The Spatiotemporal Neural Dynamics of Intersensory Attention Capture of Salient Stimuli: A Large-Scale Auditory-Visual Modeling Study

Qin Liu[1,2†], Antonio Ulloa[1,3] and Barry Horwitz[1]*

[1] Brain Imaging and Modeling Section, National Institute on Deafness and Other Communication Disorders, National Institutes of Health, Bethesda, MD, United States, [2] Department of Physics, University of Maryland, College Park, College Park, MD, United States, [3] Center for Information Technology, National Institutes of Health, Bethesda, MD, United States

The spatiotemporal dynamics of the neural mechanisms underlying endogenous (top-down) and exogenous (bottom-up) attention, and how attention is controlled or allocated in intersensory perception are not fully understood. We investigated these issues using a biologically realistic large-scale neural network model of visual-auditory object processing of short-term memory. We modeled and incorporated into our visual-auditory object-processing model the temporally changing neuronal mechanisms for the control of endogenous and exogenous attention. The model successfully performed various bimodal working memory tasks, and produced simulated behavioral and neural results that are consistent with experimental findings. Simulated fMRI data were generated that constitute predictions that human experiments could test. Furthermore, in our visual-auditory bimodality simulations, we found that increased working memory load in one modality would reduce the distraction from the other modality, and a possible network mediating this effect is proposed based on our model.

INTRODUCTION

Large-scale, biologically realistic neural modeling has become a critical tool in the effort to determine the mechanisms by which neural activity results in high-level cognitive processing, such as working memory. Our laboratory has investigated a number of working memory tasks in humans using functional neuroimaging and large-scale neural modeling (LSNM) in both the visual and auditory modalities. In this paper, we combine our visual and auditory models through a node representing the anterior insula to investigate the spatiotemporal dynamics of the neural mechanisms underlying endogenous (top-down) and exogenous (bottom-up) attention, and how attention is controlled or allocated in intersensory perception during several working memory tasks.

Attention is a crucial cognitive function enabling humans and other animals to select goal-relevant information from among a vast number of sensory stimuli in the environment. On the other hand, attention can also be captured by salient goal-irrelevant distractors. This mechanism, allowing us to focus on behavioral goals while staying vigilant to environmental changes, is

usually described as two separate types of attention: endogenous (voluntary/goal-driven) attention and exogenous (involuntary/stimulus-driven) attention (Hopfinger and West, 2006). Endogenous attention for object features is thought to be controlled by a top-down process, starting from the frontal lobe and connecting back to early sensory areas (Kastner and Ungerleider, 2000; Koechlin et al., 2003; Bichot et al., 2015; D'Esposito and Postle, 2015; Leavitt et al., 2017; Mendoza-Halliday and Martinez-Trujillo, 2017). In contrast, exogenous attention behaves primarily in a bottom-up manner, triggered by stimuli that may be task irrelevant but salient in a given context (Yantis and Jonides, 1990; Hopfinger and West, 2006; Clapp et al., 2010; Bowling et al., 2020).

Working memory, the brain process by which selected information is temporarily stored and manipulated, relies on endogenous attention for protection from distractions (Berti and Schroger, 2003; Lorenc et al., 2021). However, working memory is not completely protected but is capable of handling unexpected and salient distractions mediated by exogenous attention (Berti et al., 2004). Early studies on the relationship between working memory and attention focused mostly on the role of endogenous attention in working memory encoding and maintenance (Baddeley, 1986, 1996). Later, some functional neuroimaging and behavioral studies showed that working memory can also control exogenous attention and reduce distractions (Berti and Schroger, 2003; Berti et al., 2004; Spinks et al., 2004; SanMiguel et al., 2008; Clapp et al., 2010). However, little is known about the brain networks mediating such effects. The aim of the present study was to investigate and propose a possible neural network mechanism of how endogenous and exogenous attention interact with each other, and how working memory controls exogenous attention switching. We restricted our analysis to the storage component of working memory (i.e., short-term memory).

We, among others, believe that computational modeling is a powerful tool for helping determine the neural mechanisms mediating cognitive functions (Horwitz et al., 1999, 2005; Deco et al., 2008; Friston, 2010; Jirsa et al., 2010; Eliasmith et al., 2012; Kriegeskorte and Diedrichsen, 2016; Bassett et al., 2018; Yang et al., 2019; Ito et al., 2020; Pulvermuller et al., 2021). With respect to working memory, Tagamets and Horwitz (1998) and Horwitz and Tagamets (1999) developed a large-scale dynamic neural model of visual object short-term memory. The model consisted of elements representing the interconnected neuronal populations comprising the cortical ventral pathway that processes primarily the features of visual objects (Ungerleider and Mishkin, 1982; Mishkin et al., 1983; Haxby et al., 1991). Later an auditory object processing model was built that functioned in an analogous fashion to the visual model (Husain et al., 2004). The two LSNMs were each designed to perform a short-term recognition memory delayed match-to-sample (DMS) task. During each trial of the task, a stimulus S1 is presented for a certain amount of time, followed by a delay period in which S1 must be kept in short-term memory. When a second stimulus (S2) is presented, the model responds as to whether S2 matches S1. Recently, the visual model was extended to be able to manage distractors and multiple objects in short-term memory (Liu et al., 2017). The extended visual model successfully performed

