Received: 2 September 2022 | Revised: 20 December 2022 | Accepted: 18 January 2023

DOI: 10.1002/hbm.26219

RESEARCH ARTICLE

WILEY

The neural representations underlying asymmetric cross-modal prediction of words

Liang Shi[1] | Chuqi Liu[1] | Xiaojing Peng[1] | Yifei Cao[1] | Daniel A. Levy[2] | Gui Xue[1]

[1]State Key Laboratory of Cognitive Neuroscience and Learning and IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, People's Republic of China

[2]Baruch Ivcher School of Psychology, Interdisciplinary Center Herzliya, Herzliya, Israel

Correspondence
Gui Xue, State Key Laboratory of Cognitive Neuroscience and Learning and IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing 100875, People's Republic of China.
Email: gxue@bnu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Number: 31730038; NSFC and Israel Science Foundation (ISF) joint project, Grant/Award Number: 31861143040; Sino-German Collaborative Research Project "Crossmodal Learning", Grant/Award Number: NSFC 62061136001/DFG TRR169

Abstract

Cross-modal prediction serves a crucial adaptive role in the multisensory world, yet the neural mechanisms underlying this prediction are poorly understood. The present study addressed this important question by combining a novel audiovisual sequence memory task, functional magnetic resonance imaging (fMRI), and multivariate neural representational analyses. Our behavioral results revealed a reliable asymmetric cross-modal predictive effect, with a stronger prediction from visual to auditory (VA) modality than auditory to visual (AV) modality. Mirroring the behavioral pattern, we found the superior parietal lobe (SPL) showed higher pattern similarity for VA than AV pairs, and the strength of the predictive coding in the SPL was positively correlated with the behavioral predictive effect in the VA condition. Representational connectivity analyses further revealed that the SPL mediated the neural pathway from the visual to the auditory cortex in the VA condition but was not involved in the auditory to visual cortex pathway in the AV condition. Direct neural pathways within the unimodal regions were found for the visual-to-visual and auditory-to-auditory predictions. Together, these results provide novel insights into the neural mechanisms underlying cross-modal sequence prediction.

KEYWORDS
cross-modal prediction, fMRI, representation, sequence memory, word sequence

1 | INTRODUCTION

Making predictions in a multisensory world is critical for human survival, for example, the sound of a horn predicting an incoming car or thunder following lightning. Indeed, neocortical perceptual processes have been posited to be essentially multisensory (Ghazanfar & Schroeder, 2006). Supporting this view, distributed brain regions are found to be activated in cross-modal associative learning and memory, including primary visual cortex and somatosensory cortex (Pillai et al., 2013; Zhou & Fuster, 2000), frontoparietal regions (Ku et al., 2015; Tanabe et al., 2005; Tibon et al., 2019; Zhang et al., 2004), and the hippocampus (Borders et al., 2017; Butler & James, 2011). For example, higher-order association areas exhibited greater activation during the formation of audiovisual associations than visual–visual associations (Tanabe et al., 2005). Furthermore, recalling information from cross-modal associations elicited greater hippocampal activity than unimodal associations (Butler & James, 2011). Nevertheless, the neural representations of these

regions under cross-modal sequence prediction are still barely understood.

Focusing on unimodal sequence prediction, it has been found that learning a fixed sequence of events enables the prediction of an upcoming event, expressed as faster reaction times (Hsieh et al., 2014). The prefrontal cortex (Jenkins & Ranganath, 2010; Yokoi & Diedrichsen, 2019) and the hippocampus (Turk-Browne et al., 2010) have been found to support unimodal sequence prediction, showing strong activation during the retrieval of learned sequences (Lehn et al., 2009; Ross et al., 2009).

The predictive coding theory has posited that making predictions about a future event is accomplished by pre-activating its associated representations (Bar, 2007; Clark, 2013; Friston, 2010), which then shape our perception of future stimuli (de Lange et al., 2018). Supporting this view, predictive representation reinstatement after associative learning has been found in visual (Ekman et al., 2017; Kok et al., 2017; Reddy et al., 2015; Senoussi et al., 2020), auditory (Demarchi et al., 2019), and multimodal (e.g., movie) domains (Lee et al., 2021; Paz et al., 2010). Another study found that after learning the temporal regularity of successive events, the neural representation of the preceding items became increasingly similar to the following ones, but not vice versa (Schapiro et al., 2012). However, how the brain implements such predictive mechanisms to perform cross-modal sequence prediction has not been elucidated.

Another question concerns the neural pathways underlying cross-modal sequence prediction. Two candidate neural pathways have been proposed for cross-modal processing, including direct pathways between modality-specific regions and indirect neural pathways via higher-order areas (Arnal et al., 2009; Driver & Noesselt, 2008). Anatomical studies have found direct white matter connections and projections between primary sensory areas (Beer et al., 2011, 2013; Bieler et al., 2017), which might underlie the cross-modal interaction in modality-specific regions (Kayser & Logothetis, 2007). In addition, some studies have found feedback influence from multisensory convergence zones to sensory cortices (Macaluso et al., 2000; Macaluso & Driver, 2005). How these neural pathways are differentially involved in within-modal and cross-modal predictions remain to be examined.

To address these questions, the present study employed a novel audiovisual sequence memory paradigm, in which participants were asked to perform pre- and post-learning semantic judgment tasks and in between to learn word sequences that were presented in unimodal or cross-modal formats. Behavioral results revealed a reliable asymmetric cross-modal predictive effect, with a stronger prediction from visual to auditory than auditory to visual. Using high-resolution functional magnetic resonance imaging (fMRI) and multivariate pattern similarity analyses, we found this asymmetric predictive effect was supported by an indirect neural pathway from the visual to the auditory cortex via SPL, but no such indirect pathway was found from the auditory to the visual cortex. These results provide novel insights into our understanding of the neural mechanisms of cross-modal prediction.

2 | MATERIALS AND METHODS

2.1 | Participants

Twenty-one healthy college students (10 males, mean age = 22.4 years, range = 18–26 years) participated in the fMRI study (Exp 1). Another 37 college students (4 males, mean age = 20.9 years, range = 18–25 years) participated in the behavioral study (Exp 2). The sample size of the fMRI study was comparable with that of several previous studies using a similar paradigm (Bellmund et al., 2019; Hsieh et al., 2014). The sample size of the behavioral study was determined by the power analysis conducted in the Gpower toolbox, which indicated that 36 subjects could achieve 0.95 statistical power to detect a small effect size (0.25) at the significance level of 0.05. All participants were right-handed, had a normal or corrected-to-normal vision, and had no psychiatric or neurological disease history. Informed written consent was obtained from all participants after a full explanation of the study procedure. This study was approved by the Institutional Review Boards of the Center for MRI Research at Peking University and the State Key Laboratory of Cognitive Neuroscience and Learning at Beijing Normal University.

2.2 | Materials

A total of 80 two-character Chinese nouns were used in the experiment, including 40 words describing animate objects (e.g., dog) and 40 words describing inanimate objects from five subcategories (8 words from each subcategory), including fruits (e.g., banana), tools (e.g., knife), transports (e.g., train), musical instruments (e.g., guitar), and household appliances (e.g., refrigerator) (Table S1). These stimuli were divided into four groups of 20 words, each containing 10 words describing living objects and 10 words describing non-living objects. They were pseudo-randomly assigned to the four modality (visual vs. auditory) and sequence type (fixed vs. random) combinations. The assignment was counterbalanced across participants. The visual words were presented in white color on a black background, with a width of 250 pixels and a height of 125 pixels. For the auditory words, the voice was generated by the Text-to-speech (TTS) software Balabolka (http://balabolka.site/balabolka.htm), where a female speaker pronounced the words. Each word lasts 2 s with an intensity of 75 dB.

We constructed eight types of five-word sequences, including V-V-V-V-V, V-V-A-A-V, V-A-A-V-V, V-A-V-A-V, A-A-A-A-A, A-A-V-V-A, A-V-V-A-A, A-V-A-V-A (V refers to visual and A to auditory modality). For the fixed sequences, the same words were assigned to each sequence in the same order across repetitions. Whereas for the random sequences, across repetitions, the same words were presented in a different order while obeying the sequence structure (Figure 1a). Particularly, the presentation modality of each word does not change across repetitions for any sequence. With this design, we could effectively match the fixed and random sequences in terms of sequence structure and the familiarity of the materials. Although the words in later positions of the random sequence sequences are more

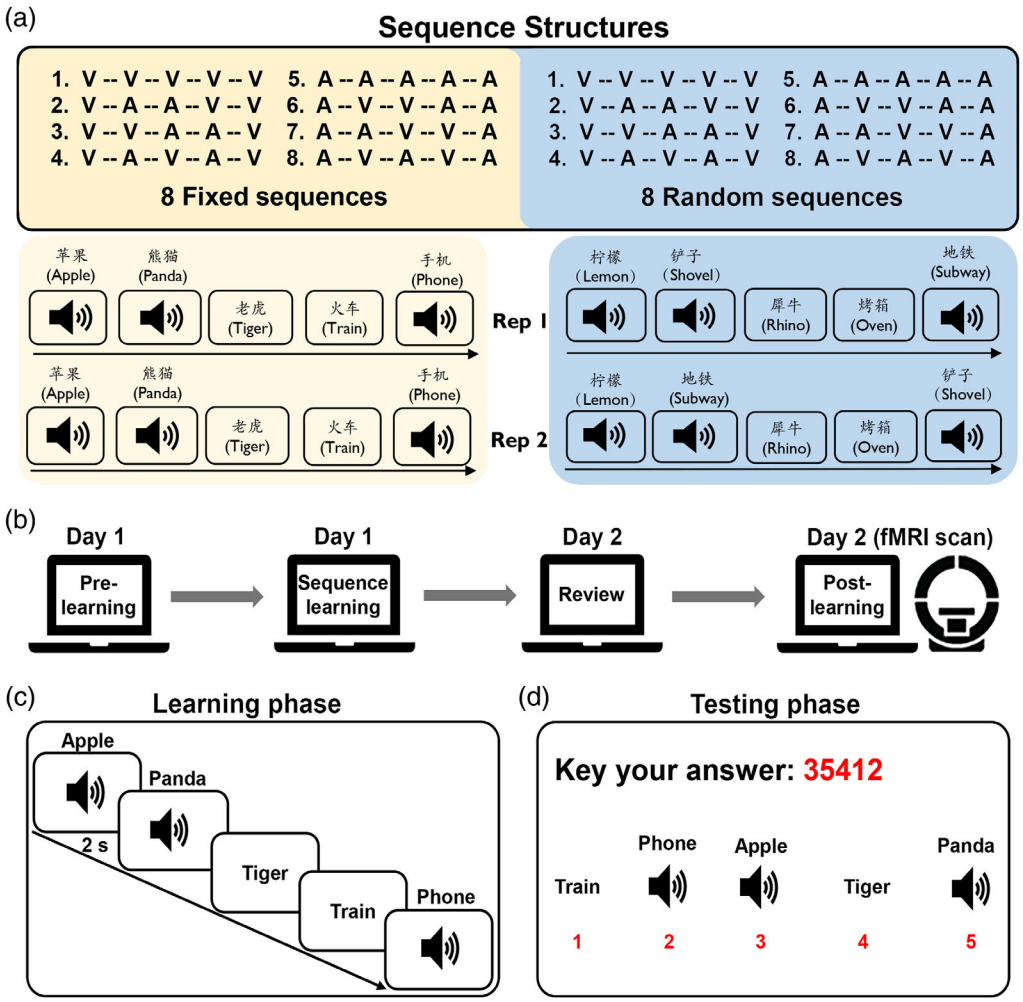

FIGURE 1 Experimental paradigm. (a) Sequence structures. Top: the same eight types of sequence structures were constructed for fixed and random sequences. Bottom: two exemplar sequences for fixed (left) and random (right) sequences. (b) Schematic depiction of the overall experimental procedure includes a pre-learning judgment task and a sequence learning task on day 1, a review task, and a post-learning semantic judgment task on day 2. (c) Task structure for the learning phase of the sequence learning task. Each word in a sequence was sequentially presented for 2 s, followed by a 0.5 s inter-stimulus interval (ISI). Participants were required to remember the sequence of their words. (d) Task structure for the test phase of the sequence learning task. Five words in a sequence were simultaneously presented on the screen in their learning modality but shuffled the order, and participants needed to reconstruct the sequence by typing the corresponding number into the text box in the correct order. Please note that for the auditory word, only a sound icon was presented on the screen during both the learning and testing phase, and the visual words were not presented to the participants. During the testing phase, participants could click the sound icon on the screen, and the sound of that word would be played.

predictable than the initial words, making such predictions based on sequence structure and previously presented items would require participants to actively track that information in working memory, which is very challenging. Our behavioral data suggest that participants indeed did not, and perhaps could not, actively track that information in random sequences to make such predictions about final items (Figure S1).

