MULTISENSORY CODING OF AUDIOVISUAL MOVIES IN THE HUMAN HIPPOCAMPUS

Omri Raccah
Department of Psychology
Wu Tsai Institute
Yale University
omri.raccah@yale.edu

Aryan Agarwal
Department of Psychology
Wu Tsai Institute
Yale University
aryan.agarwal@yale.edu

Yannan Zhu
Center for Cognitive Neuroscience
Institute for Brain Science
Duke University
yannan.zhu@duke.edu

Nicholas B. Turk-Browne
Department of Psychology
Wu Tsai Institute
Yale University
nicholas.turk-browne@yale.edu

March 1, 2026

ABSTRACT

The hippocampus receives convergent input from multiple sensory systems, yet in humans it has been studied almost exclusively through vision. Here we examine how the hippocampus contributes to sensory processing beyond the visual modality and to the multisensory integration of visual information with these other modalities. Participants were exposed to multiple repetitions of short naturalistic movie clips, each presented in four formats: auditory only, visual only, congruent audiovisual, and incongruent audiovisual (audio and video from different movies). Using high-resolution fMRI, we measured univariate activation and multivariate representations across the subfields and longitudinal axis of the human hippocampus. Whereas univariate analyses detected only visual responses across hippocampal subfields, with no activation for auditory stimuli and no benefit of congruent stimuli, multivariate analyses revealed robust representations of both auditory and visual scenes. The posterior hippocampus showed enhanced pattern similarity for congruent stimuli relative to unisensory stimuli, demonstrating multisensory facilitation. The anterior hippocampus showed crossmodal decoding between auditory and visual versions of the same clip, suggesting a more abstract representation. Finally, whole-brain searchlight analyses revealed parallel effects in cortical regions known to support multisensory integration. These findings advance understanding of auditory and multisensory coding in the human hippocampus, including the discovery of a functional dissociation along its longitudinal axis, from facilitation in posterior to generalization in anterior.

Introduction

The hippocampus and surrounding medial temporal lobe (MTL) cortex receive anatomical inputs from virtually all sensory systems [1]. In non-human animals, hippocampal neurons encode auditory, tactile, olfactory, and visual information, and represent crossmodal associations [2, 3, 4]. For example, macaque hippocampal neurons exhibit selectivity for congruent face and voice pairings of individual animals [4]. Yet, in humans, the hippocampus has been studied almost exclusively in the visual modality [5]. This leaves a fundamental gap: How does the human hippocampus represent sensory modalities beyond vision, and how does it integrate sense information across modalities? Understanding this integration is fundamental to explaining how the brain constructs coherent experiences and memories from fragmented inputs.

Multisensory learning enhances memory across species, from Drosophila to humans [6, 7, 8, 9]. However, whether hippocampal mechanisms support this mnemonic benefit remains unknown. Determining the contribution of the hippocampus to multisensory processing requires considering its broader representational capacity. Beyond its traditional role in memory, an emerging literature suggests that the human hippocampus contributes to central aspects of visual perception [10, 11], including eye movements [12, 13], imagery [14, 15], scene discrimination [16, 17], and visual attention [18, 19]. Given that the hippocampus represents complex conjunctions of visual features, it may additionally be critical for binding information across sensory modalities.

We tested this possibility for auditory and visual modalities. In line with recent approaches [11], we focused on characterizing representational coding schemes in the hippocampus, as opposed to testing a specific role in perception or memory behavior. Using high-resolution fMRI in 30 participants, we examined unisensory and multisensory processing in the whole hippocampus, along its longitudinal axis, and across its subfields. Participants were exposed to multiple repetitions of short naturalistic movie clips presented in four formats: auditory only, visual only, congruent audiovisual, and incongruent audiovisual (audio and video from different clips). We first conducted a general linear model (GLM) analysis to estimate univariate activation in each condition across hippocampal ROIs. Next, we performed representational similarity analysis (RSA) to investigate multivariate representations of unimodal and multimodal stimuli in these regions. We repeated these analyses across the whole brain to juxtapose effects in hippocampal ROIs with known cortical sites supporting multisensory integration.

This approach allows us to address three questions in this study: First, how is auditory information distributed in the hippocampus? Whereas auditory processing in the hippocampus has been widely studied in other species, evidence in humans remains limited (for review, see [5]). Only a handful of human fMRI studies have directly examined auditory processing in the human hippocampus [20, 21], and to our knowledge, none have mapped these representations across hippocampal subfields. Second, how does congruent multisensory stimulation facilitate processing in the hippocampus? Decades of research has demonstrated that congruent audiovisual stimuli enhance neural responses relative to unisensory stimuli, effects most prominent in the superior temporal sulcus (STS) and posterior superior temporal gyrus (posterior STG; e.g., [22, 23, 24, 25]). Whether the hippocampus exhibits similar multisensory facilitation effects remains unknown. Third, what is the relationship between auditory and visual representations in the hippocampus? The anterior subdivision of the hippocampus has been shown to maintain abstract (as opposed to perceptually detailed) representations in the visual modality [26, 27]. This raises the intriguing possibility that the anterior hippocampus may support generalization across sensory modalities that enables crossmodal transfer.

Results

Behavior

Participants first viewed 10 short movie clips in their original audiovisual format in the scanner prior to data acquisition (6 s each; Figure 1A). This pre-exposure was intended to limit memory demands in the main experiment by reducing novelty and learning, consistent with prior work that isolated representational content in the hippocampus [11].

During the main experiment, participants were exposed to multiple repetitions of these movie clips across four presentation formats: auditory only, visual only, congruent audiovisual, and incongruent audiovisual (mismatched audio and video from different clips; Figure 1B). All 4 formats of each of the 10 movies were presented once per run (40 stimuli per run) across nine runs, enabling leave-one-run-out cross-validation analyses. For the incongruent audiovisual condition, each video was paired once with audio from each of the other movies; these re-pairings varied across blocks to ensure that the visual and auditory content of each movie appeared as part of an incongruent trial once per block.

To ensure attention during the movies, participants were instructed to press a button within 1 s of each movie's offset (Figure 1B). Performance in this offset detection task was high across conditions (auditory: mean accuracy across participants, $M = 0.94$, standard error of the mean, $SEM = 0.006$, $P < 0.001$; visual: $M = 0.98$, $SEM = 0.006$, $P < 0.001$; congruent: $M = 0.98$, $SEM = 0.003$, $P < 0.001$; incongruent: $M = 0.99$, $SEM = 0.003$, $P < 0.001$). Response times (RTs) on correct trials were fast and consistent across conditions (auditory: $M = 379$ ms, $SEM = 13$; visual: $M = 370$ ms, $SEM = 13$; congruent: $M = 337$ ms, $SEM = 11$; incongruent: $M = 334$ ms, $SEM = 10$), confirming participants' engagement throughout the experiment.

Univariate activity

To examine univariate activity, we constructed a GLM with regressors for each condition. The resulting beta estimates were extracted from each ROI bilaterally and averaged across voxels per participant. We focused ROI analyses on the whole hippocampus, anterior and posterior hippocampus, and hippocampal subfields CA1, CA2/3, DG, and subiculum (Figure 1C). We corrected for multiple comparisons across anterior/posterior and across subfields.

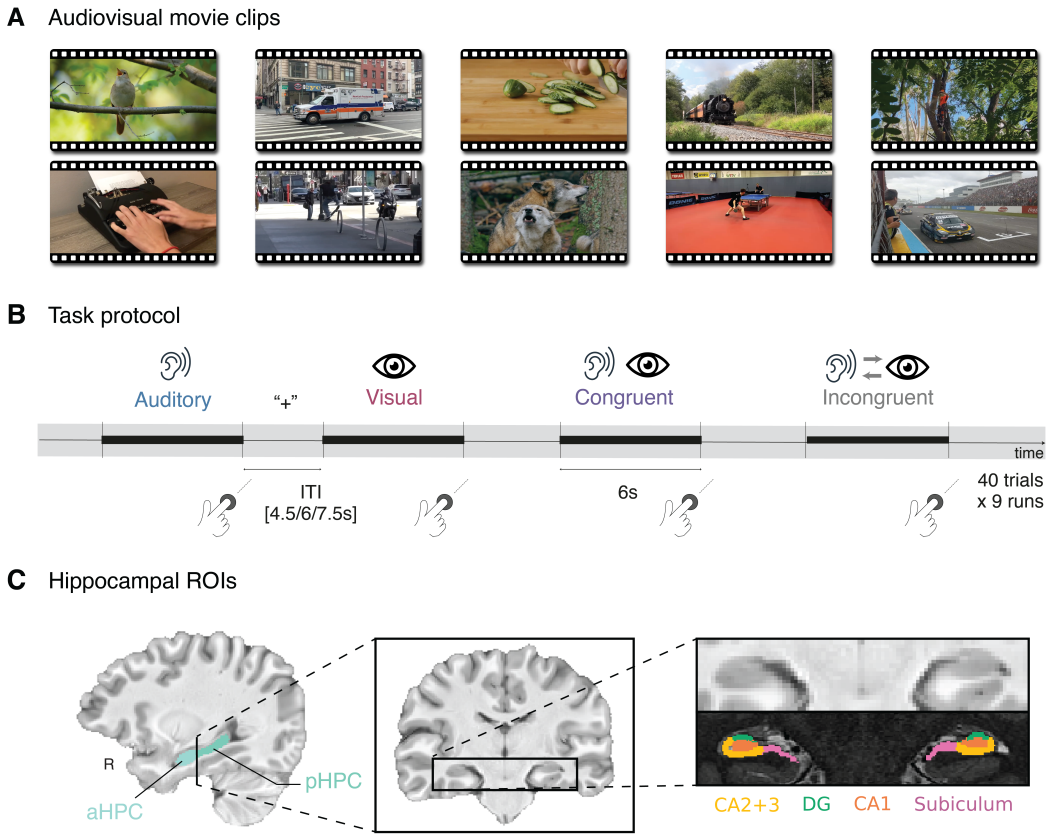

Figure 1: Experimental design and hippocampal segmentation. **(A)** The stimuli consisted of 10 naturalistic audiovisual movie clips (6 s each). **(B)** Unisensory stimuli were created by isolating the audio or visual track of each movie, resulting in 10 auditory only and 10 visual only stimuli. Incongruent audiovisual stimuli were generated by pairing audio and video tracks from different movies (90 unique pairings). In each run (9 runs total), all 10 stimuli from the auditory only, visual only, and congruent audiovisual conditions were presented, along with 10 incongruent audiovisual stimuli (40 total trials per run). Stimulus order was randomized within each run, with jittered inter-trial intervals between trials. As a cover task to ensure attention to the stimuli, participants were instructed to press a button within 1 s of each movie's offset. **(C)** The hippocampal long axis was manually segmented into posterior and anterior subdivisions based on the last coronal slice in which the uncus was visible [28]. Hippocampal subfields (CA2/3, DG, CA1, and subiculum) were segmented using the ASHS toolbox [29] trained on an atlas of manual 3T segmentations [30, 31].