the DMS task with distractors and Sternberg's recognition task (Sternberg, 1969) where subjects are asked to remember a list of items and indicate whether a probe is on the list.

Here we present a simulation study of intersensory (auditory and visual) attention switching and the interaction between endogenous and exogenous attention. The term intersensory attention refers to the ability to attend to stimuli from one sensory modality while ignoring stimuli from other modalities (Keil et al., 2016). We first combine and extend the aforementioned LSNMs to incorporate "exogenous attention" (the original models already included one type of "endogenous attention"). We add a pair of modules representing "exogenous attention" for auditory and visual processing. These two modules compete with each other based on the salience of auditory and visual stimuli and assign the value of attention together with endogenous attention. Endogenous attention is set according to task specification before each simulation. Then we simulate intersensory attention allocation and various bimodal (i.e., auditory and visual) short-term memory tasks. Simulations presented below show the "working memory load effect," i.e., higher working memory load in one modality reduces the distraction from another modality, which has been reported in a number of experimental studies (Berti and Schroger, 2003; Spinks et al., 2004; SanMiguel et al., 2008). Furthermore, we also show that higher working memory load can increase distraction from the same modality. We propose the neural mechanism that underlies intersensory attention switching and how this mechanism results in working memory load modulating attention allocation between different modalities.

MATERIALS AND METHODS

Large-scale neural network modeling aims at formulating and testing hypotheses about how the brain can carry out a specific function under investigation. Generally, the hypotheses underlying the model are instantiated in a computational framework and quantitative relationships are generated that can be explicitly compared with experimental data (Horwitz et al., 1999). Because the network paradigm now has become central in cognitive neuroscience (and especially in human studies), neural network modeling has emerged as an essential tool for interpreting neuroimaging data, and as well, integrating neuroimaging data with the other kinds of data employed by cognitive neuroscientists (Horwitz et al., 2000; Bassett et al., 2018; Kay, 2018; Naselaris et al., 2018; Pulvermuller et al., 2021).

A large assortment of neural network models has been developed, with different types aimed at addressing different questions. In two extensive reviews, Bassett and her colleagues discussed some of the various kinds of neural network models that have recently emerged (Bassett et al., 2018; Lynn and Bassett, 2019). A key distinction is made between model networks of artificial neurons and model biophysical networks (Lynn and Bassett, 2019). Deep learning networks (see LeCun et al., 2015; Saxe et al., 2021 for reviews) and recurrent neural networks (Song et al., 2016) illustrate the former, whereas a model of the ventral visual object processing pathway (Ulloa and Horwitz, 2016) and a

model of visual attention (Corchs and Deco, 2002) are examples of the latter. However, this distinction is not a binary one since there has been recent work, for instance, in using deep learning models to understand the neural basis of cognition (Cichy et al., 2016; Devereux et al., 2018), and in employing recurrent network models to investigate information maintenance and manipulation in working memory (Masse et al., 2019).

The combined auditory-visual large-scale neural network model used in this paper is a biophysical cortical network model. It consists of three types of sub-models: a structural model representing the neuroanatomical relationships between modules; a functional model indicating how each basic unit of each module represents neural activity; and a hemodynamic model indicating how neural activity is converted into BOLD fMRI activity. The simulated visual inputs to the model correspond to simple shapes, and the simulated auditory inputs correspond to frequency-time patterns. The model's outputs consist of simulated neural activity, simulated regional fMRI activity and simulated human behavior on 15 related short-term memory tasks. Because the focus of this paper is on the interaction between exogenous and endogenous attention, the auditory and visual models are connected *via* a module representing the anterior insula (see below, where details of the combined model are provided).

The Large-Scale Neural Model Network

The structural network of the combined auditory and visual model, representing the neuroanatomical relations between network modules, is shown in **Figure 1A**. Both the visual and auditory sub-models are organized as hierarchal networks, based on empirical data obtained from nonhuman primates and humans (Ungerleider and Mishkin, 1982; Ungerleider and Haxby, 1994; Rauschecker, 1997; Kaas and Hackett, 1999; Nourski, 2017).