2.3 | Experimental design

The experiment was performed on two consecutive days (Figure 1b). On day 1, participants performed the pre-learning test and completed the sequence learning task. On day 2, participants reviewed the learned sequences and then performed the post-learning test. In the fMRI study (Exp 1), the post-learning test was completed in the fMRI scanner. In the behavioral study (Exp 2), the post-learning test was conducted in the same room as day 1.

2.3.1 | Pre-learning test (day 1)

To assess baseline performance, participants were asked to perform two pre-learning tests, namely the modality judgment task (auditory vs. visual) and the semantic judgment task (living vs. non-living). The order of the two tasks was counterbalanced between participants. In each run of a given task, the 16 sequences (8 fixed and 8 random) were pseudo-randomly concatenated using a Latin square design.

Each word was presented for 2 s, and participants were required to make a modality or semantic judgment by pressing 'f' or 'j' keys with their left or right index finger. A 500 ms fixation stimulus was presented before the next trial started. The correspondence between buttons and answers ("visual/auditory", "living/non-living") was counterbalanced between participants. Participants performed four runs of each task, with each run lasting approximately 3.5 min.

2.3.2 | Sequence learning task (day 1)

The sequence learning task started after the modality and semantic judgment tasks. During the sequence learning task, participants were required to learn all eight fixed sequences. In order to match the familiarity of the stimuli, the eight random sequences were also included in the sequence learning task. The only difference was that for the fixed sequence condition, the word orders were fixed across all repetitions, whereas in the random condition, they were randomized in each repetition while still following the abstract audiovisual structure. The learning task included a familiarization phase and then a study-test phase.

During the familiarization phase, the words in a sequence were presented one by one for 2 s with 0.5 s inter-stimulus-interval (ISI). Participants were asked to memorize the sequence. Each sequence was repeated three times before moving to the next sequence.

During the study-test phase (Figure 1c), participants first studied and were then immediately tested on each sequence, and subsequently, their memory for all sequences was tested serially. During the study phase, each word in a sequence was sequentially presented to the participants for 2 s, followed by a 0.5 s inter-stimulus fixation. Participants were instructed to remember the order of stimuli. After a 2 s interval, the five words were presented simultaneously on five different spatial locations of the screen in random order. The auditory stimulus was shown as a sound icon, and the corresponding word could be played out by clicking the icon. Participants were required to reproduce the word sequence by typing the number corresponding to each word into the text box in the correct order. Participants could click any sound icon as many times as necessary and in any order before they started to input the word order. Once participants confirmed their answers by pressing the "Enter" key, the sequence was presented again as feedback.

Once all 16 sequences were studied in this fashion, participants were tested on their memory. During the final testing phase (Figure 1d), the 16 sequences were presented to participants in a randomized order. All five stimuli of each sequence were presented on the screen simultaneously, and participants were required to reconstruct the correct sequence as described above. For the random sequences, participants needed to make the response according to the last presentation. No feedback was provided.

The study-test cycles continued until participants achieved 100% accuracy for the fixed sequences, whereas the accuracy in the random sequences was not considered. On average, participants required 3.57 ± 1.21 study-testing cycles to reach that criterion.

2.3.3 | Sequence review task (day 2)

On day 2, participants were first asked to review all 16 sequences learned on day 1. During the review task, each stimulus in a sequence was sequentially presented, and then all five stimuli were presented simultaneously on the screen. Participants were asked to reconstruct the sequence by typing the number corresponding to each word into the textbox in the correct order. No feedback was provided in the review task.

2.3.4 | Post-learning semantic judgment task in the scanner (day 2)

After reviewing the sequences, participants performed the post-learning semantic judgment task in the fMRI scanner (Exp 1). The procedures were similar to the pre-learning semantic judgment task, except that the ISI was set to 4 s to better characterize the single-trial BOLD response. During this interval, participants were asked to perform an orientation judgment where they needed to judge the orientation of an arrow by pressing the corresponding key as quickly as possible. Although this filler task might interfere and reduce with the prediction, it could offer better control of participants' behaviors and reduce the unexpected variance that might complicate the behavioral patterns and interpretations. Besides, since the orientation judgment task was used for all conditions (e.g., visual and auditory words in the fixed and random sequences) and the predictive effect was obtained by subtracting the response of the random sequence from that of the fixed sequence, any influence caused by this filler task would be canceled out. A self-paced procedure was used to make this task engaging. Participants finished 2.47 ± 0.67 orientation judgments per trial interval. Each word stimulus was tested four times in four separate runs of 8 min.

To replicate the results from Exp 1 and to further examine the effect of ISI duration, we did an additional behavioral experiment (Exp 2). In this experiment, there were two types of post-learning semantic judgment tasks: one task with 0.5 s ISI and one task with 4 s ISI. The procedure of the post-learning semantic task with 0.5 s ISI was the same as the pre-learning semantic judgment task, while the procedure of the post-learning semantic task with 4 s ISI was the same as the post-learning semantic task in the fMRI scanner. Each subject needed to complete these two types of semantic tasks, and the test order was counterbalanced across participants. Participants finished 2.45 ± 0.69 orientation judgments in the semantic task trials with 4 s ISI.

2.4 | MRI acquisition

MRI scanning was conducted on a 3 T Siemens Prisma scanner (Siemens, Erlangen, Germany) with a 20-channel head coil at the Center for MRI Research at Peking University. A high-resolution simultaneous multi-slice EPI sequence was used for functional imaging (FOV = 224 mm × 224 mm, TR/TE/θ = 2000 ms/30 ms/90°, slice

thickness = 2 mm, matrix = 112 × 112, slice acceleration factor = 2). A 3D, T1-weighted MPRAGE sequence (FOV = 256 mm × 256 mm, TR/TE/θ = 2530 ms/2.98 ms/7°, slice thickness = 1 mm, matrix = 256 × 256) were used to obtain high-resolution structural images. A high-resolution T2-weighed image using a T2-SPACE sequence was acquired for hippocampus segmentation. The image plane was perpendicular to the main hippocampal axis and covered the whole MTL region (FOV = 220 mm × 220 mm; matrix = 512 × 512; slice thickness = 1.5 mm; TR/TE/θ = 13,-150 ms/82 ms/150°, 60 slices). A field map was acquired for correction of magnetic field distortions using a Gradient Echo sequence (FOV = 224 mm × 224 mm; matrix = 112 × 112; slice thickness = 2 mm; TR/TE1/TE2/θ = 620 ms/4.92 ms/7.38 ms/60°, 62 slices).

2.5 | Image preprocessing

Neuroimaging data were first converted to Brain Imaging Data Structure (BIDS) format (Gorgolewski et al., 2016). Image preprocessing was conducted following the pipeline of FMRIPrep v1.4.0 (Esteban et al., 2019). Functional images were slice-timing corrected using AFNI v16.2.07 (Cox, 1996), motion-corrected using FSL's MCFLIRT (Jenkinson et al., 2002), and registered to the T1 image using boundary-based registration with nine degrees of freedom. Each T1 volume was corrected for intensity using N4BiasFieldCorrection (Tustison et al., 2010) and skull-stripped using antsBrainExtraction.sh (OASIS template), then normalized to the ICBM 152 Nonlinear Asymmetrical template (version 2009c) through nonlinear registration with the ANTs v2.1.0 (Avants et al., 2011). The images were temporally filtered using a nonlinear high-pass filter with a 100 s cutoff. For univariate analysis, images were spatially smoothed with a 6 mm full-width-at-half-maximum (FWHM) Gaussian kernel using FSL's SUSAN and normalized to MNI standard space. For representational similarity analysis (RSA), images were aligned to participants' T1 images and kept in their native space. Slight spatial smoothing was also applied to the data using a 2 mm FWMH Gaussian kernel to obtain a high signal-to-noise ratio and anatomical specificity.

2.6 | Definition of regions-of-interest

Following previous studies, we focused our main analysis on the visual cortex, auditory cortex, and higher-order association areas. Two sensory-specific regions-of-interest (ROIs) were defined based on the univariate activation analysis of the contrast between visual and auditory modality: left ventral visual cortex (VVC) for visual word processing (Xue et al., 2006; Xue & Poldrack, 2007; Zhao et al., 2017), and bilateral superior temporal gyrus (STG) for auditory word processing (Calvert et al., 2000; Van Atteveldt et al., 2004). Three higher-order ROIs were defined based on the Harvard-Oxford probabilistic atlas (threshold at 25% probability), including bilateral superior parietal lobe (SPL; Zhang et al., 2004, 2014), bilateral inferior parietal lobe (IPL, consisting of the supramarginal gyrus and angular gyrus; Tibon et al., 2019; Yazar et al., 2017), and bilateral middle frontal gyrus (MFG; Regev et al., 2013).

2.7 | Hippocampal subfields segmentation

The hippocampus and surrounding medial temporal lobe were segmented into CA1, CA2, CA3, DG, SUB, BA35/36, ERC, and PHC using the automatic segmentation of hippocampal subfields (ASHS) toolbox with the MTL-UPenn atlas (Yushkevich et al., 2015). The CA2, CA3, and DG were combined (i.e., DGCA23) because they could not be unambiguously distinguished. These masks were resampled and co-registered to each subject's native space. As a result, four ROIs, including left and right CA1, and DGCA23 (2 mm^3 resolution, numbers of voxels: left CA1 = 152.46 ± 20.41; left DGCA23 = 96.29 ± 13.86; right CA1 = 153.68 ± 20.74; right DGCA23 = 105.93 ± 15.54), were included in the further analysis.

2.8 | Univariate activation analysis

The general linear model (GLM) was constructed using the FILM module of FSL (version 6.00). According to the sequence structure, each word within its sequence structure was assigned to one type of eight events according to three factors: predicting modality (visual vs. auditory) × sequence type (fixed vs. random) × predicted modality (visual vs. auditory). These eight types of events were modeled as regressors of interest, and the orientation trials were modeled as regressors of no interest. Both the eight task regressors and the regressors of no interest were convolved with a double gamma hemodynamic response function. The six motion parameters and the frame-wise displacement (FD) were also included as confounding regressors. In addition, each volume with an FD greater than 0.3 mm was separately modeled as a censor regressor. Each run was modeled separately in the first-level analysis. Cross-run averages for each contrast image were created using a fixed-effects model for each subject. These contrast images were then used for group-level analysis with a random-effects model. Group-level statistical results were reported using cluster detection methods, with a height threshold of $z > 2.3$ and a cluster probability of $p < .05$, corrected for whole-brain multiple comparisons using Gaussian Random Field Theory.

2.9 | Single-trial response estimation

The least-square separate (LSS) method was used to estimate single-trial response for each functional run (Mumford et al., 2012). Each trial was estimated in a separate GLM, in which the trial was modeled as a separate regressor, whereas all the other trials were modeled as another regressor. We also included six movement parameters and FD as confound regressors. Additional censor regressors were included for each volume with an FD greater than 0.3 mm. This

resulted in one t map for each trial, which was used for the representational similarity analysis.

Several methods have been proposed to correct for multiple noise and biases in RSA (Cai et al., 2019; Diedrichsen et al., 2011; Walther et al., 2016). In particular, multivariate noise normalization (i.e., prewhitening) has been shown to be effective in correcting the spatially correlated noise between voxels, although its effect has been shown to be varied in different brain regions and designs (Ritchie et al., 2021). To evaluate the effect of multivariate noise normalization, we respectively used the least-square all (LSA) method that all trials were estimated simultaneously in a GLM, the LSA plus prewhitening (LSA_Prew) method, which uses multivariate noise normalization with a regularized estimate of the spatial noise-covariance matrix (Diedrichsen et al., 2016; Walther et al., 2016), and the LSS method described above. To examine the reliability of each method, we split the four runs into two run pairs and constructed an 80*80 (each runs had all 80 unique items) cross-run representational dissimilarity matrix (RDM) for each run pair. The reliability was measured as the correlation between the two RDMs (i.e., second order correlation). Since there were three ways to split the four runs, the three Fisher *z*-transformed correlations were averaged to represent the overall reliability. We found that the LSS method outperformed than the other two methods in all ROIs (Figure S2). Therefore, the t image for each trial derived from the LSS method was used for subsequent RSA.

2.10 | Representational similarity analysis

RSA was used to estimate the neural pattern similarity across trials (Kriegeskorte et al., 2008). In our study, RSA was conducted within the five pre-defined ROIs. All representational similarity analyses were conducted between trials from different runs, which have been shown to be less affected by the temporal autocorrelation of BOLD signals (Alink et al., 2015; Henriksson et al., 2015). The *T* value from the single-trial response estimate was extracted for each trial and each voxel within the ROI, and Pearson correlation was used to quantify the pattern similarity. Only correct trials were included in this analysis. These Pearson correlations were then Fisher *Z* transformed for further analysis.