The whole hippocampus (Figure 2A) showed significant evoked responses to unisensory visual stimuli ($P < 0.001$) and both multisensory audiovisual conditions (congruent: $P < 0.001$; incongruent: $P < 0.001$), but not to unisensory auditory stimuli ($P = 0.60$). This pattern was consistent in the anterior hippocampus (visual: $P < 0.001$; auditory: $P = 0.86$; congruent: $P < 0.001$; incongruent: $P = 0.001$) and posterior hippocampus (visual: $P < 0.001$; auditory: $P = 0.51$; congruent: $P < 0.001$; incongruent: $P < 0.001$), as well as in all hippocampal subfields (Figure 2B), including CA2/3 (visual: $P < 0.001$; auditory: $P = 0.53$; congruent: $P = 0.001$; incongruent: $P = 0.006$), DG (visual: $P < 0.001$; auditory: $P = 0.69$; congruent: $P < 0.001$; incongruent: $P = 0.001$), CA1 (visual: $P = 0.002$; auditory: $P = 0.69$; congruent: $P = 0.004$; incongruent: $P = 0.010$), and subiculum (visual: $P < 0.001$; auditory: $P = 0.53$; congruent: $P < 0.001$; incongruent: $P = 0.001$). We found no evidence of multisensory facilitation in univariate activity: the congruent audiovisual condition did not reliably exceed the visual condition in any hippocampal ROI (all $P > 0.05$). There was weak evidence of suppression for the incongruent relative to the congruent audiovisual conditions prior to FDR correction in the anterior hippocampus ($P = 0.027$, uncorrected) and DG ($P = 0.036$, uncorrected). For full descriptive statistics and results from MTL cortex, see Table S1 and Figure S1.

Next, we estimated group-level effects in the whole brain using voxelwise non-parametric randomization tests (TFCE corrected $P < 0.05$). A contrast of unisensory auditory versus visual conditions showed canonical activations in auditory

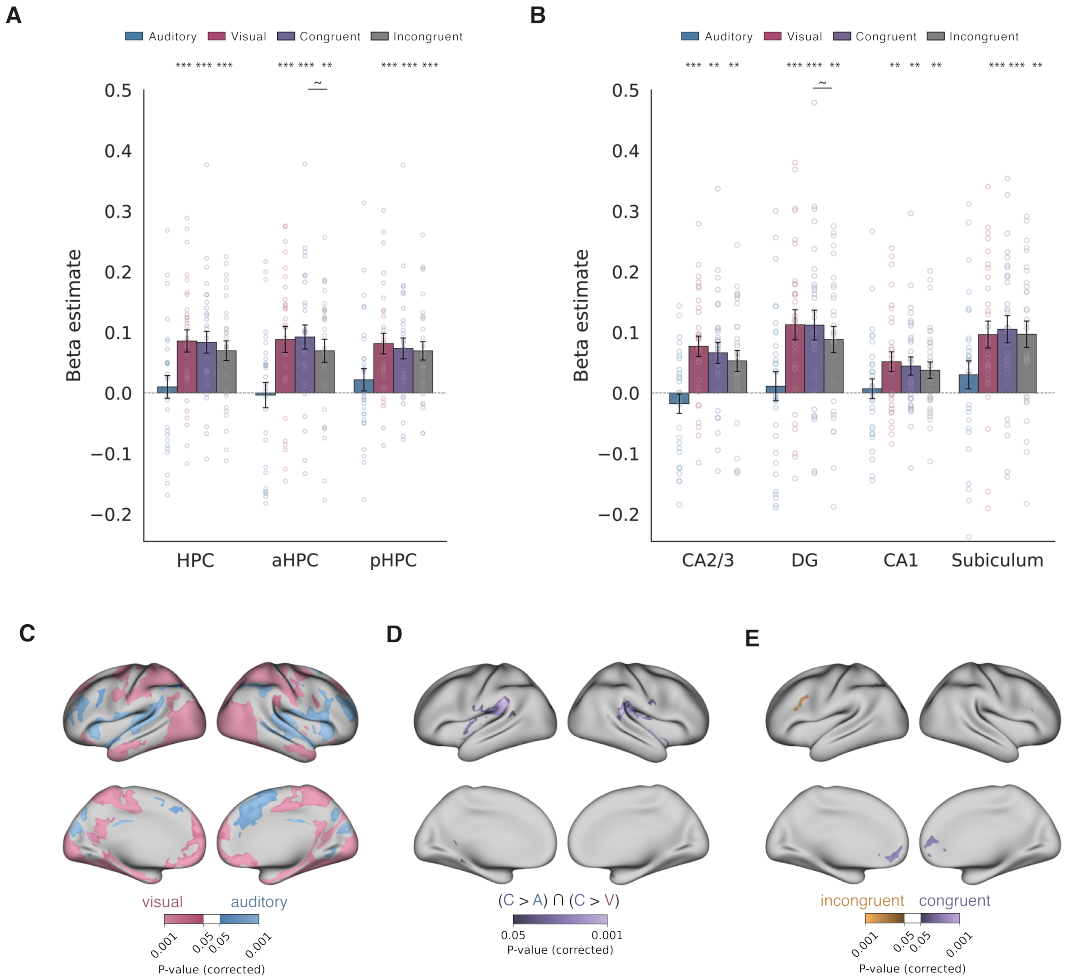

Figure 2: **Univariate activation across the hippocampus and whole brain.** **(A)** Mean beta estimates for each condition in the whole hippocampus and anterior/posterior subdivisions. **(B)** Mean beta estimates for each condition across hippocampal subfields (CA1, CA2/3, DG, subiculum). **(C)** Group-level whole-brain auditory only versus visual only contrast. **(D)** Conjunction analysis identifying regions in which congruent audiovisual activation exceeded both auditory only and visual only conditions. **(E)** Whole-brain contrast of congruent versus incongruent audiovisual conditions. Individual participant data shown as circles. Error bars indicate SEM. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, $\sim p < 0.05$ (uncorrected). Whole-brain maps were TFCE corrected at $P < 0.05$.

and visual cortices, respectively (Figure 2C). To assess multisensory facilitation, we performed a conjunctive analysis of voxels in which activation for the congruent audiovisual condition significantly exceeded both of the unisensory auditory and visual conditions. This contrast revealed activity in several regions, including STS, posterior STG, angular gyrus, and insula (Figure 2D). Finally, a contrast between congruent and incongruent audiovisual conditions revealed greater activation in bilateral medial prefrontal cortex for congruent and in left lateral prefrontal cortex for incongruent (Figure 2E). Together, these whole-brain effects validate the paradigm by replicating established patterns of unisensory and multisensory processing in the neocortex. Critically, while these univariate results confirm that the hippocampus is sensitive to the visual modality, they provide no evidence of sensitivity to the auditory modality or for multisensory facilitation.

Multivariate representations

We estimated single-trial activity patterns using GLMSingle ([32]) and performed RSA with a leave-one-run-out cross-validation approach. For a given ROI, we extracted voxel-wise beta coefficients for each movie of a condition in every run (Figure 3A). On each fold, we correlated the patterns from one held-out run with the average patterns from the other runs (Figure 3B), yielding a 10 (movie) by 10 (movie) similarity matrix for each condition. All correlation coefficients were Fisher z-transformed and the similarity matrices were averaged across folds. Diagonal elements corresponded to within-movie similarity (same movie across runs), while off-diagonal elements reflected between-movie similarity (different movies across runs). To quantify representational distinctiveness for each condition, we computed movie-specific pattern similarity (MSPS) as the signed difference between average within-movie and between-movie similarities (mean of diagonal cells minus mean of off-diagonal cells), with positive values indicating more distinctive neural representations for individual movie clips. MSPS was computed for each participant and statistical reliability for auditory, visual, and congruent conditions was calculated against zero (null hypothesis of no diagonal minus off-diagonal; i.e., that pattern similarity was not movie-specific).

We found reliable MSPS for auditory, visual, and congruent movie clips in a subset of hippocampal ROIs (Figure 3C,D). Unisensory auditory representations were observed in the whole hippocampus ($P = 0.014$), anterior hippocampus ($P = 0.006$), and subiculum ($P = 0.038$). Unisensory visual representations were observed in the posterior hippocampus ($P = 0.043$) and at an uncorrected threshold in the subiculum ($P = 0.049$, uncorrected). Congruent audiovisual representations were found in the whole hippocampus ($P < 0.001$), anterior hippocampus ($P = 0.017$), posterior hippocampus ($P < 0.001$), DG ($P = 0.021$), and at an uncorrected threshold in the subiculum ($P = 0.028$, uncorrected). The posterior hippocampus displayed multisensory facilitation, with enhanced MSPS for congruent compared to both visual ($P = 0.032$) and auditory conditions ($P = 0.003$). The whole hippocampus showed a similar but non-significant pattern of congruent facilitation relative to the visual condition ($P = 0.050$). We did not observe a facilitation effect in the anterior hippocampus nor in any hippocampal subfield (all $P > 0.05$). The full descriptive statistics and MTL cortex results are shown in Table S2, Table S3, and Figure S2.