As the basic units of our model, we use a variant of Wilson-Cowan units (Wilson and Cowan, 1972), which consists of one excitatory unit and one inhibitory unit (see **Figure 1B**). One basic unit can be considered as a simplified representation of a cortical column. Each module of the auditory subnetwork, originally developed by Husain et al. (2004), is explained in detail below. Submodules of A1 and A2 are organized as 1 × 81 arrays of basic units, and all the other modules are 9 × 9 arrays of basic units (see details below). The structure of the visual model is similar to the auditory model in many ways (one exception: the V1/V2 and V4 modules are 9 × 9 arrays of basic units); for the details of the visual model (see Tagamets and Horwitz, 1998; Ulloa and Horwitz, 2016; Liu et al., 2017). **Figure 1C** shows the visual and auditory models embedded in the human structural connectome provided by Hagmann et al. (2008). We used published empirical findings to posit the hypothetical brain regions of interest (ROIs) corresponding to each module in our computational model and the corresponding nodes in Hagmann's connectome. Then we embedded our revised model of microcircuits and network structure into the connectome (see Ulloa and Horwitz, 2016, for details). We ran the simulations using our in-house simulator in parallel with Hagmann's connectome using The Virtual Brain

(TVB) software (Sanz Leon et al., 2013) (see below). Details about our simulation framework are discussed in the **Appendix**.

Module A1

The auditory model of Husain et al. (2004) that we extended here was designed to process one kind of auditory object. As pointed out by Griffiths and Warren (2004), an individual auditory object consists of sound source and sound event information (e.g., the voice of a speaker and the word produced by the speaker). In our model, the simulated auditory object of interest consists only of the sound event component—the spectrotemporal pattern of information (what we call a tonal contour; see **Figure 2A**). The duration of these patterns is meant to represent sounds whose duration is that of a single syllable word (∼200–300 ms).
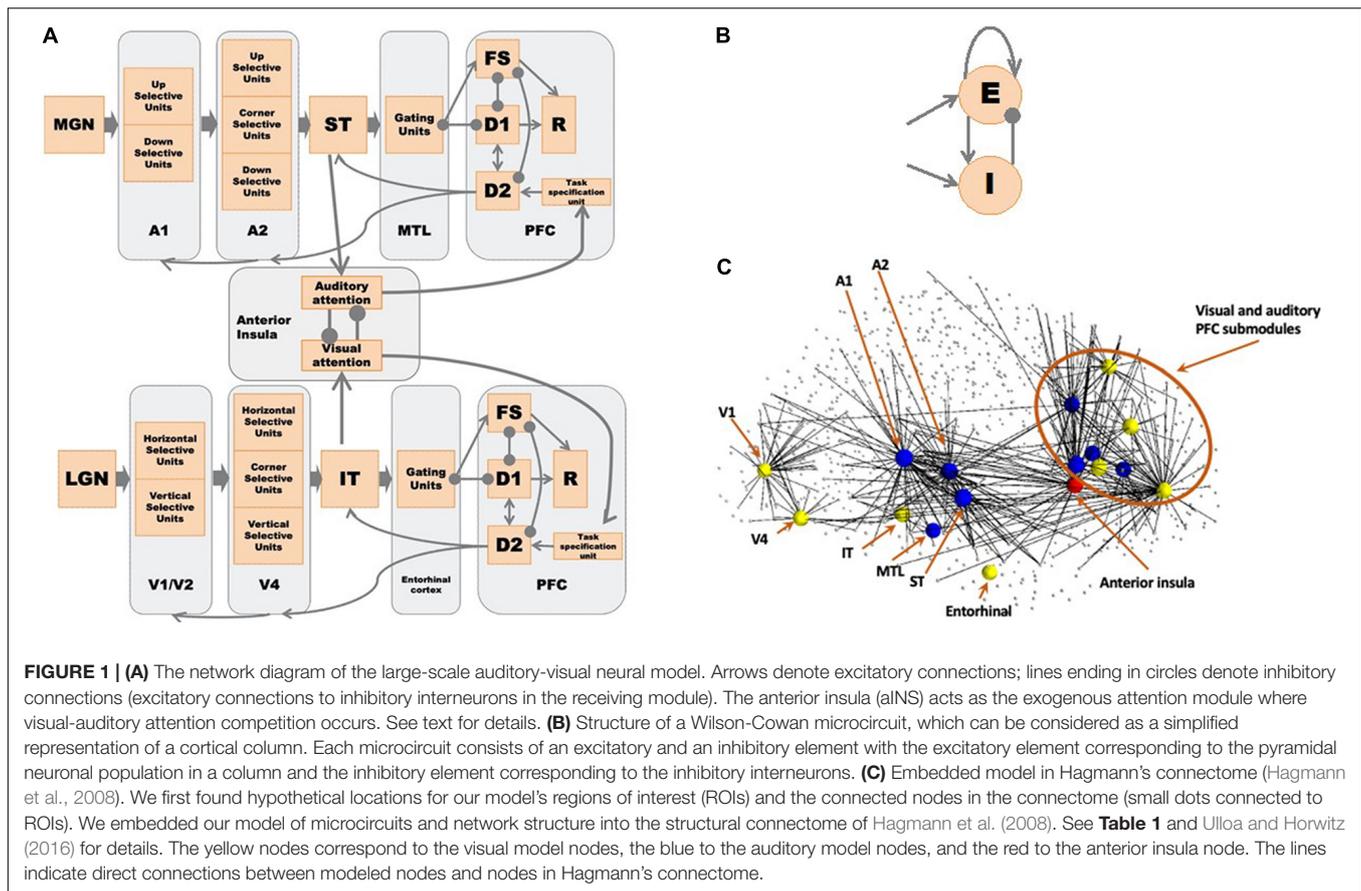
In the model the early cortical auditory areas are combined as A1, which is analogous to the V1/V2 module in the visual model; all simulated auditory (visual) inputs enter the model *via* the A1(V1/V2) module. A1 corresponds to the core/belt area in monkeys (Rauschecker, 1998; Kaas and Hackett, 1999) and the primary auditory area in the transverse temporal gyrus in human (putative Brodmann Area 41; Talairach and Tournoux, 1988). Based on experimental evidence that the neurons in early auditory areas are responsive to the direction of frequency modulated sweeps (Mendelson and Cynader, 1985; Mendelson et al., 1993; Shamma et al., 1993; Bieser, 1998; Tian and Rauschecker, 2004; Godey et al., 2005; Kikuchi et al., 2010; Hsieh et al., 2012), module A1 was designed to consist of two types of neuronal units: upward-sweep selective and downward-sweep selective units. The two submodules are organized as 1 × 81 arrays of basic units due to the fact that in auditory cortex sounds are represented on a frequency-based, one-dimensional (tonotopic) axis (Schreiner et al., 2000; Shamma, 2001).