To examine the sequence prediction representation and how they were modulated by modality, we conducted RSA between two adjacent trials (that were 1 position apart, i.e., lag 1) for each type of transition, i.e., visual–Visual (vV), auditory–Auditory (aA), visual–Auditory (vA), and auditory–Visual (aV) pairs, for the fixed and random sequences, respectively. Again, it should be emphasized that the two words in any combination were from different runs. This was enabled by the design that each sequence was presented once in a run. We first examined the sequence prediction effect in each condition separately by comparing the pattern similarity of lag 1 pairs between fixed and random sequences. Then, a predicting modality (visual vs. auditory) by predicted modality (visual vs. auditory) ANOVA was performed on the predictive effect to further examine the cross-modal predictive effect.

2.11 | Representational connectivity analysis

Representational connectivity analysis was conducted to examine the within- and between-regions informative connectivity across adjacent trials (Anzellotti & Coutanche, 2018). As above, this analysis was also conducted across runs so that the predicting words and the predicted words were from different runs. We first calculated a representational similarity matrix (RSM, 8×8) for the eight predicting words and eight predicted words in each condition (i.e., vV, aA, vA, and aV), separately for each fixed and random sequence and in each ROI (Figure 5a). Then, second-order correlation analyses were performed to calculate the informative connectivity between the two adjacent words, either within or between regions. Specifically, we correlated the RSMs of the predicting words in one region with the RSMs of the predicted words in the same or different regions (Xiao et al., 2017). Nonparametric permutation tests were performed to examine the group-level significance of these representational connectivities by shuffling one of the RSMs (8×8). Permutations were conducted 5000 times and then averaged to obtain the baseline correlation coefficient under the null hypothesis for each ROI and each subject. We also compare the connectivity value between fixed and random sequences to determine whether there was significant informative connectivity across adjacent words. Furthermore, we further conducted a sequence type (fixed vs. random) by condition (vV vs. aA vs. vA vs. aV) ANOVA on the empirical connectivity value to examine the condition-specific property of informative connectivity.

2.12 | Statistical analysis

All *t*-tests and repeated measures ANOVAs in our analyses were conducted by the afex package in R 4.1.2. We used type III sums of squares and the Greenhouse–Geisser method to correct the degrees of freedom. Error bars in the bar plot figures denote within-subject errors that account for the heterogeneity of variance. FDR correction was used for multiple comparisons between ROIs.

2.13 | Data and code availability

The code can be accessed via GitHub (https://github.com/leonepsy/RSA_sequence_prediction). Both fMRI and behavior data are available at the Open Science Framework: https://osf.io/dn6ty/.

3 | RESULTS

3.1 | Behavioral performance

Participants were required to make semantic judgments on sequentially presented words before learning on day 1 and after learning on day 2. Once participants learned the word sequences, they could predict the next word in the fixed sequence because the words were

always presented in the same order. In contrast, they could not make an accurate prediction for the random sequence because the words in the random sequence were presented in random orders. Thus, we predicted after learning, they would respond faster to words from the fixed sequences than those from the random sequences (Hsieh et al., 2014). We further predicted that this predictive effect would be stronger when the preceding word was from the same modality (i.e., within-modal prediction) than from a different modality (i.e., cross-modal prediction). We only included the last four words in each sequence in our behavioral analysis since the first word in a sequence could not be predicted from a prior sequence.

The accuracy of the semantic judgment task in the fMRI experiment was very high (88%–93%) and did not show a significant main effect of task phase (pre- vs. post-learning), sequence type (fixed vs. random), or any interaction (all *p*s > .2, uncorrected). Consequently, the following analysis focused on the reaction time of correct trials. As expected, this analysis on RT revealed a significant main effect of the task phase ($F_{1,20} = 8.475$, $p = .009$) and a task phase by sequence type interaction effect ($F_{1,20} = 7.307$, $p = .014$). Follow-up analyses revealed that although participants responded equally fast to words from the random and fixed sequences before learning ($t_{20} = -1.358$, $p = .190$), the RTs for words from the fixed sequences were significantly shorter than that from the random sequences in the post-learning phase (Figure 2a, $t_{20} = 2.264$, $p = .035$).

To test our second hypothesis regarding the modality effect, we grouped these words into four conditions according to the modality of the current word and the proceeding word (i.e., 1 position apart): vV, vA, aA, aV, where the first lowercase letter indicates the modality of the proceeding word (predicting modality) and the second uppercase letter indicates the modality of the current word (predicted modality) whose RT were analyzed. For example, in the V-V-A-A-V sequence, the word on position 2 was assigned to vV condition, and the words on positions 3, 4, and 5 were assigned to vA, aA, and aV condition, respectively. We then calculated the RT differences between the target words in the fixed sequences and that in the random sequences, before and after learning: RT difference = $(RT_{post\_random} - RT_{post\_fixed}) - (RT_{pre\_random} - RT_{pre\_fixed})$, with greater differences indicating a greater predictive effect for fixed sequences than for random sequences as a result of learning. Using the RT differences as dependent variable, a predicting modality (visual vs. auditory) by modality shift (within- vs. cross-modal) ANOVA revealed a significant main effect of predicting modality ($F_{1,20} = 20.175$, $p < .001$), and a significant interaction between predicting modality and modality shift ($F_{1,20} = 4.69$, $p = .043$), but no significant main effect of modality shift ($F_{1,20} = 0.01$, $p = .921$). Follow-up analysis found an asymmetric cross-modal predictive effect, as indicated by greater predictive effect for the vA condition than aV condition ($t_{20} = 4.262$, $p < .001$) (Figure 2b). No significant

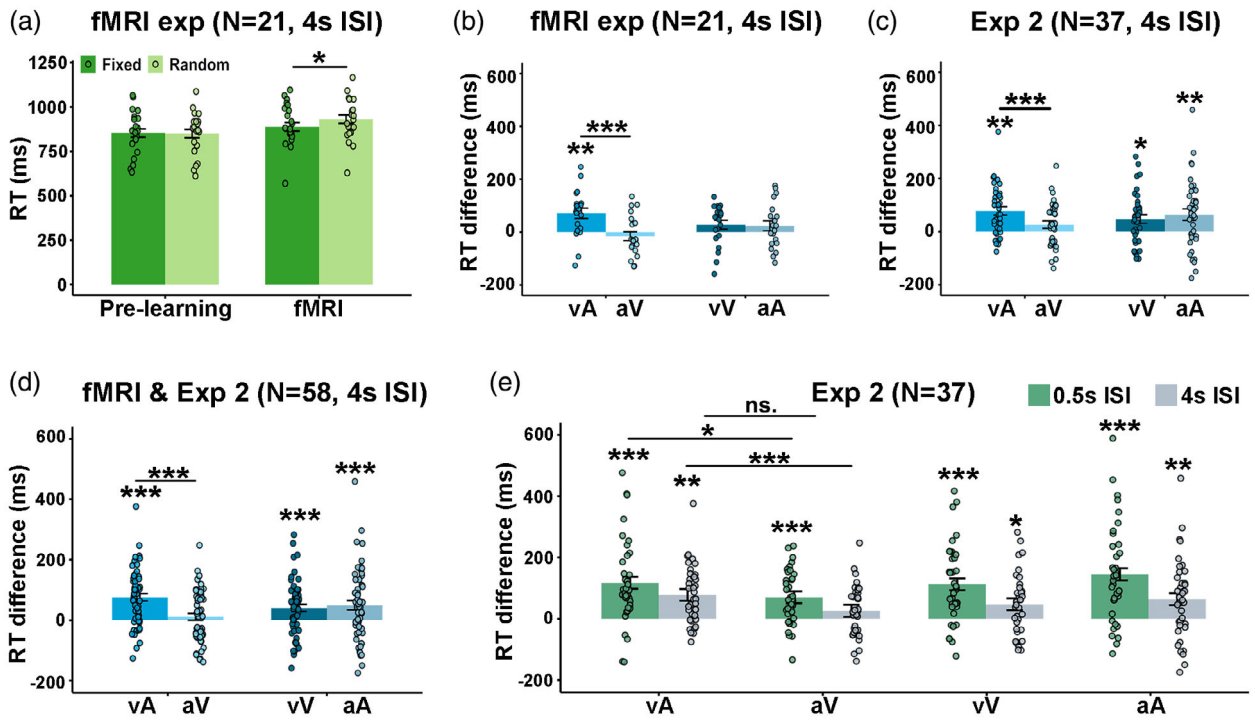

FIGURE 2 Behavioral performance. (a) The mean RT for the semantic judgment task is a function of task phase and sequence type. (b) The RT differences as a function of predicting modality (modality of the preceding item, indicated by the first lowercase letter) and predicted modality (modality of the current item, indicated by the second uppercase letter). A(a) and V(v) indicated auditory and visual modality, respectively. The RT differences were calculated using the following formula: RT difference = $(RT_{post\_random} - RT_{post\_fixed}) - (RT_{pre\_random} - RT_{pre\_fixed})$, with greater RT differences indicating a higher predicting effect as a result of sequence learning. (c) The RT differences in the additional behavioral experiment (Exp 2, 4 s ISI condition). (d) The RT differences after pooling the data from the fMRI experiment and the Exp 2 (4 s ISI condition). (e) The RT differences between 0.5 s and 4 s ISI conditions in Exp 2. Each dot represents one subject, and the bars represent the group means. Error bars indicate within-subject errors. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

difference was observed between the two within-modal conditions (vV vs. aA) ($t_{20} = 0.179$, $p = .860$), or between aV and aA conditions ($t_{20} = -1.550$, $p = .137$), or between vA and vV conditions ($t_{20} = 1.589$, $p = .128$).

To examine whether this asymmetric cross-modal predictive effect was caused by the long inter-stimulus-interval (ISI, 4 s) employed by the fMRI design, we did an additional behavioral experiment (Exp 2) on 37 participants. The whole procedure was identical to the fMRI experiment except a short ISI condition (0.5 s with no orientation task) was also included for the semantic judgment task. The order of the two ISI conditions (4 s vs. 0.5 s) was counterbalanced across participants. An ISI conditions (4 s vs. 0.5 s) by predicting modality (visual vs. auditory) by modality shift (within vs. cross-modal) three-way ANOVA was conducted on the RT difference to examine the effect of ISI on the predictive effect. Results showed a significant main effect of ISI conditions ($F_{1,36} = 8.83$, $p = .005$) and significant predicting modality by modality shift interaction ($F_{1,36} = 16.31$, $p < .001$), but no other significant main or two-way or three-way interaction effects (all ps > .128). These results replicated the asymmetric cross-modal predictive effect in both the 4 s (interaction effect: $F_{1,36} = 11.426$, $p = .002$; vA > aV: $t_{36} = 4.038$, $p < .001$) and 0.5 s conditions (interaction effect: $F_{1,36} = 6.574$, $p = .015$; vA > aV: $t_{36} = 2.244$, $p = .031$) (Figure 2c, Table S2). Since we did not find significant interaction between the fMRI study and behavioral study (4 s condition) ($F_{1,56} = 0.14$, $p = .712$), we pooled together the data from the both experiments. We found strong asymmetric cross-modal prediction effect (interaction effect: $F_{1,57} = 15.409$, $p < .001$; vA > aV: $t_{57} = 5.790$, $p < .001$) (Figure 2d). We also found that longer ISI reduced the overall predictive effect ($F_{1,36} = 6.506$, $p = .015$), but there was no interaction effect between ISI and the asymmetric cross-modal predictive effect ($F_{1,36} = 0.10$, $p = .755$) (Figure 2e).

One possible reason for the asymmetric cross-modal prediction is that the response to visual words was faster than auditory words (763 ms vs. 1001 ms, $t_{20} = 22.266$, $p < .001$), so that there was less room for improvement for the visual words. To account for this effect, we calculated the z score of RT difference using the following formula: $zRT\ difference = ([RT_{pre\_fixed} - RT_{pre\_random\_mean}]/RT_{pre\_random\_std}) - ([RT_{post\_fixed} - RT_{post\_random\_mean}]/RT_{post\_random\_std})$. Our results still revealed a significant asymmetric predictive effect for the fMRI study (interaction effect: $F_{1,20} = 4.36$, $p = .05$; vA > aV: $t_{20} = 2.658$, $p = .015$) and the behavioral study (4 s condition: interaction effect: $F_{1,36} = 4.46$, $p = .042$; vA > aV: $t_{36} = 3.200$, $p = .003$; 0.5 s condition: interaction effect: $F_{1,36} = 4.84$, $p = .034$; vA > aV: $t_{36} = 2.495$, $p = .017$). This effect was still significant when data from both studies were pooled together (interaction effect: $F_{1,57} = 8.87$, $p = .004$; vA > aV: $t_{57} = 4.183$, $p < .001$). Overall, the effect size (Cohen's d) was decreased after normalizing the RT difference (raw RT difference vs zRT difference: 0.93 vs. 0.58 in the fMRI study; 0.66 vs. 0.53 [4 s condition] and 0.37 vs. 0.41 [0.5 s condition] in the behavioral study; 0.75 vs. 0.55 when the results from both the fMRI and behavioral study [4 s condition] were pooled).