Note that this multivariate analysis could not be performed straightforwardly in the incongruent condition because each run contained a unique audiovisual combination. We did, however, compare unisensory against incongruent patterns. While we did not find any significant differences in the hippocampus (Figure S4A-B), we observed a significant suppression of MSPS for the incongruent relative to unisensory conditions in the PHC (Figure S4C) and at the whole brain level (Figure S5).

To test whether MSPS was modality-specific or amodal, we cross-correlated the voxelwise patterns for auditory and visual presentations of the same versus different movies using the cross-validated leave-one-run-out RSA approach (Figure 4A). Reliable MSPS in this analysis would indicate the existence of movie-specific representations invariant to sensory modality. The whole hippocampus ($P = 0.006$) and the anterior hippocampus ($P = 0.017$) exhibited significant crossmodal transfer (Figure 4B), whereas the posterior hippocampus did not ($P = 0.20$). In hippocampal subfields (Figure 4C)), there was weak evidence of crossmodal transfer that did not survive correction in CA2/3 ($P = 0.045$, uncorrected) and subiculum ($P = 0.030$, uncorrected). See full descriptive statistics and crossmodal MTL cortex results in Table S3 and Figure S3.

The multivariate effects reported above could in principle benefit from memory retrieval, with repeated stimulus exposure allowing for reinstatement of movie-specific details or mental imagery of the absent modality. If so, we would expect that the magnitudes of these effects should increase as stimuli are repeated across cross-validation folds. However, neither multisensory facilitation in the posterior hippocampus (congruent audiovisual > visual MSPS; Spearman's $\rho = 0.048$, $P = 0.45$) nor crossmodal transfer in the anterior hippocampus ($\rho = 0.05$, $P = 0.41$) strengthened with more stimulus repetition across runs. This lack of correlation suggests that both phenomena were relatively stable and fails to support a retrieval-based account.

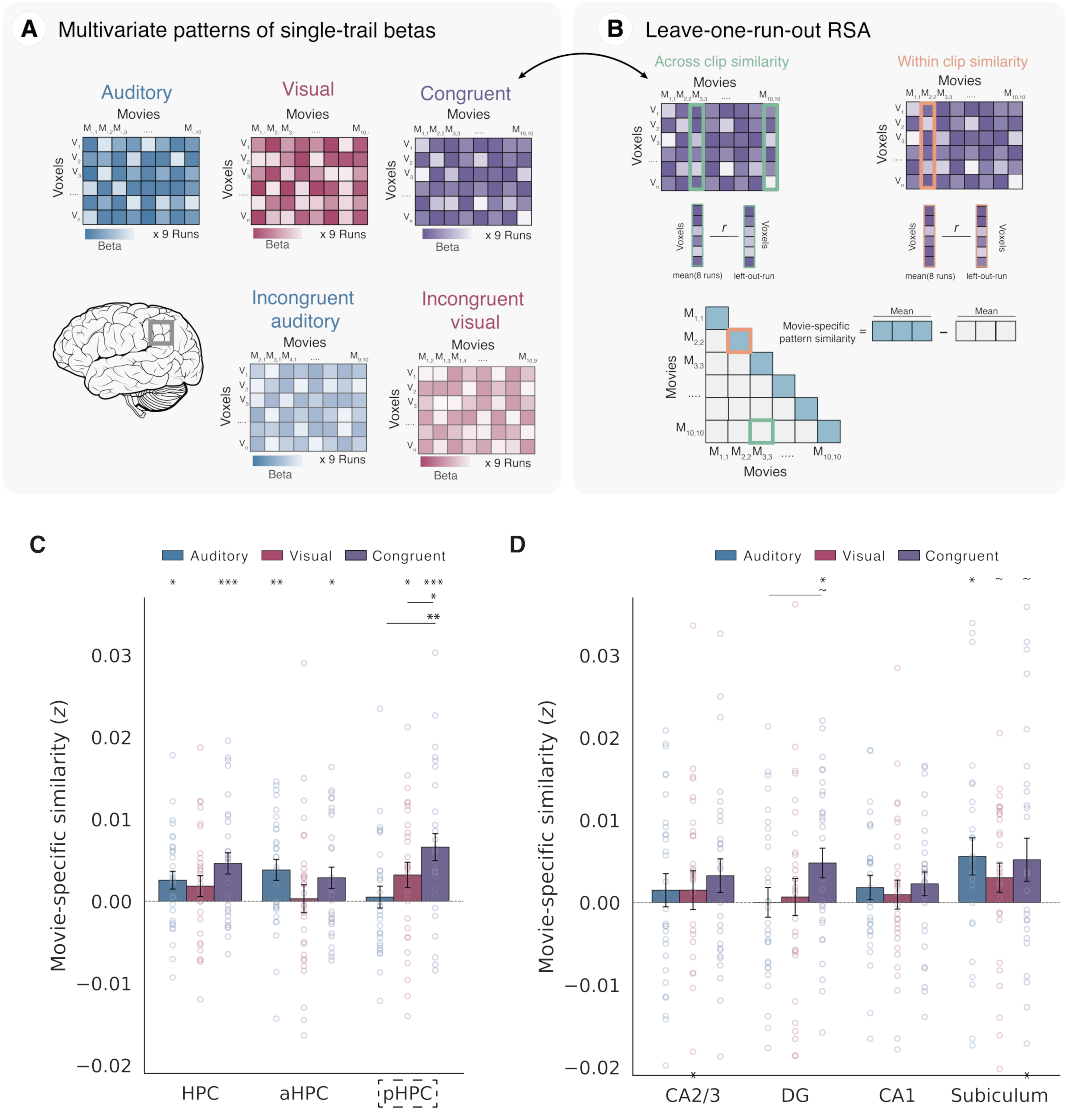

Figure 3: **Movie-specific multivariate representations in the hippocampus. (A)** To perform RSA, single-trial beta estimates were first extracted from a given ROI, generating multivariate activity patterns for each movie presentation. **(B)** We computed cross-validated RSA matrices independently for each condition using leave-one-run-out procedures (congruent audiovisual condition shown). Voxel-wise correlations were computed between movies (green) and within movies (orange) and averaged across cross-validation folds to construct the RSA matrix (bottom). To quantify movie-specific pattern similarity (MSPS), we computed the signed difference between mean diagonal (within-movie) and off-diagonal (between-movie) similarity values. **(C)** MSPS across conditions for the whole hippocampus (left) and anterior and posterior subdivisions (right). The posterior hippocampus showed enhanced pattern similarity for congruent audiovisual stimuli relative to unisensory conditions (grey hatched box). **(D)** MSPS across conditions across hippocampal subfields. Individual participant data shown as circles. Error bars indicate SEM. $\times$ markers indicate individual data points outside the plotted y-axis range. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; $\sim p < 0.05$ (uncorrected).

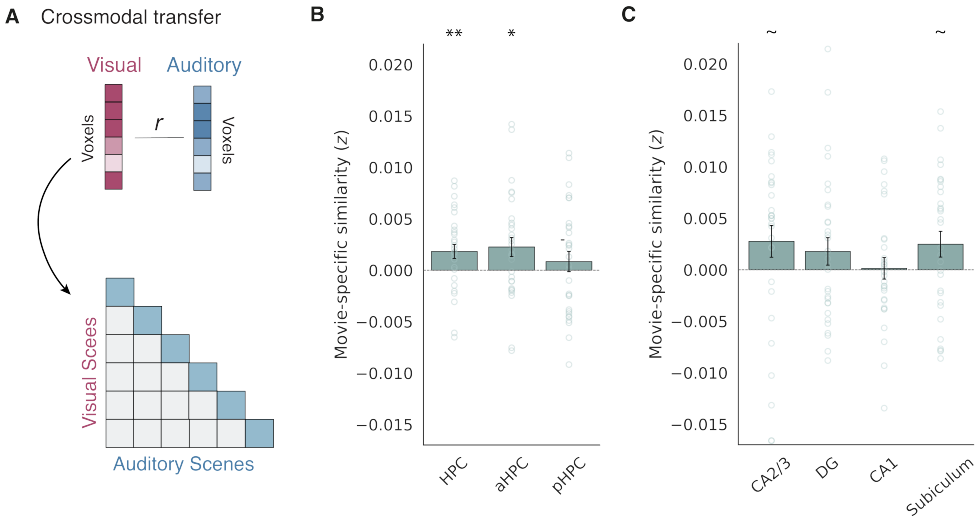

Figure 4: **Crossmodal transfer in the hippocampus. (A)** To estimate crossmodal transfer, we extended the cross-validated RSA approach by computing voxel-wise correlations between auditory only and visual only scene presentations. Movie-specific pattern similarity across modalities was quantified as the signed difference between same-movie (diagonal) and different-movie (off-diagonal) correlations. **(B)** Crossmodal MSPS in the whole hippocampus and anterior/posterior subdivisions. **(C)** Crossmodal MSPS across individual hippocampal subfields. Individual participant data shown as circles; error bars indicate SEM. $^{*}P < 0.05$, $^{**}P < 0.01$; $\sim P < 0.05$ (uncorrected).

To examine multivariate results beyond the hippocampus, we conducted an exploratory whole-brain searchlight analysis. In each participant, we applied the cross-validated leave-one-run-out RSA analysis within a cubic searchlight centered on each brain voxel and assigned the resulting MSPS value to that voxel (Figure 5A). We then assessed reliability across participants (TFCE corrected, $P < 0.05$). A contrast between unisensory auditory and visual conditions showed robust MSPS in auditory and visual cortices, respectively (Figure 5B). Next, we performed a conjunctive contrast to identify regions in which congruent audiovisual stimuli exceeded both unisensory conditions. This revealed localized MSPS in canonical multisensory sites, specifically bilateral STS and posterior superior temporal gyrus (pSTG; Figure 5C). Crossmodal transfer effects at the whole-brain level were similarly robust in STS and pSTG, as well as in higher-order visual areas and retrosplenial cortex (RSC; Figure 5D).