Module A2

The A2 module is designed to be a continuation of A1 and consists of three populations of units: upward sweep selective units, downward sweep selective units and contour selective units; the analogous module in the visual model is V4. The upward sweep selective units and downward sweep selective units have a longer spectrotemporal window of integration than those in A1 so that they are selective for longer frequency sweeps. The contour selective units are selective to changes in sweep direction, which are analogous with the corner selective units in the visual model. The A2 module represents the lateral belt/parabelt areas of primate auditory cortex. In experiments, parabelt neurons are found to be selective to band-pass noise stimuli and FM sounds of a certain rate and direction (Rauschecker, 1997).

Module ST

The third processing module of the auditory model is ST, which stands for superior temporal cortex, including superior temporal gyrus and/or sulcus and the rostral supratemporal plane. Functionally, ST is equivalent to the IT (inferior temporal) module in the visual model, and acts as a feature integrator, containing a distributed representation of the presenting stimulus (Husain et al., 2004; Hackett, 2011). This functional equivalency is supported by experimental

FIGURE 1 | **(A)** The network diagram of the large-scale auditory-visual neural model. Arrows denote excitatory connections; lines ending in circles denote inhibitory connections (excitatory connections to inhibitory interneurons in the receiving module). The anterior insula (aINS) acts as the exogenous attention module where visual-auditory attention competition occurs. See text for details. **(B)** Structure of a Wilson-Cowan microcircuit, which can be considered as a simplified representation of a cortical column. Each microcircuit consists of an excitatory and an inhibitory element with the excitatory element corresponding to the pyramidal neuronal population in a column and the inhibitory element corresponding to the inhibitory interneurons. **(C)** Embedded model in Hagmann's connectome (Hagmann et al., 2008). We first found hypothetical locations for our model's regions of interest (ROIs) and the connected nodes in the connectome (small dots connected to ROIs). We embedded our model of microcircuits and network structure into the structural connectome of Hagmann et al. (2008). See **Table 1** and Ulloa and Horwitz (2016) for details. The yellow nodes correspond to the visual model nodes, the blue to the auditory model nodes, and the red to the anterior insula node. The lines indicate direct connections between modeled nodes and nodes in Hagmann's connectome.

studies that neurons in ST respond to complex features of stimuli (Kikuchi et al., 2010; Leaver and Rauschecker, 2010) and by the findings that a lesion of ST impairs auditory delayed match-to-sample performance (Colombo et al., 1996; Fritz et al., 2005).