3.2 | Multiple frontoparietal regions were involved in sequence prediction

After demonstrating the overall predictive effect as a result of sequence learning, we then turned to examine its underlying neural basis. To reveal the brain regions involved in the predictive effect and the modality effect, we did a three-way ANOVA on the words from positions 2–5 in a sequence, including the predicted modality (visual vs. auditory), modality shift (within- vs. cross-modal) and sequence type (fixed vs. random). This analysis revealed no significant two-way or three-way interaction effects, but significant main effects of sequence type and of predicted modality. In particular, we found greater activation for visual words than auditory words in the left ventral visual cortex (VVC, MNI coordinates: $x = -46$, $y = -72$, $z = -16$, $Z = 5.95$) and right VVC (MNI coordinates: $x = 46$, $y = -70$, $z = -18$, $Z = 5.65$) (Figure 3a). In contrast, there was greater activation for auditory words than visual words in the bilateral superior temporal gyrus (STG, left: MNI coordinates: $x = -52$, $y = -16$, $z = 5$, $Z = 7.02$; right: MNI coordinates: $x = 58$, $y = -14$, $z = 0$, $Z = 6.69$), left supplementary motor area (MNI coordinates: x = −4, y = 8, z = 55, Z = 4.71), and left precuneus (MNI coordinates: $x = -8$, $y = -62$, $z = 58$, $Z = 3.82$) (Figure 3b). Similar activation patterns were observed when we conducted a predicting modality, modality shift and sequence type three-way ANOVA, focusing on words from sequence positions 1–4 (Figure S3).

More importantly, frontoparietal regions were involved in cross-modal prediction, showing greater activation for words from the fixed sequences than for those from the random sequences. These regions were superior parietal lobe (SPL, MNI coordinates: $x = -32$, $y = -56$, $z = 51$, $Z = 4.93$), left middle frontal gyrus (MFG, MNI coordinates: $x = -34$, $y = 60$, $z = 12$, $Z = 4.86$), and right precentral gyrus (MNI coordinates: $x = 50$, $y = 6$, $z = 35$, $Z = 4.77$) (Figure 3c). Only medial superior frontal gyrus (MNI coordinates: $x = -10$, $y = 58$, $z = -4$, $Z = 3.25$) showed a reversed pattern (Figure 3d).

Given the significant asymmetric cross-modal predicting effect, we directly compared the vA vs. aV conditions. Sequence type (fixed vs. random) by condition (vA vs. aV) ANOVA revealed no significant main effect or interaction.

In addition, given the differences in RT between trials, we did a further univariate analysis which regressed out the trial-wise RTs in the first-level models. This analysis revealed similar results (Figure S4, Table S3), confirming the robustness of our findings.

3.3 | Predictive coding of item representations in the fixed sequences

Having identified the involvement of unimodal and higher-order areas in the sequence prediction, we further examined the predictive representation in these regions. One possible mechanism underlying the predictive effect (i.e., faster RT for items in the fixed sequences than in random sequences) is the predictive activation of the next item in a

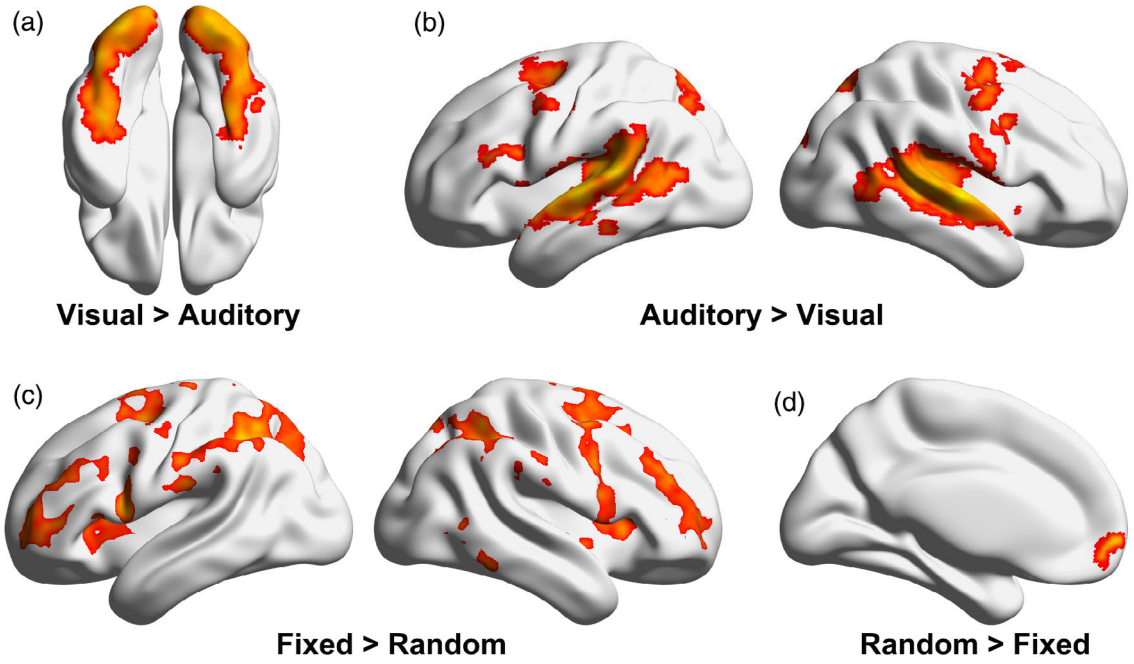

FIGURE 3 Whole-brain univariate activation analysis on predicted words (words on sequence positions 2–5). (a) Univariate results for visual > auditory contrast. (b) Univariate results for auditory > visual contrast. (c) Univariate results for fixed > random contrast. (d) Univariate results for random > fixed contrast. Images were thresholded using cluster detection statistics, with a height threshold of $z > 2.3$ and a cluster probability of $p < .05$, corrected for multiple comparisons across the whole brain using Gaussian Random Field Theory.

sequence. If this were the case, we would predict greater pattern similarity for adjacent items (i.e., lag 1 pairs) in fixed sequences than in random sequences (Figure 4a). Furthermore, the effect would be modulated by the modality of predicting items due to the asymmetric cross-modal predictive effect.

To test this hypothesis, RSA was used to calculate the pattern similarity between adjacent items within a sequence but from different runs. We grouped all lag 1 pairs into 4 conditions, that is, vV, aA, vA, and aV, and then examined the predictive effect for each condition separately. Following previous studies and univariate results, we focused our main analysis on the unimodal regions (i.e., VVC and STG) and the higher-order association areas (i.e., MFG, IPL, and SPL).

We found that the higher-order areas showed a significant predictive effect for one or both cross-modal conditions, but not for the within-modal conditions. In particular, MFG showed a significant predictive effect for the aV ($t_{20} = 3.351$, corrected $p = .005$) and vA conditions ($t_{20} = 2.323$, corrected $p = .046$) (Figure 4b), and IPL showed a significant predictive effect for the aV ($t_{20} = 4.163$, corrected $p = .001$) condition (Figure 4c), whereas SPL showed significant predictive effect for the vA condition ($t_{20} = 4.561$, corrected $p = .001$) (Figure 4d). No other predictive effect was significant in these regions (all $ps > .078$, uncorrected). In contrast, the unimodal regions showed significant within-modal predictions. In particular, the VVC showed a significant predictive effect for the vV condition ($t_{20} = 2.579$, corrected $p = .036$) (Figure 4e), whereas the STG showed a significant predictive effect for aA condition ($t_{20} = 2.546$, corrected $p = .038$) (Figure 4f). No other predictive effect was significant (all $ps > .247$, uncorrected).

To directly compare the predictive effect between different modalities, we conducted a predicting modality (visual vs. auditory) by predicted modality (visual vs. auditory) ANOVA. This analysis revealed a significant interaction between predicting modality and predicted modality ($F_{1,20} = 20.36$, corrected $p = .001$) in SPL. No significant main effect or interaction effect was found in the other regions (all $ps > .129$, corrected). Follow-up analyses revealed that mirroring the asymmetric cross-modal predictive effect in the behavioral data, SPL exhibited a greater predicting effect for vA than aV pairs ($t_{20} = 4.557$, $p < .001$) (Figure 4d). The correlational analysis further showed that a greater neural predicting effect for the vA pairs was significantly correlated with the behavioral predicting effect in the vA condition ($r = 0.525$, $p = .015$) (Figure 4g), suggesting that SPL plays an important role in the visual-to-auditory prediction.

3.4 | Direct neural pathways supported within-modal prediction

The above analyses revealed that the unimodal and higher-order areas were respectively involved in the within-modal and cross-modal sequence prediction. We then further examined the neural pathways underlying successful within-modal and cross-modal sequence predictions. Two possible neural pathways were examined: the direct pathway, which involved the direct connectivity within and between the unimodal regions, and the indirect pathway mediated by the higher-order areas. We predicted that the within-modal predictions would be supported by a direct pathway within modality-specific regions,

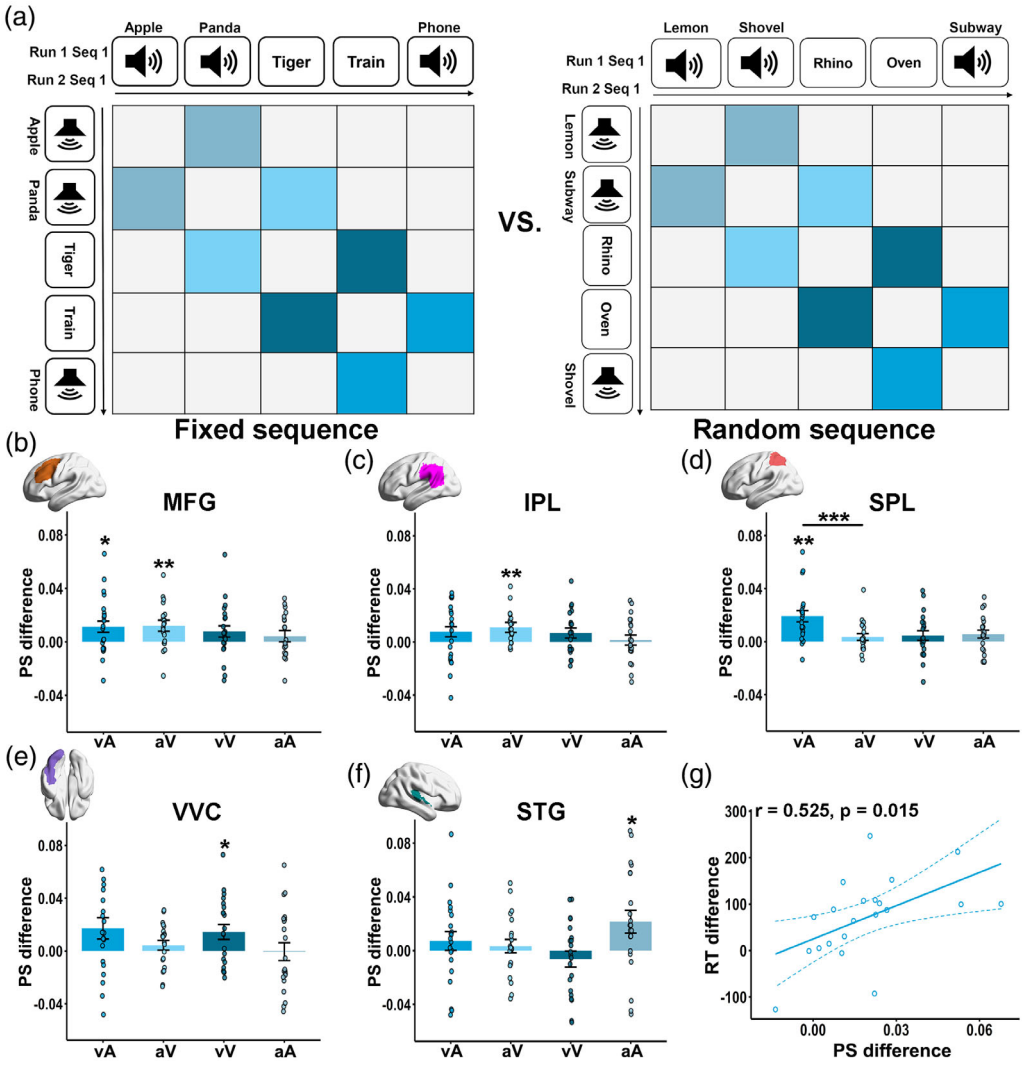

FIGURE 4 The neural predicting effect is a result of cross-modal sequence learning. (a) Schematic depiction of the pattern similarity (PS) analysis, which was done between adjacent words (i.e., lag 1), separately for each of the four types of word pair, i.e., visual–Visual (vV), auditory–Auditory (aA), visual–Auditory (vA), and auditory–Visual (aV) pairs, and for the fixed (left panel) and random sequence (right panel). Please note that the two adjacent words were from different runs. (b–f) The neural predicting effect in the five pre-defined ROIs, separately for each pair type. The PS difference was obtained by substrating the pattern similarity between pairs of adjacent words in the random sequences from that in the fixed sequences, which was used to index the neural predicting effect. The SPL showed an asymmetric cross-modal predictive effect, as indicated by a greater PS difference for the vA condition than the aV condition. Each dot represents one subject, and the bars represent group means. Error bars indicate within-subject errors. G. Correlation between the neural predicting effect (PS difference) in the SPL and behavioral predicting effect (RT difference) in the vA condition. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

whereas the cross-modal predictions would be supported by an indirect pathway mediated by higher-order areas (i.e., SPL). To test these hypotheses, representational connectivity analysis was used to examine the neural pathways (Figure 5a, see Methods) separately for vV, Aa, vA, and aV conditions.