Discussion

This study focused on the representation of unisensory and multisensory content in the human hippocampus. We found evidence of auditory and visual coding, multisensory facilitation, and crossmodal transfer along the hippocampal long axis. Namely, visual processing was facilitated by the presence of a congruent auditory track in the posterior hippocampus, whereas visual and auditory versions of the same clip shared an abstract representation in the anterior hippocampus. These findings are consistent with a posterior-to-anterior gradient in the hippocampus from fine-grained to abstract representations, respectively [26, 27].

Auditory scene representation in the medial temporal lobe

We observed a clear distinction between analytical approaches: univariate analyses detected only visual responses across the hippocampus, with no activation for auditory stimuli, whereas multivariate pattern analyses revealed representations of auditory scenes. Specifically, there was significant movie-specific pattern similarity for auditory only stimuli in the whole hippocampus, anterior hippocampus, and subiculum. This suggests that auditory content can be encoded in distributed voxel-wise patterns and not readily captured in the mean activity across these voxels, underscoring the need for multivariate approaches to detect non-visual representations in the hippocampus.

Prior evidence for auditory coding in the human MTL is limited (cf. [33, 34]). Yet, an extensive body of work has established a role for the hippocampus in visual scene perception [10, 11]. Our findings provide initial evidence that the hippocampus may additionally support auditory scene perception in the absence of explicit memory demands, consistent with accounts of *domain-general* relational coding in the human hippocampus [35, 36]. Testing this hypothesis more

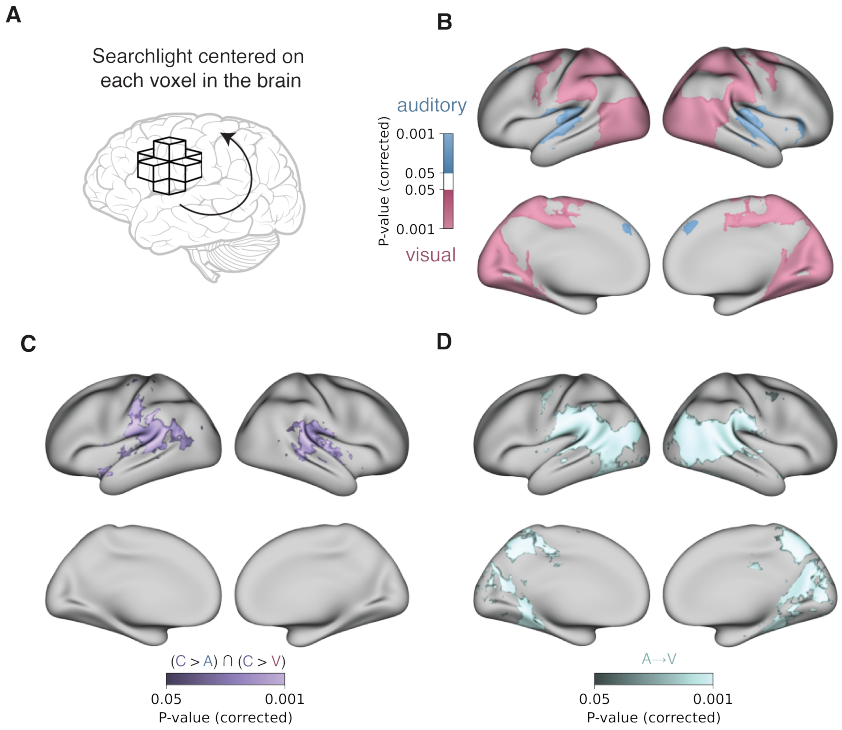

Figure 5: **Whole-brain searchlight RSA results. (A)** An exploratory searchlight was applied to estimate multivariate RSA effects in regions across the whole brain **(B)** Group-level auditory only versus visual only MSPS contrast. **(C)** Conjunctive map showing regions in which congruent audiovisual MSPS exceeded both auditory only and visual only conditions. **(D)** Above-baseline crossmodal transfer MSPS. All maps were TFCE corrected at $P < 0.05$.

directly will require establishing a preference in the hippocampus for auditory scenes over other forms of auditory stimuli, such as auditory objects or simple acoustic stimuli, extending experiments from the visual modality [16, 37]. It would further be interesting to test whether the hippocampus is specifically sensitive to the global structure or local details of auditory scenes, as has been examined in the visual modality [17].

We additionally observed robust auditory coding in PHC — a major input structure to the hippocampus via the entorhinal cortex. Although traditionally considered a visual region involved in place and scene processing, our findings align with recent evidence that the PHC responds to spoken descriptions of spatial information and landmarks [33]. Importantly, these effects may reflect crossmodal interactions, as evidenced by stronger visual effects and robust crossmodal generalization in our data. Nevertheless, these findings suggest a potential role for PHC in representing auditory "landscapes" that parallels its established function in visual scene processing.

Multisensory facilitation in the posterior hippocampus

We found enhanced movie-specific pattern similarity in the posterior hippocampus for congruent audiovisual stimuli relative to both unisensory modalities. Notably, pattern similarity was reliably above baseline only for visual and not auditory stimuli. Incongruent audiovisual stimuli did not show enhanced pattern similarity relative to unisensory conditions, further supporting the sensitivity of posterior hippocampus to multisensory congruence. We did not observe multisensory facilitation in the anterior hippocampus or in any individual hippocampal subfields. Together, we interpret these findings as facilitation of visual processing in the posterior hippocampus when a matching auditory track is present. This aligns with prior evidence that the posterior hippocampus exhibits stronger connectivity with sensory cortices and preferential coding of fine-grained visual features [38, 26, 39].

A whole-brain searchlight contrast between congruent audiovisual and unisensory conditions revealed localized effects in STS and posterior STG. This aligns with an extensive literature implicating these regions in audiovisual integration (e.g., [22, 23, 24, 25]). These findings situate the posterior hippocampus within a broader network supporting multisensory processing, raising questions about the distinct functional roles of these regions. For instance, does audiovisual

facilitation have different consequences for multisensory perception versus memory across these sites? Furthermore, the functional interaction between the hippocampus and these cortical sites represents an important direction for future research. The current event-related design precludes a rigorous analysis of the interactions across these regions, which could be addressed using longer-duration audiovisual stimuli to enable connectivity analyses.

Crossmodal generalization in the anterior hippocampus

Studies of multisensory perception interpret crossmodal transfer as reflecting "equivalence representations," reflecting abstracted, modality-invariant contents [40, 41, 42, 43]. The anterior (but not posterior) hippocampus exhibited crossmodal transfer, such that multivariate patterns belonging to the same scene's unisensory auditory and visual stimuli were more similar to each other than to patterns evoked by unisensory stimuli from different scenes. In the visual modality, the anterior aspects of the hippocampus have been implicated in representing abstract associations as opposed to perceptual details [38, 27]. Our whole-brain searchlight analysis found crossmodal transfer in the STS, pSTG, as well as high-level visual areas [44].

It is important to acknowledge that such effects can emerge from a range of shared contents across sensory modalities [40], including shared semantic content, low-level temporal coherence (e.g., energy fluctuations), or reinstatement of the absent modality via pattern completion or imagery. The present movie-based paradigm is not well-suited for adjudicating between these interpretations, but other paradigms that systematically manipulate stimulus features could be informative in future studies.

Conclusions

Despite decades of research on the human hippocampus, we lack a basic understanding of how this brain structure processes non-visual perceptual information and integrates across senses. By providing a comprehensive initial account of how multisensory coding is organized in the human hippocampus, this work provides foundation for bridging across perception, memory, language, and other domains. A critical next step will be to link these unisensory and multisensory representations to the formation of episodic memories.

Methods

Participants

Thirty-two young adults with normal or normal-to-corrected vision were recruited from Yale University and the New Haven communities. Two participants were excluded due to excessive head motion. The remaining participants comprised our final sample (N = 30; 16 female; mean age = 24.78 years, age range = 18.80–37.53). Although the effect sizes were unknown preventing a power analysis, this sample size is consistent with, or greater than, other studies of hippocampal function [45]. All participants provided informed consent and were compensated for their time ($25 per hour). The study was approved by the Yale University Institutional Review Board.

Stimuli generation

To generate the stimuli, we sourced 12 naturalistic audiovisual movies. Nine movies were from YouTube and three were from Pexels (`https://www.pexels.com`). In searching for stimuli, we prioritized movies that varied in content (animate/inanimate, indoor/outdoor, etc.) and that maximized audiovisual coherence (i.e., the visual and auditory components were clearly aligned to specific events). Only movies that contained high-fidelity video (1080p or 4K) and audio (48 kHz) were included. Movies containing legible speech were not considered. The final 12 movies included a ping-pong game, wolves howling, waves breaking, birds chirping, race cars passing by, and other multisensory scenes (Figure 1A).

Using Premiere Pro 2024 (Adobe Inc.), we cropped each movie into a 6-s segment. These segments were loudness normalized to the EBU R128 loudness standard with a target loudness of -23 LUFS. The movies were exported at 1280 x 720 resolution at 30 frames per second, with H.264 video codec and AAC audio codec. The audio was configured as stereo across all movies at 24 kHz.