Module MTL

The module MTL, a new module that we added to the original auditory model of Husain et al. (2004), represents the medial temporal lobe. It serves as a gate between ST and PFC and is incorporated so as to avoid the short-term memory representation of one stimulus being overwritten by later-arriving stimuli. MTL is analogous to the EC (entorhinal cortex) module in the visual model of Liu et al. (2017). Anatomical studies on monkeys (Munoz et al., 2009) have revealed that medial temporal lobe ablation disconnects the rostral superior temporal gyrus from its downstream targets in thalamus and frontal lobe. In our model, several groups of neurons in MTL are designed to competitively inhibit one another so that only one group of gating neurons will be activated when a stimulus comes in. Once the item is stored in this working memory buffer, an inhibitory feedback from PFC to MTL cortex will suppress the active gating neurons and will release other gating neurons so that the remaining gating neurons are ready for new stimuli. We assume that each group of MTL gating neurons can be used only once during a task trial.

Module PFC

The module PFC represents the prefrontal cortex in both the visual and auditory models. In the visual model, neurons in the PFC module can be delineated into four types based on experimental data acquired during a delayed response task by Funahashi et al. (1990). In our auditory model, the same four types of neuronal populations are employed analogously (Husain et al., 2004). Submodule FS contains cue-sensitive units that in general reflect the activities in the ST (IT) module. D1 and D2 submodules form the short-term memory units that excite one another during the delay period. Recently, we have built multiple sets of D1 and D2 submodules into the visual model (Liu et al., 2017) and successfully implemented tasks that hold more than one item in short-term memory; in the present study we employ the same extension in the auditory model. Submodule R serves as a response module (output). It responds when a displayed stimulus (probe) matches the cue stimulus that is being held in short-term memory. Note that we assume that there are a limited number of gating units and a similarly limited number of D1-D2 units, since empirical studies indicate that only a limited number of items can be simultaneously kept in short-term memory[e.g., the so-called 7 ± 2 (Miller, 1956); others have proposed a more limited capacity such as 3 or 4 (Cowan, 2001); however, see Ma et al., 2014 for a somewhat alternative view]. For computational simplicity, in this paper we will employ no more than three items.

Module aINS

The newly added aINS (anterior insula) module is represented by a pair of mutually inhibited modules. The outputs of the visual and auditory processing streams are taken as inputs for the two modules respectively and are used to generate an exogenous attention signal. The mutual inhibition between the two modules is designed to reflect the competition between modalities in salience computation. The insula area is known for its role in accumulating sensory evidence in perceptual decision-making, bottom-up saliency detection and attentional processes (Seeley et al., 2007; Menon and Uddin, 2010; Ham et al., 2013; Uddin, 2015; Lamichhane et al., 2016). In the current study, aINS processes the visual-auditory bimodality competition that leads to involuntary attention switching.

In both the visual and auditory models, a task specification module is used to provide low-level, diffuse incoming activity that can be interpreted as an attention level to the respective D2 module in the prefrontal area. We located this module arbitrarily in the superior frontal gyrus of the Virtual Brain model. The attention level/task parameter can be modulated by the outputs of the aINS module. When the attention level is low, the working memory modules are not able to hold a stimulus throughout the delay period.

The Talairach coordinates (Talairach and Tournoux, 1988) and the closest node in Hagmann's connectome (Hagmann et al., 2008) for each of the modules discussed above (as well as for the visual model) were identified (see **Table 1**) based on experimental findings. As to the PFC module, which contains four submodules (FS, D1, D2, R), we used the Talairach coordinates of the prefrontal cortex in Haxby et al. (1995) for the D1 submodule in the visual model and assigned the locations of adjacent nodes for the other submodules (FS, D2, R) (see **Table 1**); similarly, for the PFC module in the auditory model, we used the Talairach

coordinates from Husain et al. (2004) for the auditory model D1 units. This arrangement is due to the fact, as mentioned above, that the four types of neuronal populations were based on the experimental findings in monkey PFC during a visual delayed response task (Funahashi et al., 1990) and we assume that auditory working memory possesses the analogous mechanism as visual working memory. It is not known if the four neuronal types were found in separate anatomical locations in PFC or were found in the same brain region, although recent findings from Mendoza-Halliday and Martinez-Trujillo (2017) suggest that visual perceptual and mnemonic coding neurons in PFC are close (within millimeters) to one another.

See the **Appendix** for the mathematical aspects of the network model. All the computer code for the combined auditory-visual model can be found at https://github.com/NIDCD, as can information about how our modeling software is integrated with software for the Virtual Brain.