To test the direct pathway, we examined the informative connectivity of two consecutive words within the unimodal regions (e.g., VVC-VVC for vV condition and VVC-STG for vA condition). Supporting our hypothesis, we found significant direct connectivity within the VVC for the vV condition (fixed vs. baseline: $t_{20} = 2.570$, $p = .018$; fixed vs random: $t_{20} = 2.197$, $p = .040$) (Figure 5b), and within the STG for the aA condition (fixed vs. baseline: $t_{20} = 2.686$, $p = .014$; fixed vs random: $t_{20} = 2.225$, $p = .038$) (Figure 5c). However, no significant connectivity was found for aV or vA conditions (all $ps > .911$, uncorrected) (Figure 5d,e). Confirming these dissociative effects, the condition (vV vs. aA vs. vA vs. aV) by sequence type (fixed vs. random) ANOVAs revealed a significant interaction effect for the VVC–VVC connectivity ($F_{3,60} = 3.271$, $p = .032$), and the STG–STG connectivity ($F_{3,60} = 3.979$, $p = .012$). Follow-up analyses found stronger VVC–VVC connectivity for fixed than random sequence only in the vV condition but not for the other three conditions (all $ps > .095$, uncorrected). Similarly, stronger STG–STG connectivity was observed for fixed than for random sequence in the aA condition but not for the other three conditions (all $ps > .101$, uncorrected).

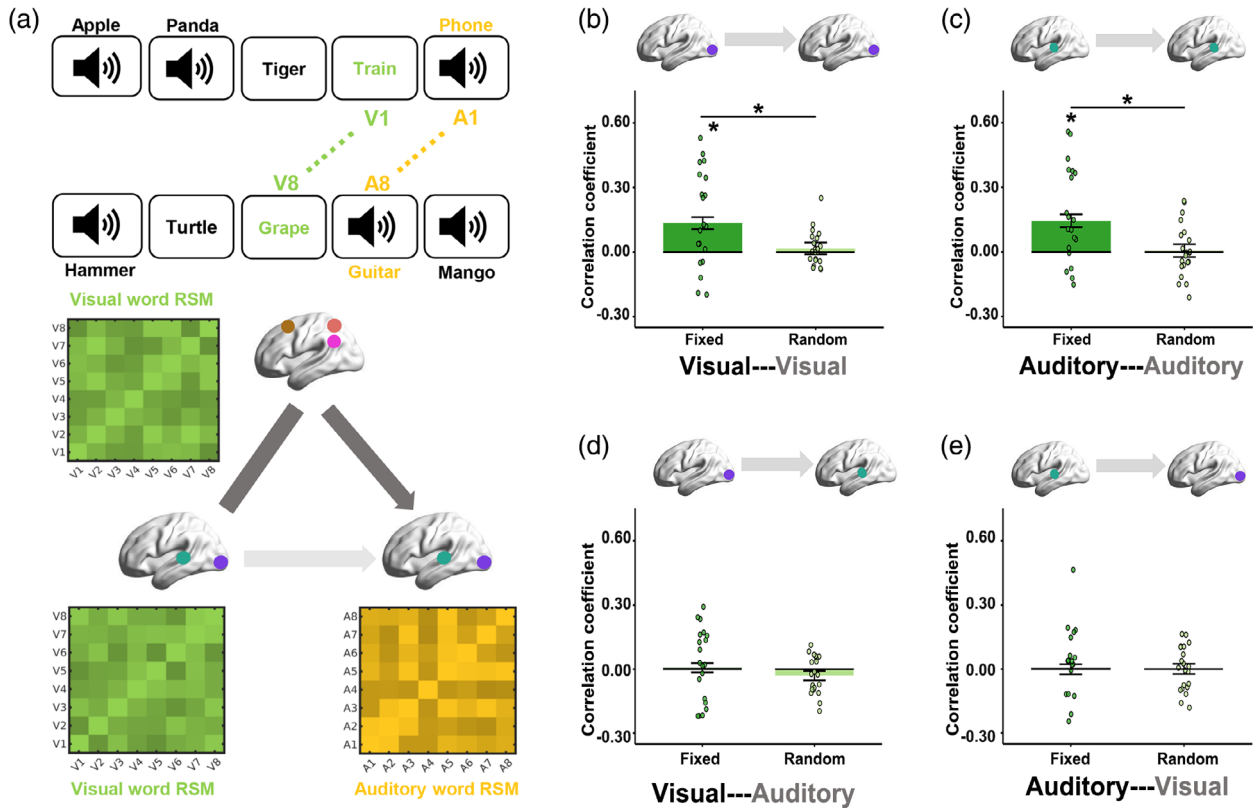

FIGURE 5 Representational connectivity analysis and direct neural pathways. (a) Schematic depiction of the representational connectivity analysis. Take the vA condition as an example. We constructed an 8 × 8 representational similarity matrix (RSM) for the predicting words (V1 to V8) and an 8 × 8 RSM for the predicted words (A1 to A8) in both the unimodal and higher-order areas. The information connectivity (Pearson correlation) could be calculated within- or cross-regions and within- and cross-stages. Please note that the predicting words and the predicted words were from different runs. (b) Direct connectivity within VVC underlies vV prediction. (c) Direct connectivity within STG underlies aA prediction. (d) Direct connectivity between VVC and STG underlies vA prediction. (e) Direct connectivity between STG and VVC underlies aV prediction. The *y* axis represents the Pearson correlation coefficients between the RSMs. The black horizontal line indicates the baseline (mean shuffled similarity value). Each dot represents one subject, and the bars represent group means. Error bars indicate within-subject errors. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

3.5 | Indirect neural pathway supported cross-modal prediction

For the indirect pathway, we examined whether there was significant informative connectivity between the predicting words (e.g., the visual word in vA condition) in the SPL and that of the predicted words (e.g., the auditory word in vA condition) in the STG for the vA condition. Again, informative connectivity was significant if the correlation for the fixed sequence was greater than the baseline and also greater than that for the random sequence. In addition, there should also be significant informative connectivity between the VVC and the SPL for the predicting words, and this connectivity should not differ between fixed and random sequences.

Again supporting our hypothesis, we found a significant indirect pathway from VVC to STG via SPL for the vA condition (Figure 6a). In particular, there was significant informative connectivity between VVC and SPL for the predicting words in both the fixed ($t_{20} = 4.056$, corrected $p = .002$) and random sequence ($t_{20} = 4.020$, corrected $p = .001$), and no significant difference between them was found ($t_{20} = 1.284$, $p = .428$). Importantly, there was significant informative connectivity between the predicting words in the SPL and the predicted words in the STG for the vA condition (fixed vs. baseline: $t_{20} = 2.910$, corrected $p = .027$; fixed vs. random: $t_{20} = 3.607$, corrected $p = .006$). A condition (vV vs. aA vs. vA vs. aV) by sequence type (fixed vs. random) ANOVA on the SPL–STG connectivity revealed a significant interaction effect ($F_{2.42,48.34} = 5.10$, $p = .007$), and follow-up analysis found that this SPL–STG connectivity was specific to the vA condition, but not found in the other three conditions (all *p*s > .098, uncorrected),

Consistent with the weak prediction from auditory to visual stimuli, we found the STG-SPL-VVC indirect pathway was not significant (all *p*s > .05, uncorrected) (Figure 6b). Additionally, there was no significant indirect pathway for the unimodal conditions (i.e., VVC–SPL–VVC and STG–SPL–STG for vV and aA, respectively) (Figure S5). Furthermore, supporting the unique role of the SPL in the indirect pathway, we found no significant indirect pathway through the MFG or IPL for the cross-modal conditions (all *p*s > .108, uncorrected) (Figure S6) or the within-modal conditions (all *p*s > .069, uncorrected) (Figure S7).

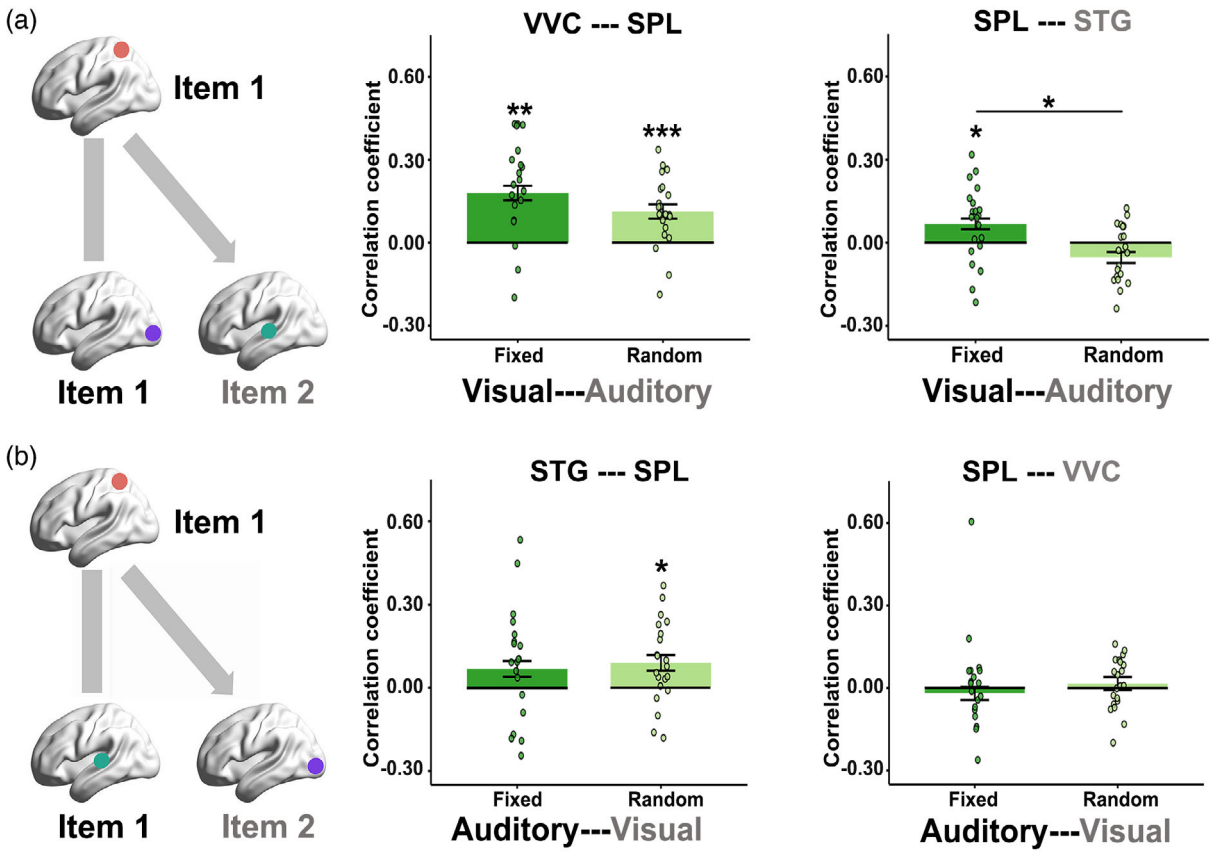

FIGURE 6 Indirect neural pathways underlying cross-modal predictions. (a) (Left): Schematic depiction of an indirect neural pathway via SPL (red circle) connecting the unimodal regions under vA condition. (Middle): VVC–SPL connection for the predicting words in the vA condition. (Right): The informative connection between SPL for the predicting words and STG for the predicted words in the vA condition. (b) (Left): Schematic depiction of an indirect neural pathway via SPL (red circle) connecting the unimodal regions under aV condition. (Middle): STG–SPL connection for the predicting words in the aV condition. (Right): The informative connection between SPL for the predicting words and VVC for the predicted words in the aV condition. The *y* axis represents the Pearson correlation coefficients between the RSMs. The black horizontal line indicates the baseline (mean shuffled similarity value). Each dot represents one subject, and the bars represent group means. Error bars indicate within-subject errors. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Together, the connectivity results revealed dissociable neural pathways for within-modal and cross-modal predictions and highlighted the role of SPL in the indirect pathway supporting visual-to-auditory prediction.