From these processed movies, we selected two movies to serve as stimuli for the practice block, while the remaining 10 movies made up the experimental stimuli. To construct the task conditions, we used Premiere Pro to modify the movies. For the visual only condition, we extracted the video component while removing the audio. For the auditory only condition, we extracted the audio component while removing the video (Figure 1B). To create the incongruent condition, we re-paired the audio and visual components from different movies. This resulted in 90 possible incongruent

stimuli (10 videos × 9 non-matching audio tracks each). For the practice block, we created two incongruent stimuli using the same mismatching procedure. In total, the experimental stimulus set comprised 120 stimuli: 10 congruent audiovisual movies (original movies), 10 visual only, 10 auditory only, and 90 incongruent audiovisual movies. The practice block contained a total of 6 stimuli. Source URLs for all movies and the specific time segments extracted are available on the OSF project (`https://osf.io/jcg8a`).

Task design and procedures

The experiment consisted of 9 experimental runs, each containing 40 movie presentations: 10 congruent audiovisual, 10 auditory only, 10 visual only, and 10 incongruent audiovisual. All congruent, auditory only, and visual only stimuli were presented in every run. For the incongruent condition, each run contained all 10 video components and all 10 audio components, each paired with a mismatched audio or visual track. That is, each video was paired with audio from a different movie, with pairings varying across runs (e.g., ambulance video paired with ping-pong audio in one block, with waves audio in another, and so on). This ensured that the visual and auditory content from all movies appeared with equal frequency in every run while varying the specific incongruent pairings across runs, enabling leave-one-run-out cross-validation comparable to the other conditions. The order of stimuli in each run was pseudo-randomized by shuffling trials in each half of the run, ensuring an equal number of stimuli from all conditions in both halves. Finally, the order of the runs themselves was randomized across participants, to prevent order effects for the incongruent condition (the runs were all identical for the congruent, auditory only, and visual only conditions). In each run, the movies were shown sequentially with a jittered ITI (4.5, 6, and 7.5 s; Figure 1B). The onset of each movie was triggered by the start of a brain volume acquisition to ensure that any timing delays were accounted for and did not accumulate across the experimental session.

Participants performed a simple cover task while viewing the movies to ensure attention. They were instructed to press a button with their index finger within 1 s of each movie's offset (across all conditions). Visual feedback was provided after each trial, with a red fixation cross indicating an incorrect response (button press during the movie or >1 s after offset), while a green fixation indicated a correct response. At the end of each run, participants were shown their cover task accuracy (% correct responses). This paradigm was designed to maintain participants' attention throughout each movie while imposing similar task demands across conditions. Participants made all responses on an MRI-compatible button box.

In the scanner, and prior to the experimental runs, participants passively viewed the 10 movies in their original audiovisual form to gain familiarity with the auditory and visual components. This pre-exposure intended to reduce memory demands during the main experiment, consistent with prior work [11]. Participants then completed one practice run that mirrored the experimental procedures but used the two held-out practice stimuli presented in each format.

Audio delivery procedures

To deliver high-fidelity audio, we applied noise attenuation methods to overcome scanner noise. We used OptoActive inflatable headphones with active noise cancellation (OptoAcoustics, Mazor, Israel; `https://www.optoacoustics.com`), providing approximately 30 dB of noise suppression. Prior to the experimental session, we calibrated the active noise cancellation system based on the specific noise from the fMRI sequence while the participant was positioned in the scanner. Participants also wore semi-open earplugs, intended to provide an additional 10 dB of passive noise attenuation, while preserving stimulus transmission from the headphones. To maintain the spectral fidelity of auditory stimuli during active noise cancellation, we applied frequency-dependent equalization through the OptoAcoustics system. Following active noise cancellation calibration, we adjusted the volume level based on participants' feedback when listening to an audiovisual clip not included in the experiment. This volume level was held constant throughout the experimental session.

MRI acquisition

Data were acquired on a Siemens Prisma 3T scanner with a 64-channel volume head coil at BrainWorks in the Wu Tsai Institute at Yale University. Functional scans were collected using an echo-planar imaging (EPI) sequence with the following parameters: repetition time (TR) = 1,500 ms; echo time (TE) = 32 ms; 90 axial slices; voxel size = 1.5 × 1.5 × 1.5 mm; flip angle = 64°; multiband factor = 6. We also collected a pair of spin-echo scans with opposite phase-encoding directions for distortion correction (TR = 11,390 ms; TE = 66 ms). We acquired two anatomical scans: a high-resolution 3D T1-weighted (T1w) MPRAGE sequence (TR = 2,400 ms; TE = 2.19 ms; voxel size = 1 × 1 × 1 mm; 208 sagittal slices; flip angle = 8°), and a T2-weighted (T2w) turbo spin-echo sequence (TR = 11,170 ms; TE = 93 ms; 54 coronal slices; voxel size = 0.44 × 0.44 × 1.5 mm; distance factor = 20%; flip angle = 150°) positioned perpendicular

to the anterior-posterior axis of the hippocampus to assist with automated segmentation of medial temporal lobe (MTL) regions of interest (ROIs). These sequence parameters have been used in recent studies of the hippocampus [46, 45].

fMRI preprocessing

fMRI data were preprocessed using fMRIPrep version 23.2.1 [47], based on Nipype 1.8.6 [48]. The T1-weighted anatomical image was corrected for intensity non-uniformity using N4BiasFieldCorrection [49] distributed with ANTs 2.5.0 [50]. The corrected T1w image was skull-stripped using an implementation of the antsBrainExtraction.sh workflow. Brain tissue segmentation of cerebrospinal fluid (CSF), white matter (WM), and gray matter (GM) was performed on the brain-extracted T1w image using FAST in FSL ([51]). Brain surfaces were reconstructed using recon-all from FreeSurfer 7.3.2 [52]. Volume-based spatial normalization to standard space was performed through nonlinear registration with antsRegistration (ANTs 2.5.0) to the MNI152 standard template [53].

For each of the nine functional runs, slice-timing correction was applied. Head motion was corrected using mcflirt in FSL ([54]), and head-motion parameters (six rotation and translation parameters) were estimated for use as nuisance regressors. Susceptibility distortion correction was performed using a fieldmap-based approach, with field inhomogeneities estimated from spin-echo images with opposite phase-encoding directions using topup in FSL [55]. Fieldmaps were not acquired correctly for the first three participants and so susceptibility distortion correction was not applied in these cases. The corrected BOLD reference image was co-registered to the T1w reference using bbregister in FreeSurfer, which implements boundary-based registration [56] with six degrees of freedom.

We excluded trials in which framewise displacement exceeded 1.5 mm on any TR. Runs with more than two excluded trials were removed entirely, and participants with more than two excluded runs were removed from the dataset. This procedure resulted in the exclusion of two participants. In the remaining sample (N = 30), one run was excluded from each of two participants and 20 individual trials total were excluded across all participants.

Defining regions of interest

MTL structures were segmented from each participant's T1-weighted and T2-weighted anatomical images using the Automatic Segmentation of Hippocampal Subfields (ASHS; [29]) machine learning toolbox. The ASHS algorithm was trained on 24 manual segmentations of hippocampal subfields [31, 30]. This procedure generated participant-specific masks for hippocampal subfields CA1, CA2/3, and dentate gyrus (DG), as well as extrahippocampal MTL regions including entorhinal cortex (EC), perirhinal cortex (PRC), and parahippocampal cortex (PHC). Anterior and posterior hippocampal ROIs were manually defined for each participant on their native-space T1w image as the most posterior coronal slice in which the uncus was visible [26, 28].

Univariate GLM analysis

We constructed event-related general linear models (GLMs) using Nilearn version 0.9.2 [57]. For each participant and run, we included four condition regressors: auditory only, visual only, congruent audiovisual, and incongruent audiovisual. Each trial was modeled as a 6-s boxcar function aligned to stimulus onset and convolved with a canonical hemodynamic response function (HRF) using the Glover model with temporal and dispersion derivatives [58]. Additional regressors included six rigid-body motion parameters from fMRIPrep and nuisance regressors derived from high-variance voxels and high-pass filtering. The GLM was fit using an autoregressive AR(1) noise model with 5 mm FWHM spatial smoothing applied during estimation.

For ROI analyses, we extracted subject-specific beta estimates from each ROI and averaged across voxels within region. For whole-brain analyses, subject-level z-statistic maps were transformed from native T1w space to MNI space using ANTs [50] and non-parametric randomization tests were applied using randomise in FSL ([59]).

Representational similarity analyses

We estimated single-trial activity patterns using GLMSingle [32], which modeled each 6-s movie presentation and generated regularized beta estimates for each trial and voxel. GLMSingle applied voxel-wise HRF optimization. Furthermore, using GLMdenoise, this procedure automatically estimated nuisance regressors and executed noise correction [60]. Finally, regularization was performed to improve the final beta estimates.

For each ROI, we extracted single-trial beta estimates for each voxel, generating a multivariate activity pattern for each trial. For example, each of the 10 auditory only stimuli was presented once per run across 9 experimental runs, yielding 90 beta patterns for the auditory only condition (assuming no trial exclusions due to motion). Similarly, we obtained patterns each for the congruent audiovisual, visual only, and incongruent audiovisual conditions (Figure 3A).

For each condition, we computed a cross-validated representational similarity matrix using leave-one-run-out procedures (Figure 3B). We iteratively left out one run and correlated it with the average of the remaining eight runs, repeating this procedure across all cross-validation folds. All correlation coefficients were Fisher z-transformed before averaging across folds. To quantify pattern distinctiveness for each condition and participant, we computed the difference between the mean diagonal z-value (within-scene similarity) and the mean off-diagonal z-value (between-scene similarity). Higher values indicate more distinct multivariate representations for individual movies.

To assess crossmodal transfer, we applied the same RSA procedure but compared patterns between the auditory only and visual only conditions (Figure 4A). Specifically, we correlated multivariate patterns for the same scene presented in different modalities (diagonal elements) and for different scenes across modalities (off-diagonal elements). As before, we computed the difference between mean within-scene similarity (diagonal z-values) and mean between-scene similarity (off-diagonal z-values) to quantify crossmodal transfer within individual scenes.