Simulated Experiments

We used the combined auditory-visual model to perform several simulated experiments that included not only one stimulus, but other stimuli as well, some of which could be considered to be distractors. We created 10 "subjects" by varying interregional structural connection weights slightly through a Gaussian perturbation. For each experiment, 20 trials are implemented for each "subject." The complete set of simulated experiments is the following:

Simulated Auditory Short-Term Memory Experiments:

a. Auditory delayed match-to-sample task. This experiment implemented the original delayed match-to-sample (DMS) task to demonstrate that the extended auditory model [with an added module—the MTL (medial temporal lobe), and the linkage between visual and auditory models *via* the aINS (anterior insula) module] continues to perform the DMS task and gives essentially the same results as the original model (Husain et al., 2004). One typical DMS trial consists of the presentation of a stimulus, an ensuing delay period, presentation of a probe (the same or a new stimulus); the simulated subjects need to decide whether the probe is the same as the first stimulus presented (see **Figure 2B**). The attention/task parameter is set to high (0.3) during a trial.
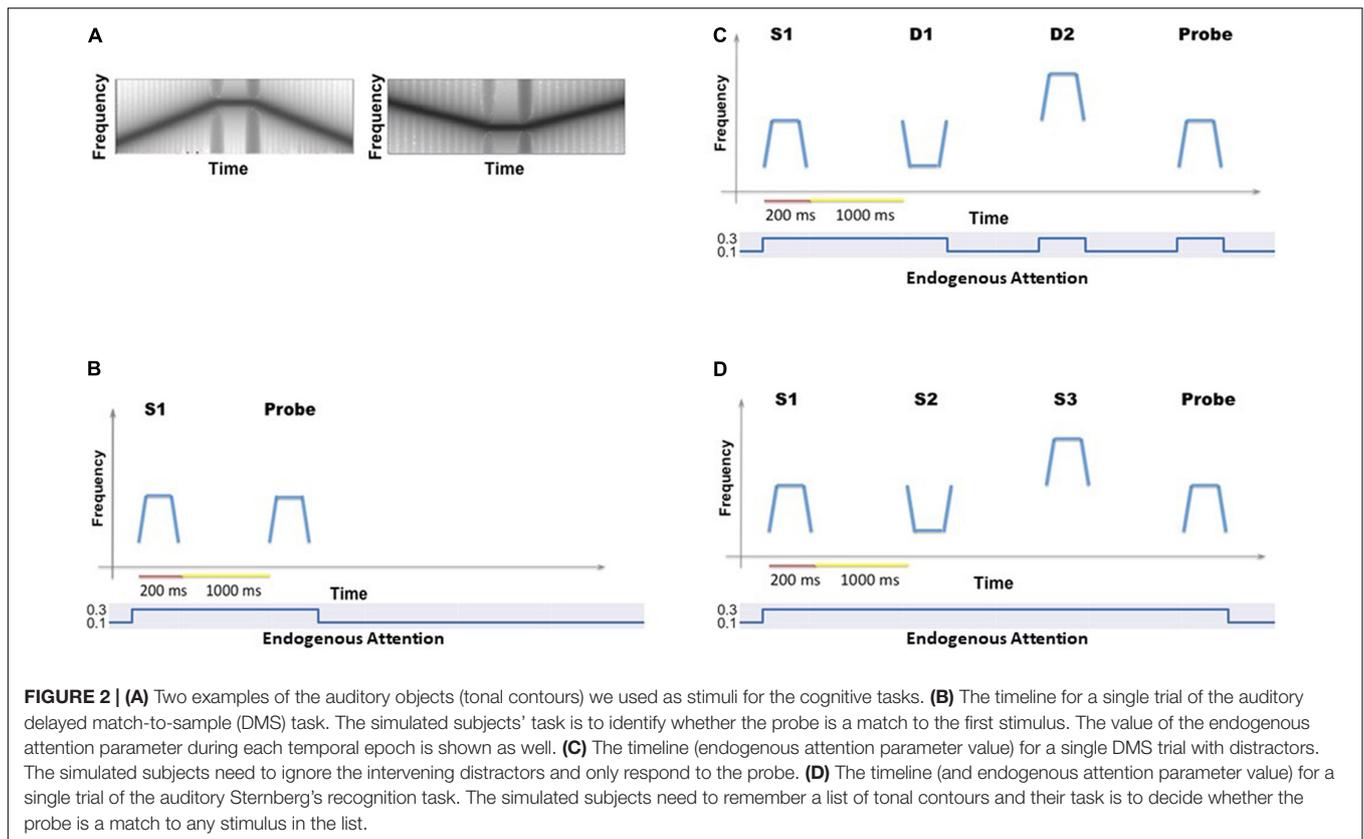
b. Auditory delayed match-to-sample task with distractors. These auditory short-term memory simulations are employed to demonstrate that the extended auditory sub-model within the combined auditory-visual model performs the analogous tasks as the visual model of Liu et al. (2017). The simulated subjects are presented with two distractors (visual or auditory) before the probe stimulus is presented (see **Figure 2C**). The attention/task parameter is set to high (0.3) at stimulus onset and decreases to low (0.05) following the presentation of the distractors.