3.6 | The involvement of hippocampus in cross-modal sequence prediction

To examine the contribution of the hippocampal subfields to the cross-modal sequence prediction, we conducted the above analyses in the CA1 and DGCA23 subregions (Figure 7). Only 20 participants were included in those analyses as the T2-weighted high-resolution image was missing for one subject. Univariate analysis revealed no significant differences between the fixed and random sequences in any hippocampal subregions. RSA of adjacent stimuli showed that left CA1 exhibited a marginally significant predictive effect for the vV condition ($t_{19} = 2.238$, uncorrected $p = .0374$) but not for the other three conditions (all *p*s > .078, uncorrected) (Figure 7). No significant predictive effect was found in the other three regions (all *p*s > .109, uncorrected).

4 | DISCUSSION

The present study aimed to advance our understanding of cross-modal sequence learning and prediction. Extant studies on within-modal sequence prediction have shown that once participants have learned an event sequence, they exhibit a faster response to items that could be predicted by the preceding item than to those that could not (Hsieh et al., 2014). Our results extended this finding to cross-modal sequence prediction. Interestingly, the behavioral predicting effect is comparable between within-modal and visual-to-auditory cross-modal conditions.

In line with previous studies which have implicated higher-order areas in cross-modal associations (Tanabe et al., 2005; Tibon et al., 2019; Zhang et al., 2004), we found that multiple frontoparietal regions, including SPL, SFG, and IFG, showed greater activation for

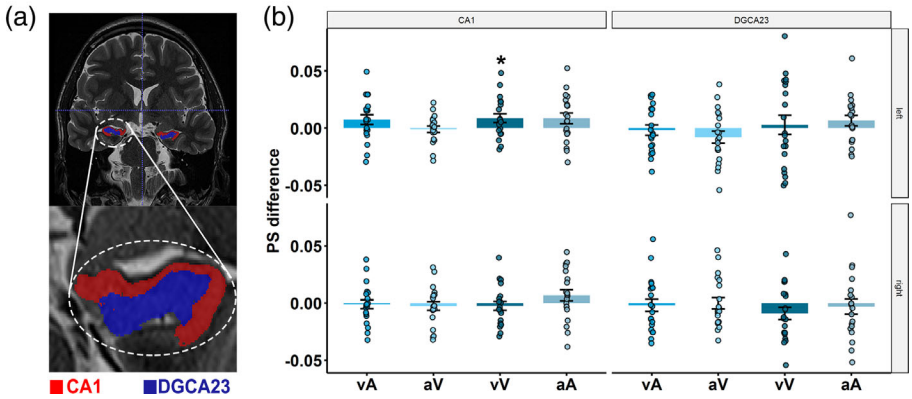

FIGURE 7 The role of hippocampal subfields in cross-modal sequence prediction. (a) A representative image of hippocampal subfields from one subject, including CA1 (red) and DGCA23 (purple), which are overlaid onto the subject's T2 image. (b) The predictive representation effect was examined in the subregions of the Hippocampus, separately for each pair type (i.e., vV, aA, vA, aV). The result showed there was a neural predicting effect on vV condition in the left CA1. Each dot represents one subject, and the bars represent group means. Error bars indicate within-subject errors. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

items in the fixed sequences than that in the random sequences. This could be explained by the anticipatory role of higher-order areas in event prediction (Alexander & Brown, 2014; Brunec & Momennejad, 2022; Lee et al., 2021). However, we failed to find involvement of any hippocampus subfields in the cross-modal sequence prediction, which is somewhat inconsistent with previous studies which reported significant hippocampal involvement in unimodal sequence prediction (Paz et al., 2010; Schapiro et al., 2012). One factor that may contribute to this discrepancy is the experimental materials. While images of objects (Reddy et al., 2015; Senoussi et al., 2020) or movies (Lee et al., 2021; Paz et al., 2010) were used in the previous studies, familiar words were used in the current study. Existing studies have shown that the neural representation of words is more likely to involve higher-order areas than pictures (Devereux et al., 2013; Liuzzi et al., 2021; Vandenberghe et al., 1996).

Our results further suggest within- and cross-modal sequence predictions may involve distinct cognitive and neural mechanisms. First, we found that within-modal sequence prediction elicited preactivation of the next item in the modality-specific regions, in line with the previous study, which found predictive representations in the visual (Ekman et al., 2017; Kok et al., 2013, 2017) and auditory cortex (Demarchi et al., 2019) during unimodal association. In contrast, cross-modal sequence prediction exhibited a similar predictive activation pattern in the higher-order areas. This may be related to the information coding nature of higher-order areas, which support modality-independent representations of cross-modal stimuli (Handjaras et al., 2017; Jung et al., 2018). Thus, our findings suggest the predictive mechanism of cross-modal prediction activates the abstract-level information representation in the higher-order areas.

Moreover, connectivity analysis revealed dissociable neural pathways for within-modal and cross-modal predictions. The within-modal prediction was supported by direct pathways within the unimodal regions. This suggests that one possible mechanism for within-modal prediction is that the sensory cortices could develop a direct association at the sensory level between adjacent items with repeated training, as proposed by the chaining model (Davachi & DuBrow, 2015; Lashley, 1951). This chaining mechanism could support the binding of elements (e.g., letters of a word and notes of music) to form holistic representations, which is commonly observed in both the visual and auditory cortices (Andrews et al., 2010; Fitch & Martins, 2014; Schiltz & Rossion, 2006; Wong & Gauthier, 2010).

In contrast, the cross-modal prediction is supported by an indirect pathway mediated by a higher-order area, that is, the SPL. The SPL is located at the dorsal portion of the posterior parietal cortex and receives input from sensory cortices. Anatomical and resting-state fMRI studies have found that the SPL has fiber connections and functional connectivity with the visual and auditory cortex (Lin et al., 2021; Makris et al., 2005, 2013; Wang et al., 2015). Previous studies have implicated the SPL in the visual–spatial attention shift (Ciaramelli et al., 2008; Corbetta & Shulman, 2002; Hutchinson et al., 2009). Several studies consistently found SPL activation when the attention was shifted between two sensory modalities (Renier et al., 2009; Shomstein & Yantis, 2004), suggesting a general role of attentional shift.

In addition, the SPL is also involved in memory processes, particularly in cross-modal conditions. First, the involvement of SPL in cross-modal integration has been extensively demonstrated (Molholm et al., 2006; Nakashita et al., 2008; Williams et al., 2015). Second, it shows stronger activation in the cross-modal working memory retrieval (Zhang et al., 2014) and cross-modal temporal order memory (Zhang et al., 2004) than in the unimodal conditions. Third, the SPL is engaged in retrieving spatial ("where") information (Kwok et al., 2012; Kwok & Macaluso, 2015). Finally, the SPL is also involved in the top-down prediction process (Balser et al., 2014; Liu et al., 2010; Mayer et al., 2015). For example, expectations about the identity of letters elicited increased power of pre-stimulus alpha oscillations in the SPL (Mayer et al., 2015), and the experts showed greater SPL activation than novices during the motion anticipation task (Balser et al., 2014).

These functions might underlie the important role of SPL in cross-modal prediction. An existing study has also found that the multisensory integration in the SPL could be accounted for by both the feedforward and feedback connections (Moran et al., 2008). The feedback connections from higher-order areas to primary sensory cortices were related to the top-down prediction process within the framework of predictive coding (Bar, 2007; Bastos et al., 2012; Clark, 2013; Friston, 2010).

Intriguingly, the current study found a novel asymmetric cross-modal predictive effect, with a stronger visual to auditory prediction than vice versa. This effect was not specific to our fMRI design, as it was found under both long and short ISI conditions. Since the predictive effect was calculated as the RT differences between the target words in the fixed sequences and that in the random sequences before and after learning, this asymmetry predicting effect should be not be contributed to the RT difference in predicting or predicted modality. It should be noted that we found that the overall predictive effect was reduced in the 4 s condition than in the 0.5 s condition, likely due to the fMRI environment, the longer ISI, and the visual judgment task during the long ISI. Nevertheless, we did not find significant ISI by predicting modality by modality shift three-way interaction in the behavioral study, suggesting a similarly asymmetric effect in both ISI conditions and the robustness of our results.

Several factors might have contributed to the asymmetric cross-modal predictive effect. First, existing studies have implicated the role of phonology in semantic access during Chinese word reading (Perfetti & Tan, 1998; Spinks et al., 2000). The phonological information activates rapidly and automatically (Tan & Perfetti, 1999) and precedes the direct access of meaning from orthography (Braun et al., 2015; Frost, 1998). This compulsory activation of phonological information might facilitate the prediction of the next auditory word. However, since the processing of semantic words and simple stimuli involves different cognitive neural mechanisms, future studies could manipulate the materials (words vs. simple lights/textures and tones) and task (semantic vs. non-semantic) to examine this hypothesis further.

Second, compared to the auditory system, the visual system is a more dominant sensory system (Spence et al., 2012) and response time to visual targets is usually faster than auditory targets under multisensory situations (Egeth & Sager, 1977; Koppen & Spence, 2007). Consistently, we found that the RT for visual words (~760 ms) was much faster than that for auditory words (~1000 ms). This could have two effects that contribute to the asymmetric cross-modal prediction. On the one hand, the processing of visual items is easier than auditory items, allowing more cognitive resources to be assigned to predict and process the following items. On the other hand, there is less room for improvement and thus less predicting effect for the visual words. Consistent with this notion, after accounting for the RT differences, the asymmetric effect was reduced. Notably, since the first effect benefits the VV condition and the second benefits the AA condition, there was a comparable predicting effect for VV and AA conditions.

Finally, previous studies have posited that the visual system is an active sensing system for seeking and acquiring information (Schroeder et al., 2010), whereas the auditory system is a passive system that mainly reacts to input (Golumbic et al., 2012). For example, visual leading context (e.g., lip movement) may prompt participants to simulate the corresponding sounds actively, but listening to auditory leading context does not prompt active engagement to the same degree (Sánchez-García et al., 2011). Consequently, the leading auditory information may serve as a general alerting mechanism, whereas the leading visual information can generate specific predictions and thus facilitate cross-modal predictions (Thorne & Debener, 2014). Nevertheless, auditory information could direct visual attention in space in some ecological conditions (e.g., a crash sound predicting a car accident). Future studies should examine how the information coded in different modalities could affect cross-modal predictions.

In contrast to the evidence for a pathway supporting visual-to-auditory prediction, our study failed to reveal a neural pathway from the auditory to the visual cortex. It should be noted that although we did not find significant auditory-to-visual prediction in the 4 s ISI condition, significant (although weaker) auditory-to-visual prediction was found in the 0.5 s ISI condition. Perhaps the indirect pathway mediated by higher-order areas can support behavioral facilitation under the long ISI condition, whereas other mechanisms, such as item-context binding, might support prediction under the short ISI condition (Davachi & DuBrow, 2015; Howard & Kahana, 2002). Future EEG or MEG studies could test this hypothesis by examining how the item-context binding and the indirect neural pathway contribute to cross-modal prediction under different ISI conditions.

To summarize, the current study found a novel and robust pattern of asymmetric cross-modal sequence prediction and further revealed distinct neural mechanisms underlying within- and cross-modal sequence predictions. These results emphasize the role of the SPL in supporting an indirect pathway for visual-to-auditory cross-modal prediction. Future studies are required to examine cross-modal prediction under different conditions, which will have significant implications for our understanding of psychiatric diseases with impairments in cross-modal prediction, such as autism (Stevenson et al., 2014), schizophrenia (Stekelenburg et al., 2013), or dyslexia (Blau et al., 2009).