Whole-brain searchlight analyses

For exploratory analyses across the whole brain, we performed the cross-validated RSA approach in searchlights centered on every brain voxel. The searchlight analysis was implemented using BrainIAK (Brain Imaging Analysis Kit; [61]) and parallelized across voxels using the Message Passing Interface (MPI). For each participant, we swept a cubic searchlight (radius = 3 voxels) throughout the whole-brain mask. Searchlights were required to contain at least 80% voxels within the participant's whole-brain mask to be included in the analysis. At each searchlight location, we performed our RSA procedure for each condition (auditory only, visual only, congruent audiovisual, and incongruent audiovisual) as well as for crossmodal transfer. We also computed participant-level conjunctive contrast maps to assess multisensory facilitation (congruent audiovisual greater both unisensory conditions) and the disruptive effect of audiovisual incongruence (incongruent audiovisual less than mean of unisensory conditions). The resulting correlation coefficient maps were Fisher z-transformed and aligned into MNI152 standard space using ANTs [50] with linear interpolation for group-level analysis. Across participants, statistical maps were computed for each condition and contrast using FSL randomise with threshold-free cluster enhancement (TFCE; [59]).

Statistical analyses

Statistical analyses were conducted in Python (version 3.7.9) using SciPy [62] (version 1.7.3), Statsmodels [63] (version 0.14.1), and custom code. All ROI tests used non-parametric sign-flipping randomization (10,000 iterations) with $\alpha = 0.05$. Univariate GLM effects were evaluated using two-tailed tests; RSA-based pattern similarity was evaluated using one-tailed (greater than) tests given an priori hypotheses of movie-specific representational distinctiveness. Post-hoc pairwise comparisons between conditions used one-tailed randomization tests on the within-participant differences. FDR correction for multiple comparisons was applied separately across anatomical divisions of the ROIs: anterior vs. posterior hippocampus, hippocampal subfields (CA1, CA2/3, DG, and Subiculum), and MTL cortex (PHC, PRC, and EC). For whole-brain analyses (GLM and RSA searchlight), we used randomise in FSL (10,000 iterations; [59]) with TFCE correction to generate statistical maps.

Data and code availability

The raw and preprocessed neuroimaging data will be released publicly upon acceptance. The task code is publicly available at `https://github.com/omriraccah/multimem_fMRI_expt`. The analysis code is publicly available at `https://github.com/omriraccah/multimem_fMRI_analysis`.

Author contributions

O.R. and N.B.T-B. conceived of the study. O.R., Y.Z., and N.B.T-B. designed the experiment. O.R. and Y.Z. implemented the experiment and collected the data. O.R. and A.A. analyzed the data with support from Y.Z. O.R. and N.B.T-B. wrote the manuscript. All authors edited the article. N.B.T-B. provided supervision for all aspects of the study.

Acknowledgments

O.R. was supported by a NIH Ruth L. Kirschstein Postdoctoral Individual National Research Service Award from the National Eye Institute (F32 EY035941). N.B.T-B. was supported by an NIH grant (R01 MH069456) and the Canadian Institute for Advanced Research.

We thank Roeland Hancock, Alexander Forrence, Irene Zhou, Erica Busch, Lillian Behm, and Juliana Trach for their help with data collection. We also thank all Turk-Browne lab members for their extensive feedback and support.

Supplementary Material

	Visual		Auditory		Congruent AV		Incongruent AV	
Region	M (SEM)	P	M (SEM)	P	M (SEM)	P	M (SEM)	P
			Hippocampus					
HPC	0.086 (0.018)	<**0.001**	0.010 (0.019)	0.60	0.083 (0.018)	<**0.001**	0.070 (0.016)	<**0.001**
aHPC	0.088 (0.021)	<**0.001**	–0.004 (0.021)	0.86	0.092 (0.020)	<**0.001**	0.070 (0.019)	**0.001**
pHPC	0.081 (0.017)	<**0.001**	0.022 (0.018)	0.51	0.073 (0.017)	<**0.001**	0.069 (0.015)	<**0.001**
			Hippocampal subfields					
CA1	0.052 (0.016)	**0.002**	0.007 (0.016)	0.69	0.044 (0.015)	**0.004**	0.038 (0.014)	**0.010**
CA2/3	0.077 (0.017)	<**0.001**	–0.018 (0.016)	0.53	0.066 (0.017)	**0.001**	0.053 (0.017)	**0.006**
DG	0.113 (0.025)	<**0.001**	0.011 (0.024)	0.69	0.112 (0.025)	<**0.001**	0.088 (0.022)	**0.001**
Subiculum	0.096 (0.022)	<**0.001**	0.030 (0.023)	0.53	0.105 (0.022)	<**0.001**	0.097 (0.022)	**0.001**
			MTL cortex					
PHC	0.297 (0.026)	<**0.001**	–0.060 (0.022)	**0.025**	0.318 (0.026)	<**0.001**	0.303 (0.022)	<**0.001**
PRC	0.062 (0.026)	**0.037**	–0.036 (0.024)	0.22	0.078 (0.025)	**0.005**	0.047 (0.026)	0.11
EC	0.039 (0.033)	0.25	–0.031 (0.035)	0.38	0.058 (0.036)	0.12	0.029 (0.037)	0.43

Table S1: **Univariate activation relative to baseline.** M = mean β coefficient across participants. SEM = standard error of the mean. P = FDR-corrected statistical significance, bolded if < 0.05.

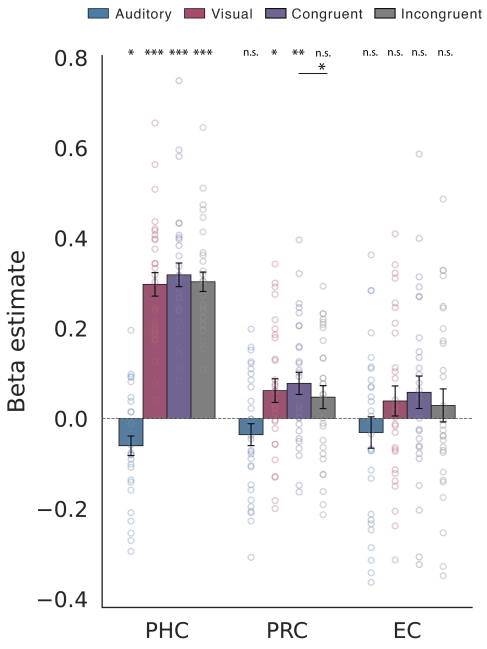

Figure S1: **Univariate activation in MTL cortical ROIs.** PHC showed a significant response to the visual only, congruent audiovisual, and incongruent audiovisual conditions, while displaying a below-baseline response for auditory only stimuli. PRC showed a significant response to the visual only and congruent audiovisual conditions, as well as a significant drop for the incongruent condition relative to congruent. EC showed no responses above baseline. Individual participant data shown as circles. Error bars indicate SEM. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

	Visual		Auditory		Congruent AV	
Region	M (SEM)	P	M (SEM)	P	M (SEM)	P
			Hippocampus			
HPC	0.002 (0.001)	0.085	0.003 (0.001)	**0.014**	0.005 (0.001)	<**0.001**
aHPC	0.000 (0.002)	0.43	0.004 (0.001)	**0.006**	0.003 (0.001)	**0.017**
pHPC	0.003 (0.002)	**0.043**	0.001 (0.001)	0.36	0.007 (0.002)	<**0.001**
			Hippocampal subfields			
CA1	0.001 (0.002)	0.38	0.002 (0.001)	0.23	0.002 (0.001)	0.065
CA2/3	0.002 (0.002)	0.38	0.002 (0.002)	0.30	0.003 (0.002)	0.065
DG	0.001 (0.002)	0.38	0.000 (0.002)	0.49	0.005 (0.002)	**0.021**
Subiculum	0.003 (0.002)	0.20	0.006 (0.002)	**0.038**	0.005 (0.003)	0.056
			MTL cortex			
PHC	0.045 (0.002)	<**0.001**	0.005 (0.002)	**0.005**	0.044 (0.003)	<**0.001**
PRC	0.002 (0.002)	0.16	0.002 (0.002)	0.39	0.002 (0.002)	0.19
EC	–0.000 (0.003)	0.57	–0.004 (0.003)	0.95	0.003 (0.003)	0.19

Table S2: **Multivariate pattern similarity relative to baseline.** M = mean movie-specific pattern similarity across participants. SEM = standard error of the mean. P = FDR-corrected statistical significance, bolded if < 0.05.

	Congruent > Visual		Congruent > Auditory		Crossmodal Transfer	
Region	ΔM (SEM)	P	ΔM (SEM)	P	M (SEM)	P
			Hippocampus			
HPC	0.003 (0.002)	0.050	0.002 (0.002)	0.13	0.002 (0.001)	**0.006**
aHPC	—	—	—	—	0.002 (0.001)	**0.017**
pHPC	0.003 (0.002)	**0.032**	0.006 (0.002)	**0.003**	0.001 (0.001)	0.20
			Hippocampal subfields			
CA1	0.001 (0.002)	0.29	0.000 (0.002)	0.40	0.000 (0.001)	0.45
CA2/3	0.002 (0.003)	0.29	0.002 (0.003)	0.40	0.003 (0.002)	0.089
DG	0.004 (0.003)	0.25	0.005 (0.002)	0.056	0.002 (0.001)	0.13
Subiculum	—	—	—	—	0.002 (0.001)	0.089
			MTL cortex			
PHC	–0.001 (0.002)	0.63	0.039 (0.003)	<**0.001**	0.009 (0.001)	<**0.001**
PRC	–0.000 (0.002)	0.63	0.001 (0.003)	0.44	0.001 (0.001)	0.24
EC	0.003 (0.005)	0.63	0.007 (0.004)	0.054	0.003 (0.002)	0.13

Table S3: **Multisensory effects.** Multivariate evidence for multisensory facilitation (congruent > visual; congruent > auditory) and crossmodal transfer. $(\Delta)M$ = (difference in) mean movie-specific pattern similarity across participants. SEM = standard error of the mean. P = FDR-corrected statistical significance, bolded if < 0.05. Facilitation was tested only in regions where congruent numerically exceeded both unisensory conditions (— = not tested).