c. Auditory version of Sternberg's recognition task. An auditory variant of Sternberg's recognition task (Sternberg, 1966, 1969) is used. On each trial of the simulation, three auditory stimuli are presented sequentially, followed by a delay period and then a probe. The subjects' task is to decide whether the probe is a match to any of the three stimuli presented earlier (see **Figure 2D**). The Sternberg paradigm with visual/auditory objects

TABLE 1 | Locations of nodes in TVB chosen to host our modules.

Modules	Source	Host connectome node Talairach location
A1	Husain et al., 2004	(51 −24, 8)
A2	Husain et al., 2004	(61, −36, 12)
ST	Husain et al., 2004	(59, −20, 1)
MTL	Tanabe et al., 2005	(22, −30, −12)
V1/V2	Haxby et al., 1995	(14, −86, 7)
V4	Haxby et al., 1995	(30, −70, −7)
IT	Haxby et al., 1995	(31, −39, −6)
EC	Hagmann et al., 2008	(25, −12, −25)
FS	Location selected for illustrative purposes	(35, 19, 13), (47, 19, 9)
D1	Haxby et al., 1995; Husain et al., 2004	(51, 12, 10), (43, 29, 21)
D2	Location selected for illustrative purposes	(32, 29, 8), (42, 39, 2)
aINS	Gu et al., 2013	(48,12,4)
R	Location selected for illustrative purposes	(33, 13, 28), (29, 25, 40)

We embedded our model into The Virtual Brain connectome based on anatomical sources listed in the table.

FIGURE 2 | (A) Two examples of the auditory objects (tonal contours) we used as stimuli for the cognitive tasks. **(B)** The timeline for a single trial of the auditory delayed match-to-sample (DMS) task. The simulated subjects' task is to identify whether the probe is a match to the first stimulus. The value of the endogenous attention parameter during each temporal epoch is shown as well. **(C)** The timeline (endogenous attention parameter value) for a single DMS trial with distractors. The simulated subjects need to ignore the intervening distractors and only respond to the probe. **(D)** The timeline (and endogenous attention parameter value) for a single trial of the auditory Sternberg's recognition task. The simulated subjects need to remember a list of tonal contours and their task is to decide whether the probe is a match to any stimulus in the list.

has been used in many studies, and thus allows us to compare our simulated results with experimental results.

Simulated Visual-Auditory Bimodality Experiments

a. Bimodality DMS task with various exogenous attention settings (see **Figure 3A**): A block of visual DMS trials and a block of auditory DMS trials are implemented simultaneously. The saliency of visual stimuli and auditory stimuli varies from trial to trial. The attention/task parameter assigned to each modality is determined based on the real-time output of the anterior insula module (aINS). In general, higher saliency of one stimulus will result in higher attention to the corresponding modality. Essentially, this experiment is asking a subject to perform a DMS task on the sensory modality that is more salient.

b. Bimodality distraction task with different working memory loads (see **Figure 3B**). The simulated subjects are asked to remember 1~3 visual stimuli before an auditory distractor occurs. The endogenous attention is set to attend to visual stimuli.