ACKNOWLEDGMENTS

This work was sponsored by the National Natural Science Foundation of China (31730038), the Sino-German Collaborative Research Project "Crossmodal Learning" (NSFC 62061136001/DFG TRR169), the NSFC, and Israel Science Foundation (ISF) joint project (31861143040).

CONFLICT OF INTEREST STATEMENT

The authors declare no competing financial interests.

DATA AVAILABILITY STATEMENT

Both fMRI and behavior data are available at the Open Science Framework: https://osf.io/dn6ty/.

ORCID

Liang Shi https://orcid.org/0000-0002-7394-3857

REFERENCES

Alexander, W. H., & Brown, J. W. (2014). A general role for medial prefrontal cortex in event prediction. *Frontiers in Computational Neuroscience*, *8*, 69.

Alink, A., Walther, A., Krugliak, A., & Van Den, J. J. F. (2015). Mind the drift – Improving sensitivity to fMRI pattern information by accounting for temporal pattern drift. *bioRxiv*, 17–19.

Andrews, T. J., Davies-Thompson, J., Kingstone, A., & Young, A. W. (2010). Internal and external features of the face are represented holistically in face-selective regions of the visual cortex. *The Journal of Neuroscience*, *30*, 3544–3552. https://doi.org/10.1523/JNEUROSCI.4863-09.2010

Anzellotti, S., & Coutanche, M. N. (2018). Beyond functional connectivity: Investigating networks of multivariate representations. *Trends in Cognitive Sciences*, *22*, 258–269. https://doi.org/10.1016/j.tics.2017.12.002

Arnal, L. H., Morillon, B., Kell, C. A., & Giraud, A.-L. (2009). Dual neural routing of visual facilitation in speech processing. *The Journal of Neuroscience*, *29*, 13445–13453.

Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A., & Gee, J. C. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*, *54*, 2033–2044.

Balser, N., Lorey, B., Pilgramm, S., Stark, R., Bischoff, M., Zentgraf, K., Williams, A. M., & Munzert, J. (2014). Prediction of human actions: Expertise and task-related effects on neural activation of the action observation network. *Human Brain Mapping*, *35*, 4016–4034. https://doi.org/10.1002/hbm.22455

Bar, M. (2007). The proactive brain: Using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, *11*, 280–289. https://doi.org/10.1016/j.tics.2007.05.005

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*, 695–711.

Beer, A. L., Plank, T., & Greenlee, M. W. (2011). Diffusion tensor imaging shows white matter tracts between human auditory and visual cortex. *Experimental Brain Research*, *213*, 299–308.

Beer, A. L., Plank, T., Meyer, G., & Greenlee, M. W. (2013). Combined diffusion-weighted and functional magnetic resonance imaging reveals a temporal-occipital network involved in auditory-visual object processing. *Frontiers in Integrative Neuroscience*, *7*, 5.

Bellmund, J. L., Deuker, L., & Doeller, C. F. (2019). Mapping sequence structure in the human lateral entorhinal cortex. *eLife*, *8*, 1–20. https://doi.org/10.7554/elife.45333

Bieler, M., Sieben, K., Schildt, S., Röder, B., & Hanganu-Opatz, I. L. (2017). Visual–tactile processing in primary somatosensory cortex emerges before cross-modal experience. *Synapse*, *71*, e21958.

Blau, V., van Atteveldt, N., Ekkebus, M., Goebel, R., & Blomert, L. (2009). Reduced neural integration of letters and speech sounds links phonological and reading deficits in adult dyslexia. *Current Biology*, *19*, 503–508.

Borders, A. A., Aly, M., Parks, C. M., & Yonelinas, A. P. (2017). The hippocampus is particularly important for building associations across stimulus domains. *Neuropsychologia*, *99*, 335–342.

Braun, M., Hutzler, F., Münte, T. F., Rotte, M., Dambacher, M., Richlan, F., & Jacobs, A. M. (2015). The neural bases of the pseudohomophone effect: Phonological constraints on lexico-semantic access in reading. *Neuroscience*, *295*, 151–163. https://doi.org/10.1016/j.neuroscience.2015.03.035

Brunec, I. K., & Momennejad, I. (2022). Predictive representations in hippocampal and prefrontal hierarchies. *The Journal of Neuroscience*, *42*, 299–312. https://doi.org/10.1523/JNEUROSCI.1327-21.2021

Butler, A. J., & James, K. H. (2011). Cross-modal versus within-modal recall: Differences in behavioral and brain responses. *Behavioural Brain Research*, *224*, 387–396.

Cai, M. B., Schuck, N. W., Pillow, J. W., & Niv, Y. (2019). Representational structure or task structure? Bias in neural representational similarity analysis and a Bayesian method for reducing bias. *PLoS Computational Biology*, *15*, 1–30. https://doi.org/10.1371/journal.pcbi.1006299

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, *10*, 649–657. https://doi.org/10.1016/S0960-9822(00)00513-3

Ciaramelli, E., Grady, C. L., & Moscovitch, M. (2008). Top-down and bottom-up attention to memory: A hypothesis (AtoM) on the role of the posterior parietal cortex in memory retrieval. *Neuropsychologia*, *46*, 1828–1851. https://doi.org/10.1016/j.neuropsychologia.2008.03.022

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences*, *36*, 181–204.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews. Neuroscience*, *3*, 201–215. https://doi.org/10.1038/nrn755

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*, 162–173.

Davachi, L., & DuBrow, S. (2015). How the hippocampus preserves order: The role of prediction and context. *Trends in Cognitive Sciences*, *19*, 92–99. https://doi.org/10.1016/j.tics.2014.12.004

de Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception? *Trends in Cognitive Sciences*, *22*, 764–779. https://doi.org/10.1016/j.tics.2018.06.002

Demarchi, G., Sanchez, G., & Weisz, N. (2019). Automatic and feature-specific prediction-related neural activity in the human auditory system. *Nature Communications*, *10*, 1–11. https://doi.org/10.1038/s41467-019-11440-1

Devereux, B. J., Clarke, A., Marouchos, A., & Tyler, L. K. (2013). Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. *The Journal of Neuroscience*, *33*, 18906–18916. https://doi.org/10.1523/JNEUROSCI.3809-13.2013

Diedrichsen, J., Provost, S., & Zareamoghaddam, H. (2016). On the distribution of cross-validated Mahalanobis distances. *arXiv*, 1–24.

Diedrichsen, J., Ridgway, G. R., Friston, K. J., & Wiestler, T. (2011). Comparing the similarity and spatial structure of neural representations: A pattern-component model. *NeuroImage*, *55*, 1665–1678. https://doi.org/10.1016/j.neuroimage.2011.01.044

Driver, J., & Noesselt, T. (2008). Multisensory interplay reveals crossmodal influences on "sensory-specific" brain regions, neural responses, and judgments. *Neuron*, *57*, 11–23. https://doi.org/10.1016/j.neuron.2007.12.013

Egeth, H. E., & Sager, L. C. (1977). On the locus of visual dominance. *Perception & Psychophysics*, *22*, 77–86.

Ekman, M., Kok, P., & De Lange, F. P. (2017). Time-compressed preplay of anticipated events in human primary visual cortex. *Nature Communications*, *8*, 1–9. https://doi.org/10.1038/ncomms15276

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., & Snyder, M. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*, 111–116.

Fitch, W. T., & Martins, M. D. (2014). Hierarchical processing in music, language, and action: Lashley revisited. *Annals of the New York Academy of Sciences*, *1316*, 87–104. https://doi.org/10.1111/nyas.12406

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews. Neuroscience*, *11*, 127–138. https://doi.org/10.1038/nrn2787

Frost, R. (1998). Toward a strong phonological theory of visual word recognition: True issues and false trails. *Psychological Bulletin*, *123*, 71–99.

Ghazanfar, A. A., & Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends in Cognitive Sciences*, *10*, 278–285.

Golumbic, E. M. Z., Poeppel, D., & Schroeder, C. E. (2012). Temporal context in speech processing and attentional stream selection: A behavioral and neural perspective. *Brain and Language*, *122*, 151–161.

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., & Halchenko, Y. O. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, *3*, 1–9.

Handjaras, G., Leo, A., Cecchetti, L., Papale, P., Lenci, A., Marotta, G., Pietrini, P., & Ricciardi, E. (2017). Modality-independent encoding of individual concepts in the left parietal cortex. *Neuropsychologia*, *105*, 39–49. https://doi.org/10.1016/j.neuropsychologia.2017.05.001

Henriksson, L., Khaligh-Razavi, S. M., Kay, K., & Kriegeskorte, N. (2015). Visual representations are dominated by intrinsic fluctuations correlated between areas. *NeuroImage*, *114*, 275–286. https://doi.org/10.1016/j.neuroimage.2015.04.026

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*, 269–299. https://doi.org/10.1006/jmps.2001.1388

Hsieh, L. T., Gruber, M. J., Jenkins, L. J., & Ranganath, C. (2014). Hippocampal activity patterns carry information about objects in temporal context. *Neuron*, *81*, 1165–1178. https://doi.org/10.1016/j.neuron.2014.01.015

Hutchinson, J. B., Uncapher, M. R., & Wagner, A. D. (2009). Posterior parietal cortex and episodic retrieval: Convergent and divergent effects of attention and memory. *Learning & Memory*, *16*, 343–356. https://doi.org/10.1101/lm.919109

Jenkins, L. J., & Ranganath, C. (2010). Prefrontal and medial temporal lobe activity at encoding predicts temporal context memory. *The Journal of Neuroscience*, *30*, 15558–15565. https://doi.org/10.1523/JNEUROSCI.1337-10.2010

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, *17*, 825–841.

Jung, Y., Larsen, B., & Walther, D. B. (2018). Modality-independent coding of scene categories in prefrontal cortex. *The Journal of Neuroscience*, *38*, 5969–5981.

Kayser, C., & Logothetis, N. K. (2007). Do early sensory cortices integrate cross-modal information? *Brain Structure & Function*, *212*, 121–132. https://doi.org/10.1007/s00429-007-0154-0

Kok, P., Brouwer, G. J., van Gerven, M. A. J., & de Lange, F. P. (2013). Prior expectations bias sensory representations in visual cortex. *The Journal of Neuroscience*, *33*, 16275–16284. https://doi.org/10.1523/JNEUROSCI.0742-13.2013

Kok, P., Mostert, P., & De Lange, F. P. (2017). Prior expectations induce prestimulus sensory templates. *Proceedings of the National Academy of Sciences of the United States of America*, *114*, 10473–10478. https://doi.org/10.1073/pnas.1705652114

Koppen, C., & Spence, C. (2007). Seeing the light: Exploring the Colavita visual dominance effect. *Experimental Brain Research*, *180*, 737–754.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4.

Ku, Y., Zhao, D., Hao, N., Hu, Y., Bodner, M., & Di, Z. Y. (2015). Sequential roles of primary somatosensory cortex and posterior parietal cortex in tactile-visual cross-modal working memory: A single-pulse transcranial magnetic stimulation (spTMS) study. *Brain Stimulation*, *8*, 88–91. https://doi.org/10.1016/j.brs.2014.08.009

Kwok, S. C., & Macaluso, E. (2015). Immediate memory for "when, where and what": Short-delay retrieval using dynamic naturalistic material. *Human Brain Mapping*, *36*, 2495–2513. https://doi.org/10.1002/hbm.22787

Kwok, S. C., Shallice, T., & Macaluso, E. (2012). Functional anatomy of temporal organisation and domain-specificity of episodic memory retrieval. *Neuropsychologia*, *50*, 2943–2955. https://doi.org/10.1016/j.neuropsychologia.2012.07.025

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior* (pp. 112–131). Wiley.

Lee, C. S., Aly, M., & Baldassano, C. (2021). Anticipation of temporally structured events in the brain. *eLife*, *10*, 1–15. https://doi.org/10.7554/ELIFE.64972

Lehn, H., Steffenach, H. A., Van Strien, N. M., Veltman, D. J., Witter, M. P., & Håberg, A. K. (2009). A specific role of the human hippocampus in recall of temporal sequences. *The Journal of Neuroscience*, *29*, 3475–3484. https://doi.org/10.1523/JNEUROSCI.5370-08.2009

Lin, Y.-H., Dadario, N. B., Hormovas, J., Young, I. M., Briggs, R. G., MacKenzie, A. E., Palejwala, A. H., Fonseka, R. D., Kim, S. J., & Tanglay, O. (2021). Anatomy and white matter connections of the superior parietal lobule. *Operative Neurosurgery*, *21*, E199–E214.