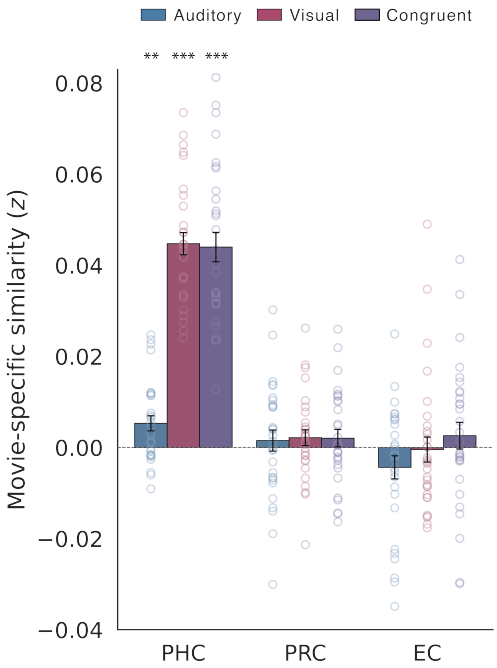

Figure S2: **Multivariate pattern similarity in MTL cortical ROIs.** PHC showed movie-specific pattern similarity for auditory only, visual only, and congruent audiovisual stimuli. There were no significant effects above baseline for PRC or EC. Individual participant data shown as circles; error bars indicate SEM. $^{**}p < 0.01$, $^{***}p < 0.001$.

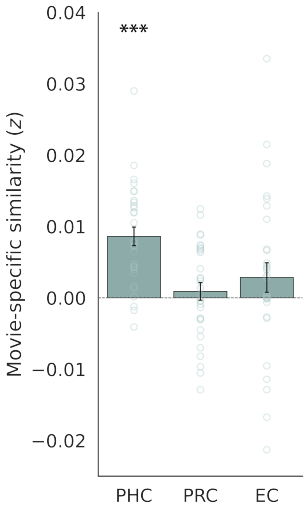

Figure S3: **Crossmodal transfer in MTL cortical ROIs.** PHC showed robust crossmodal transfer, whereas PRC and EC did not. Individual participant data shown as circles; error bars indicate SEM. $^{***}p < 0.001$

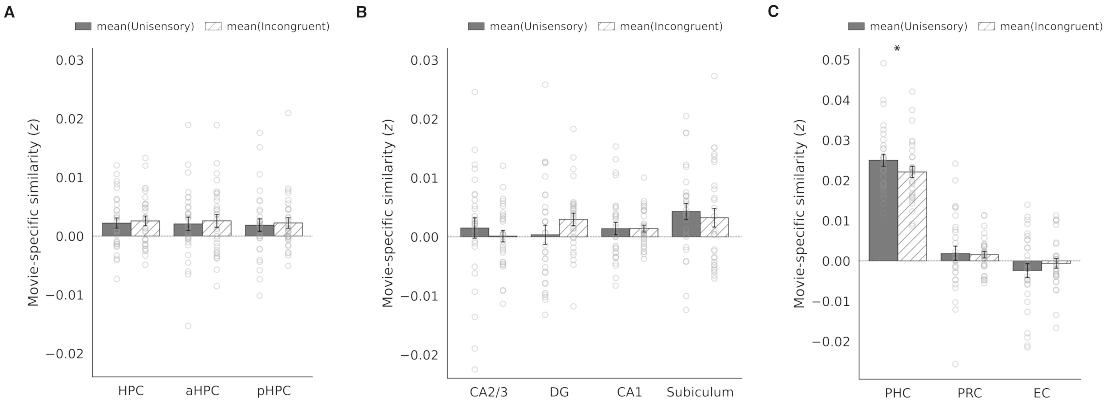

Figure S4: **Incongruent pattern similarity in the hippocampus.** No significant differences were observed between incongruent and unisensory (average of auditory and visual) MSPS **(A)** in the whole hippocampus and anterior/posterior hippocampus, or **(B)** in hippocampal subfields. **(C)** PHC showed greater MSPS for unisensory relative to incongruent conditions, suggesting a catastrophic interaction between non-matching auditory and visual content. There were no significant differences in PRC or EC. Individual participant data shown as circles; error bars indicate SEM. $^{*}p < 0.05$

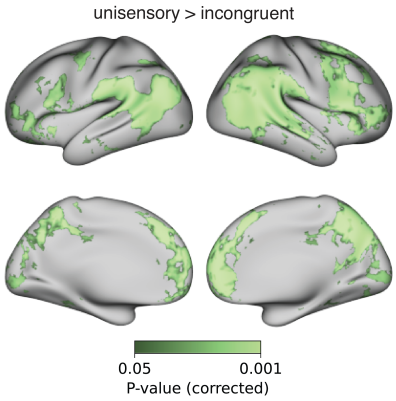

Figure S5: **Whole-brain searchlight of incongruent pattern similarity.** MSPS was significantly reduced for incongruent relative to unisensory (average of auditory and visual) throughout association cortex, including STS, posterior STG, angular gyrus, posteromedial cortex, and medial prefrontal cortex. Corrected using TFCE at $p < 0.05$.

References

[1] P. Lavenex and D. G. Amaral, "Hippocampal-neocortical interaction: A hierarchy of associativity," *Hippocampus*, vol. 10, no. 4, pp. 420–430, 2000.

[2] T. Li, A. Arleo, and D. Sheynikhovich, "Modeling place cells and grid cells in multi-compartment environments: Entorhinal–hippocampal loop as a multisensory integration circuit," *Neural Networks*, vol. 121, pp. 37–51, 2020.

[3] B. A. Radvansky, J. Y. Oh, J. R. Climer, and D. A. Dombeck, "Behavior determines the hippocampal spatial mapping of a multisensory environment," *Cell Reports*, vol. 36, no. 5, 2021.

[4] T. J. Tyree, M. Metke, and C. T. Miller, "Cross-modal representation of identity in the primate hippocampus," *Science*, vol. 382, no. 6669, pp. 417–423, 2023.

[5] A. J. Billig, M. Lad, W. Sedley, and T. D. Griffiths, "The hearing hippocampus," *Progress in Neurobiology*, vol. 218, p. 102326, 2022.

[6] Z. Okray, P. F. Jacob, C. Stern, K. Desmond, N. Otto, C. B. Talbot, P. Vargas-Gutierrez, and S. Waddell, "Multisensory learning binds neurons into a cross-modal memory engram," *Nature*, vol. 617, no. 7962, pp. 777–784, 2023.

[7] A. Thelen, P. J. Matusz, and M. M. Murray, "Multisensory context portends object memory," *Current Biology*, vol. 24, no. 16, pp. R734–R735, 2014.

[8] S. E. Duarte, A. P. Yonelinas, S. Ghetti, and J. J. Geng, "Multisensory processing impacts memory for objects and their sources," *Memory & Cognition*, vol. 53, no. 2, pp. 646–665, 2025.

[9] L. Shams and A. R. Seitz, "Benefits of multisensory learning," *Trends in Cognitive Sciences*, vol. 12, no. 11, pp. 411–417, 2008.

[10] N. B. Turk-Browne, "The hippocampus as a visual area organized by space and time: A spatiotemporal similarity hypothesis," *Vision Research*, vol. 165, pp. 123–130, 2019.

[11] C. B. Martin and M. D. Barense, "Perception and memory in the ventral visual stream and medial temporal lobe," *Annual Review of Vision Science*, vol. 9, no. 1, pp. 409–434, 2023.

[12] M. Nau, T. Navarro Schröder, J. L. Bellmund, and C. F. Doeller, "Hexadirectional coding of visual space in human entorhinal cortex," *Nature neuroscience*, vol. 21, no. 2, pp. 188–190, 2018.

[13] D. E. Hannula and C. Ranganath, "The eyes have it: hippocampal activity predicts expression of memory in eye movements," *Neuron*, vol. 63, no. 5, pp. 592–599, 2009.

[14] A. E. Comrie, L. M. Frank, and K. Kay, "Imagination as a fundamental function of the hippocampus," *Philosophical Transactions of the Royal Society B*, vol. 377, no. 1866, p. 20210336, 2022.

[15] D. Hassabis, D. Kumaran, S. D. Vann, and E. A. Maguire, "Patients with hippocampal amnesia cannot imagine new experiences," *Proceedings of the National Academy of Sciences*, vol. 104, no. 5, pp. 1726–1731, 2007.

[16] C. J. Hodgetts, N. L. Voets, A. G. Thomas, S. Clare, A. D. Lawrence, and K. S. Graham, "Ultra-high-field fmri reveals a role for the subiculum in scene perceptual discrimination," *Journal of Neuroscience*, vol. 37, no. 12, pp. 3150–3159, 2017.

[17] M. Aly, C. Ranganath, and A. P. Yonelinas, "Detecting changes in scenes: The hippocampus is critical for strength-based perception," *Neuron*, vol. 78, no. 6, pp. 1127–1137, 2013.

[18] M. M. Chun and E. A. Phelps, "Memory deficits for implicit contextual information in amnesic subjects with hippocampal damage," *Nature neuroscience*, vol. 2, no. 9, pp. 844–847, 1999.

[19] M. G. Stokes, K. Atherton, E. Z. Patai, and A. C. Nobre, "Long-term memory prepares neural activity for perception," *Proceedings of the National Academy of Sciences*, vol. 109, no. 6, pp. E360–E367, 2012.

[20] S. Kumar, H. M. Bonnici, S. Teki, T. R. Agus, D. Pressnitzer, E. A. Maguire, and T. D. Griffiths, "Representations of specific acoustic patterns in the auditory cortex and hippocampus," *Proceedings of the Royal Society B: Biological Sciences*, vol. 281, no. 1791, p. 20141000, 2014.

