that the obtained results are able to test the stated hypotheses or answer the stated research question(s)

A major issue in the current study design and analysis plan is that authors assume that active exploration is the only difference between Active and Passive conditions. However, several differences between the two groups could be introduced by the yoked design (see Gureckis & Markant, 2012). For instance, Passive participants will likely be less engaged in the task, potentially paying less attention to the screen and/or not looking at the scene in front of them. Participants in the Active condition may also be testing specific hypotheses or have particular expectations guiding their exploration, a process that would be disrupted in the Passive viewing condition. Broadly, Active participants would be more likely to have experiences directly connected to their internal stream of thoughts, which could facilitate the integration of new information within the existing knowledge base. The authors should explain how they will control for alternative explanations, clarify how they will adjust their interpretation of the results given this limitation, or address it with additional experimental conditions.

Response: Thank you for raising this point. We agree that various factors – including task engagement and specific patterns of attention / visual exploration – certainly might differ across the Active and Passive Groups; however, we do not necessarily perceive these factors to be confounds. Our key theoretical question is whether Passive Group participants will still benefit from their curiosity even in the absence of the ability to actively engage in curiosity-driven exploration. If they do, then that finding would suggest that curiosity could facilitate factors such as task engagement (e.g., Passive Group participants might be more engaged on high-curiosity trials than on low-curiosity trials, leading to curiosity-related memory benefits). If they do not, then that finding would suggest that curiosity does not necessarily facilitate factors such as task engagement if participants are solely engaged in passive exploration.

Because we do not yet know whether the benefits of curiosity depend on the ability to engage in curiosity-driven exploration, either result would be of theoretical interest. In either case, follow-up research could then help elucidate the precise mechanistic factors (e.g., differences in motivation vs. visual attention) that underlie our observed effects.

Minor comments and typos

It would be useful to define "environmental memory" as the authors intend it early in the manuscript, as the phrase might have different meanings in different fields or be new to the reader.

Response: Thank you for raising this point. We agree that this term was previously unclear; in our revision, and for consistency with Cen et al. (2024), we have updated all instances of 'environmental memory' to 'cognitive map formation'. Specifically, we are assessing the levels of accuracy and detail present in participants' cognitive maps of explored environments (as indexed by their room drawings).

Path RE and Head-Direction RE are not capitalized as in the rest of the manuscript on page 13.

Response: Thank you for catching this – we have corrected it in our revision.