Liu, J., Li, J., Zhang, H., Rieth, C. A., Huber, D. E., Li, W., Lee, K., & Tian, J. (2010). Neural correlates of top-down letter processing. *Neuropsychologia*, *48*, 636–641. https://doi.org/10.1016/j.neuropsychologia.2009.10.024

Liuzzi, A. G., Ubaldi, S., & Fairhall, S. L. (2021). Representations of conceptual information during automatic and active semantic access. *Neuropsychologia*, *160*, 107953. https://doi.org/10.1016/j.neuropsychologia.2021.107953

Macaluso, E., & Driver, J. (2005). Multisensory spatial interactions: A window onto functional integration in the human brain. *Trends in Neurosciences*, *28*, 264–271.

Macaluso, E., Frith, C. D., & Driver, J. (2000). Modulation of human visual cortex by crossmodal spatial attention. *Science (80-)*, *289*, 1206–1208.

Makris, N., Kennedy, D. N., McInerney, S., Sorensen, A. G., Wang, R., Caviness, V. S., & Pandya, D. N. (2005). Segmentation of subcomponents within the superior longitudinal fascicle in humans: A quantitative, in vivo, DT-MRI study. *Cerebral Cortex*, *15*, 854–869. https://doi.org/10.1093/cercor/bhh186

Makris, N., Preti, M. G., Wassermann, D., Rathi, Y., Papadimitriou, G. M., Yergatian, C., Dickerson, B. C., Shenton, M. E., & Kubicki, M. (2013). Human middle longitudinal fascicle: Segregation and behavioral-clinical implications of two distinct fiber connections linking temporal pole and superior temporal gyrus with the angular gyrus or superior parietal lobule using multi-tensor tractography. *Brain Imaging and Behavior*, *7*, 335–352. https://doi.org/10.1007/s11682-013-9235-2

Mayer, A., Schwiedrzik, C. M., Wibral, M., Singer, W., & Melloni, L. (2015). Expecting to see a letter: Alpha oscillations as carriers of top-down sensory predictions. *Cerebral Cortex*, *26*, 3146–3160.

Molholm, S., Sehatpour, P., Mehta, A. D., Shpaner, M., Gomez-Ramirez, M., Ortigue, S., Dyke, J. P., Schwartz, T. H., & Foxe, J. J. (2006). Audiovisual multisensory integration in superior parietal lobule revealed by human intracranial recordings. *Journal of Neurophysiology*, *96*, 721–729. https://doi.org/10.1152/jn.00285.2006

Moran, R. J., Molholm, S., Reilly, R. B., & Foxe, J. J. (2008). Changes in effective connectivity of human superior parietal lobule under multisensory and unisensory stimulation. *The European Journal of Neuroscience*, *27*, 2303–2312. https://doi.org/10.1111/j.1460-9568.2008.06187.x

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, *59*, 2636–2643.

Nakashita, S., Saito, D. N., Kochiyama, T., Honda, M., Tanabe, H. C., & Sadato, N. (2008). Tactile-visual integration in the posterior parietal cortex: A functional magnetic resonance imaging study. *Brain Research Bulletin*, *75*, 513–525. https://doi.org/10.1016/j.brainresbull.2007.09.004

Paz, R., Gelbard-Sagiv, H., Mukamel, R., Harel, M., Malach, R., & Fried, I. (2010). A neural substrate in the human hippocampus for linking suc-

cessive events. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 6046–6051. https://doi.org/10.1073/pnas.0910834107

Perfetti, C. A., & Tan, L. H. (1998). The time course of graphic, phonological, and semantic activation in Chinese character identification. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *24*, 101–118.

Pillai, A. S., Gilbert, J. R., & Horwitz, B. (2013). Early sensory cortex is activated in the absence of explicit input during crossmodal item retrieval: Evidence from MEG. *Behavioural Brain Research*, *238*, 265–272.

Reddy, L., Poncet, M., Self, M. W., Peters, J. C., Douw, L., Van Dellen, E., Claus, S., Reijneveld, J. C., Baayen, J. C., & Roelfsema, P. R. (2015). Learning of anticipatory responses in single neurons of the human medial temporal lobe. *Nature Communications*, *6*, 8556. https://doi.org/10.1038/ncomms9556

Regev, M., Honey, C. J., Simony, E., & Hasson, U. (2013). Selective and invariant neural responses to spoken and written narratives. *The Journal of Neuroscience*, *33*, 15978–15988.

Renier, L. A., Anurova, I., De Volder, A. G., Carlson, S., VanMeter, J., & Rauschecker, J. P. (2009). Multisensory integration of sounds and vibrotactile stimuli in processing streams for "what" and "where". *The Journal of Neuroscience*, *29*, 10950–10960.

Ritchie, J. B., Lee Masson, H., Bracci, S., & Op de Beeck, H. P. (2021). The unreliable influence of multivariate noise normalization on the reliability of neural dissimilarity. *NeuroImage*, *245*, 118686. https://doi.org/10.1016/j.neuroimage.2021.118686

Ross, R. S., Brown, T. I., & Stern, C. E. (2009). The retrieval of learned sequences engages the hippocampus: Evidence from fMRI. *Hippocampus*, *19*, 790–799. https://doi.org/10.1002/hipo.20558

Sánchez-García, C., Alsius, A., Enns, J. T., & Soto-Faraco, S. (2011). Cross-modal prediction in speech perception. *PLoS One*, *6*, e25198.

Schapiro, A. C., Kustner, L. V., & Turk-Browne, N. B. (2012). Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Current Biology*, *22*, 1622–1627. https://doi.org/10.1016/j.cub.2012.06.056

Schiltz, C., & Rossion, B. (2006). Faces are represented holistically in the human occipito-temporal cortex. *NeuroImage*, *32*, 1385–1394. https://doi.org/10.1016/j.neuroimage.2006.05.037

Schroeder, C. E., Wilson, D. A., Radman, T., Scharfman, H., & Lakatos, P. (2010). Dynamics of active sensing and perceptual selection. *Current Opinion in Neurobiology*, *20*, 172–176.

Senoussi, M., VanRullen, R., & Reddy, L. (2020). Anticipatory reinstatement of expected perceptual events during visual sequence learning. *bioRxiv*. https://doi.org/10.1101/2020.11.28.402123

Shomstein, S., & Yantis, S. (2004). Control of attention shifts between vision and audition in human cortex. *The Journal of Neuroscience*, *24*, 10702–10706. https://doi.org/10.1523/JNEUROSCI.2939-04.2004

Spence, C., Parise, C., & Chen, Y.-C. (2012). The Colavita visual dominance effect. In *The neural bases of multisensory processes*. CRC Press/Taylor & Francis.

Spinks, J. A., Ying, L., Perfetti, C. A., & Li, H. T. (2000). Reading Chinese characters for meaning: The role of phonological information. *Cognition*, *76*, B1–B11.

Stekelenburg, J. J., Maes, J. P., Van Gool, A. R., Sitskoorn, M., & Vroomen, J. (2013). Deficient multisensory integration in schizophrenia: An event-related potential study. *Schizophrenia Research*, *147*, 253–261.

Stevenson, R. A., Siemann, J. K., Schneider, B. C., Eberly, H. E., Woynaroski, T. G., Camarata, S. M., & Wallace, M. T. (2014). Multisensory temporal integration in autism spectrum disorders. *The Journal of Neuroscience*, *34*, 691–697.

Tan, L. H., & Perfetti, C. A. (1999). Phonological activation in visual identification of Chinese two-character words. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *25*, 382–393.

Tanabe, H. C., Honda, M., & Sadato, N. (2005). Functionally segregated neural substrates for arbitrary audiovisual paired-association learning. *The Journal of Neuroscience*, *25*, 6409–6418. https://doi.org/10.1523/JNEUROSCI.0636-05.2005

Thorne, J. D., & Debener, S. (2014). Look now and hear what's coming: On the functional role of cross-modal phase reset. *Hearing Research*, *307*, 144–152. https://doi.org/10.1016/j.heares.2013.07.002

Tibon, C. R., Fuhrmann, D., Levy, D. A., Simons, J. S., & Henson, R. N. (2019). Multimodal integration and vividness in the angular gyrus during episodic encoding and retrieval. *The Journal of Neuroscience*, *39*, 4365–4374. https://doi.org/10.1523/JNEUROSCI.2102-18.2018

Turk-Browne, N. B., Scholl, B. J., Johnson, M. K., & Chun, M. M. (2010). Implicit perceptual anticipation triggered by statistical learning. *The Journal of Neuroscience*, *30*, 11177–11187. https://doi.org/10.1523/JNEUROSCI.0858-10.2010

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging*, *29*, 1310–1320.

Van Atteveldt, N., Formisano, E., Goebel, R., & Blomert, L. (2004). Integration of letters and speech sounds in the human brain. *Neuron*, *43*, 271–282. https://doi.org/10.1016/j.neuron.2004.06.025

Vandenberghe, R., Price, C., Wise, R., Josephs, O., & Frackowiak, R. S. J. (1996). Functional anatomy of a common semantic system for words and pictures. *Nature*, *383*, 254–256. https://doi.org/10.1038/383254a0

Wang, J., Yang, Y., Fan, L., Xu, J., Li, C., Liu, Y., Fox, P. T., Eickhoff, S. B., Yu, C., & Jiang, T. (2015). Convergent functional architecture of the superior parietal lobule unraveled with multimodal neuroimaging approaches. *Human Brain Mapping*, *36*, 238–257.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, *137*, 188–200. https://doi.org/10.1016/j.neuroimage.2015.12.012

Williams, J. T., Darcy, I., & Newman, S. D. (2015). Modality-independent neural mechanisms for novel phonetic processing. *Brain Research*, *1620*, 107–115. https://doi.org/10.1016/j.brainres.2015.05.014

Wong, Y. K., & Gauthier, I. (2010). Holistic processing of musical notation: Dissociating failures of selective attention in experts and novices. *Cognitive, Affective, & Behavioral Neuroscience*, *10*, 541–551. https://doi.org/10.3758/CABN.10.4.541

Xiao, X., Dong, Q., Gao, J., Men, W., Poldrack, R. A., & Xue, G. (2017). Transformed neural pattern reinstatement during episodic memory retrieval. *The Journal of Neuroscience*, *37*, 2986–2998. https://doi.org/10.1523/JNEUROSCI.2324-16.2017

Xue, G., Chen, C., Jin, Z., & Dong, Q. (2006). Language experience shapes fusiform activation when processing a logographic artificial language: An fMRI training study. *NeuroImage*, *31*, 1315–1326.

Xue, G., & Poldrack, R. A. (2007). The neural substrates of visual perceptual learning of words: Implications for the visual word form area hypothesis. *Journal of Cognitive Neuroscience*, *19*, 1643–1655.

Yazar, Y., Bergström, Z. M., & Simons, J. S. (2017). Reduced multimodal integration of memory features following continuous theta burst stimulation of angular gyrus. *Brain Stimulation*, *10*, 624–629. https://doi.org/10.1016/j.brs.2017.02.011

Yokoi, A., & Diedrichsen, J. (2019). Neural organization of hierarchical motor sequence representations in the human neocortex. *Neuron*, *103*, 1178–1190.e7. https://doi.org/10.1016/j.neuron.2019.06.017

Yushkevich, P. A., Pluta, J. B., Wang, H., Xie, L., Ding, S., Gertje, E. C., Mancuso, L., Kliot, D., Das, S. R., & Wolk, D. A. (2015). Automated volumetry and regional thickness analysis of hippocampal subfields and medial temporal cortical structures in mild cognitive impairment. *Human Brain Mapping*, *36*, 258–287.

Zhang, D., Zhang, X., Sun, X., Li, Z., Wang, Z., He, S., & Hu, X. (2004). Cross-modal temporal order memory for auditory digits and visual

locations: An fMRI study. *Human Brain Mapping*, *22*, 280–289. https://doi.org/10.1002/hbm.20036

Zhang, Y., Hu, Y., Guan, S., Hong, X., Wang, Z., & Li, X. (2014). Neural substrate of initiation of cross-modal working memory retrieval. *PLoS One*, *9*, e103991.

Zhao, L., Chunhui, C., Shao, L., Wang, Y., Xiao, X., Chuansheng, C., Yang, J., Zevin, J., & Xue, G. (2017). Orthographic and phonological representations in the fusiform cortex. *Cerebral Cortex*, *27*, 5197–5210. https://doi.org/10.1093/cercor/bhw300

Zhou, Y.-D., & Fuster, J. M. (2000). Visuo-tactile cross-modal associations in cortical somatosensory cells. *Proceedings of the National Academy of Sciences*, *97*, 9777–9782.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Shi, L., Liu, C., Peng, X., Cao, Y., Levy, D. A., & Xue, G. (2023). The neural representations underlying asymmetric cross-modal prediction of words. *Human Brain Mapping*, *44*(6), 2418–2435. https://doi.org/10.1002/hbm.26219