[21] S. Kumar, S. Joseph, P. E. Gander, N. Barascud, A. R. Halpern, and T. D. Griffiths, "A brain system for auditory working memory," *Journal of Neuroscience*, vol. 36, no. 16, pp. 4492–4505, 2016.

[22] F. Setti, G. Handjaras, D. Bottari, A. Leo, M. Diano, V. Bruno, C. Tinti, L. Cecchetti, F. Garbarini, P. Pietrini, *et al.*, "A modality-independent proto-organization of human multisensory areas," *Nature Human Behaviour*, vol. 7, no. 3, pp. 397–410, 2023.

[23] M. S. Beauchamp, B. D. Argall, J. Bodurka, J. H. Duyn, and A. Martin, "Unraveling multisensory integration: patchy organization within human sts multisensory cortex," *Nature neuroscience*, vol. 7, no. 11, pp. 1190–1192, 2004.

[24] Y. Hu and Y. Mohsenzadeh, "Neural processing of naturalistic audiovisual events in space and time," *Communications Biology*, vol. 8, no. 1, p. 110, 2025.

[25] S. N. Pushpita and L. Wehbe, "Two cortical mechanisms of audiovisual processing in the human brain," *bioRxiv*, 2025.

[26] J. Poppenk, H. R. Evensmoen, M. Moscovitch, and L. Nadel, "Long-axis specialization of the human hippocampus," *Trends in Cognitive Sciences*, vol. 17, no. 5, pp. 230–240, 2013.

[27] S. H. Collin, B. Milivojevic, and C. F. Doeller, "Memory hierarchies map onto the hippocampal long axis in humans," *Nature neuroscience*, vol. 18, no. 11, pp. 1562–1564, 2015.

[28] O. Warrington, N. N. Graedel, M. F. Callaghan, and P. Kok, "Communication of perceptual predictions from the hippocampus to the deep layers of the parahippocampal cortex," *Science Advances*, vol. 11, no. 21, p. eads4970, 2025.

[29] P. A. Yushkevich, J. B. Pluta, H. Wang, L. Xie, S.-L. Ding, E. C. Gertje, L. Mancuso, D. Kliot, S. R. Das, and D. A. Wolk, "Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment," *Human Brain Mapping*, vol. 36, no. 1, pp. 258–287, 2015.

[30] M. Aly and N. B. Turk-Browne, "Attention stabilizes representations in the human hippocampus," *Cerebral Cortex*, vol. 26, no. 2, pp. 783–796, 2016.

[31] N. C. Hindy, F. Y. Ng, and N. B. Turk-Browne, "Linking pattern completion in the hippocampus to predictive coding in visual cortex," *Nature neuroscience*, vol. 19, no. 5, pp. 665–667, 2016.

[32] J. S. Prince, I. Charest, J. W. Kurzawski, J. A. Pyles, M. J. Tarr, and K. N. Kay, "Improving the accuracy of single-trial fmri response estimates using glmsingle," *eLife*, vol. 11, p. e77599, 2022.

[33] C. O. Häusler, S. B. Eickhoff, and M. Hanke, "Processing of visual and non-visual naturalistic spatial information in the" parahippocampal place area"," *Scientific Data*, vol. 9, no. 1, p. 147, 2022.

[34] A. Y. Li, N. Ladyka-Wojcik, H. Qazilbash, A. Golestani, D. B. Walther, C. B. Martin, and M. D. Barense, "Experience transforms crossmodal object representations in the anterior temporal lobes," *eLife*, vol. 13, p. e83382, 2024.

[35] L. Davachi, "Item, context and relational episodic encoding in humans," *Current Opinion in Neurobiology*, vol. 16, no. 6, pp. 693–700, 2006.

[36] C. Ranganath and M. Ritchey, "Two cortical systems for memory-guided behaviour," *Nature Reviews Neuroscience*, vol. 13, no. 10, pp. 713–726, 2012.

[37] C. J. Hodgetts, J. Shine, A. Lawrence, P. Downing, and K. Graham, "Evidencing a place for the hippocampus within the core scene processing network," *Human Brain Mapping*, vol. 37, no. 11, pp. 3779–3794, 2016.

[38] J. Robin and M. Moscovitch, "Details, gist and schema: hippocampal–neocortical interactions underlying recent and remote episodic and spatial memory," *Current Opinion in Behavioral Sciences*, vol. 17, pp. 114–123, 2017.

[39] M. B. Bone and B. R. Buchsbaum, "Detailed episodic memory depends on concurrent reactivation of basic visual features within the posterior hippocampus and early visual cortex," *Cerebral Cortex Communications*, vol. 2, no. 3, p. tgab045, 2021.

[40] D. J. Lewkowicz and G. Turkewitz, "Cross-modal equivalence in early infancy: Auditory–visual intensity matching.," *Developmental Psychology*, vol. 16, no. 6, p. 597, 1980.

[41] A. Mroczko-Wąsowicz and R. Grush, *Sensory individuals: Unimodal and multimodal perspectives*. Oxford University Press, 2023.

[42] A. Amedi, K. von Kriegstein, N. M. van Atteveldt, M. Beauchamp, and M. J. Naumer, "Functional imaging of human crossmodal identification and object recognition," *Experimental Brain Research*, vol. 166, no. 3, pp. 559–571, 2005.

[43] G. A. Calvert, M. J. Brammer, and S. D. Iversen, "Crossmodal identification," *Trends in Cognitive Sciences*, vol. 2, no. 7, pp. 247–253, 1998.

[44] M. S. Beauchamp, "See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex," *Current Opinion in Neurobiology*, vol. 15, no. 2, pp. 145–153, 2005.

[45] J. Wammes, K. A. Norman, and N. Turk-Browne, "Increasing stimulus similarity drives nonmonotonic representational change in hippocampus," *eLife*, vol. 11, p. e68344, 2022.

[46] B. E. Sherman and N. B. Turk-Browne, "Statistical prediction of the future impairs episodic encoding of the present," *Proceedings of the National Academy of Sciences*, vol. 117, no. 37, pp. 22760–22770, 2020.

[47] O. Esteban, C. J. Markiewicz, R. W. Blair, C. A. Moodie, A. I. Isik, A. Erramuzpe, J. D. Kent, M. Goncalves, E. DuPre, M. Snyder, *et al.*, "fmriprep: a robust preprocessing pipeline for functional mri," *Nature methods*, vol. 16, no. 1, pp. 111–116, 2019.

[48] K. Gorgolewski, C. D. Burns, C. Madison, D. Clark, Y. O. Halchenko, M. L. Waskom, and S. S. Ghosh, "Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python," *Frontiers in neuroinformatics*, vol. 5, p. 13, 2011.

[49] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4itk: improved n3 bias correction," *IEEE transactions on medical imaging*, vol. 29, no. 6, pp. 1310–1320, 2010.

[50] B. B. Avants, N. Tustison, G. Song, *et al.*, "Advanced normalization tools (ants)," *Insight j*, vol. 2, no. 365, pp. 1–35, 2009.

[51] S. M. Smith, M. Jenkinson, M. W. Woolrich, C. F. Beckmann, T. E. Behrens, H. Johansen-Berg, P. R. Bannister, M. De Luca, I. Drobnjak, D. E. Flitney, *et al.*, "Advances in functional and structural mr image analysis and implementation as fsl," *Neuroimage*, vol. 23, pp. S208–S219, 2004.

[52] B. Fischl, "Freesurfer," *Neuroimage*, vol. 62, no. 2, pp. 774–781, 2012.

[53] V. S. Fonov, A. C. Evans, R. C. McKinstry, C. R. Almli, and D. Collins, "Unbiased nonlinear average age-appropriate brain templates from birth to adulthood," *NeuroImage*, vol. 47, p. S102, 2009.

[54] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, "Improved optimization for the robust and accurate linear registration and motion correction of brain images," *Neuroimage*, vol. 17, no. 2, pp. 825–841, 2002.

[55] J. L. Andersson, S. Skare, and J. Ashburner, "How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging," *Neuroimage*, vol. 20, no. 2, pp. 870–888, 2003.

[56] D. N. Greve and B. Fischl, "Accurate and robust brain image alignment using boundary-based registration," *Neuroimage*, vol. 48, no. 1, pp. 63–72, 2009.

[57] A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, and G. Varoquaux, "Machine learning for neuroimaging with scikit-learn," *Frontiers in neuroinformatics*, vol. 8, p. 14, 2014.

[58] G. H. Glover, "Deconvolution of impulse response in event-related bold fmri1," *Neuroimage*, vol. 9, no. 4, pp. 416–429, 1999.

[59] A. M. Winkler, G. R. Ridgway, M. A. Webster, S. M. Smith, and T. E. Nichols, "Permutation inference for the general linear model," *Neuroimage*, vol. 92, pp. 381–397, 2014.

[60] K. N. Kay, A. Rokem, J. Winawer, R. F. Dougherty, and B. A. Wandell, "Glmdenoise: a fast, automated technique for denoising task-based fmri data," *Frontiers in neuroscience*, vol. 7, p. 247, 2013.

[61] M. Kumar, M. J. Anderson, J. W. Antony, C. Baldassano, P. P. Brooks, M. B. Cai, P.-H. C. Chen, C. T. Ellis, G. Henselman-Petrusek, D. Huberdeau, *et al.*, "Brainiak: the brain imaging analysis kit," *Aperture neuro*, vol. 1, no. 4, pp. 10–52294, 2022.

[62] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, *et al.*, "Scipy 1.0: fundamental algorithms for scientific computing in python," *Nature methods*, vol. 17, no. 3, pp. 261–272, 2020.

[63] S. Seabold, J. Perktold, *et al.*, "Statsmodels: econometric and statistical modeling with python.," *SciPy*, vol. 7, no. 1, pp. 92–96, 2010.