Simulated fMRI Experiments

Simulated fMRI signals can be calculated for each task discussed above. The direct outputs are the electrical activity of simulated neuronal units. Prior to generating fMRI BOLD time series, we first calculate the integrated synaptic activity by spatially integrating over each module and temporally over 50 ms (Horwitz and Tagamets, 1999). Using the integrated synaptic activity of select regions of interests (ROIs) as the input to the fM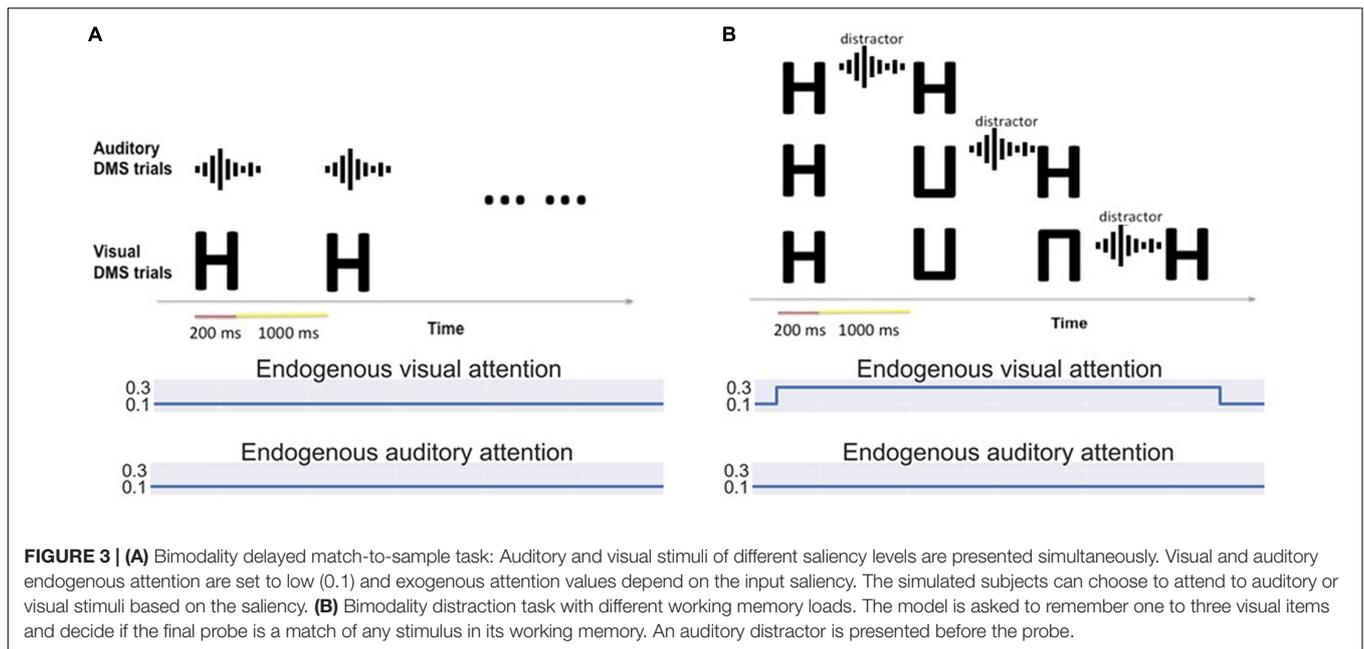RI BOLD balloon model of hemodynamic response (Stephan et al., 2007; Ulloa and Horwitz, 2016), we calculate the simulated fMRI signal time-series for all our ROIs and then down-sample the time-series to correspond to a TR value of 1 s (for more mathematical and other technical details, see the **Appendix**; Ulloa and Horwitz, 2016).

In simulating an fMRI experiment for the aforementioned cognitive tasks, we implemented two types of design schemes: block design and event-related design. In an experiment with block design, one stimulus is followed by a 1-s delay period, and the model alternately performs a block of task trials (3 trials) and a block of control trials (3 trials). The control trials use passive perception of degraded shapes and random noises. With an event-related design, the delay period following each stimulus is extended to 20 s in order to show a more complete response curve in the BOLD signal.

RESULTS

Auditory Short-Term Memory Experiments

Our first simulated experiments test whether the combined auditory-visual model can produce appropriate results (both neurally and behaviorally) for simulated auditory stimuli. Our results demonstrated that the model successfully performed the auditory DMS task (with and without distractors) and Sternberg's recognition task with similar accuracy as the visual

FIGURE 3 | (A) Bimodality delayed match-to-sample task: Auditory and visual stimuli of different saliency levels are presented simultaneously. Visual and auditory endogenous attention are set to low (0.1) and exogenous attention values depend on the input saliency. The simulated subjects can choose to attend to auditory or visual stimuli based on the saliency. **(B)** Bimodality distraction task with different working memory loads. The model is asked to remember one to three visual items and decide if the final probe is a match of any stimulus in its working memory. An auditory distractor is presented before the probe.

tasks (Liu et al., 2017; **Table 2**). **Figure 4A** shows the electrical activities of simulated neuronal units of the different modules during a DMS task. The input stimuli, represented by medial geniculate nucleus activity, are first processed by feature-selective modules in A1 and A2. A2 neurons have longer spectrotemporal windows of integration than A1 neurons and thus A2 is responsive to longer frequency sweeps. The ST module contains the distributed representation of the presenting stimulus and feeds the presentation forward to the gating module MTL and then to PFC. A working memory representation is held in the D1 and D2 modules through the delay period. The probe stimulus for **Figure 4A** is a match with the presented stimulus so that the R module responds.

The auditory model also can handle the DMS task with distractors, for which the electrical activities are illustrated in **Figure 4B**. The first stimulus is the target that the model needs to remember, and it is followed by two distractors. The endogenous attention/task-specification unit is set to only remember the first item. The distractors also evoke some activity in the working memory modules (D1, D2), but that activity is not strong enough to overwrite the representation of the first stimulus, so the model successfully holds its response until the matched probe appears.

TABLE 2 | Model performance of the auditory tasks.

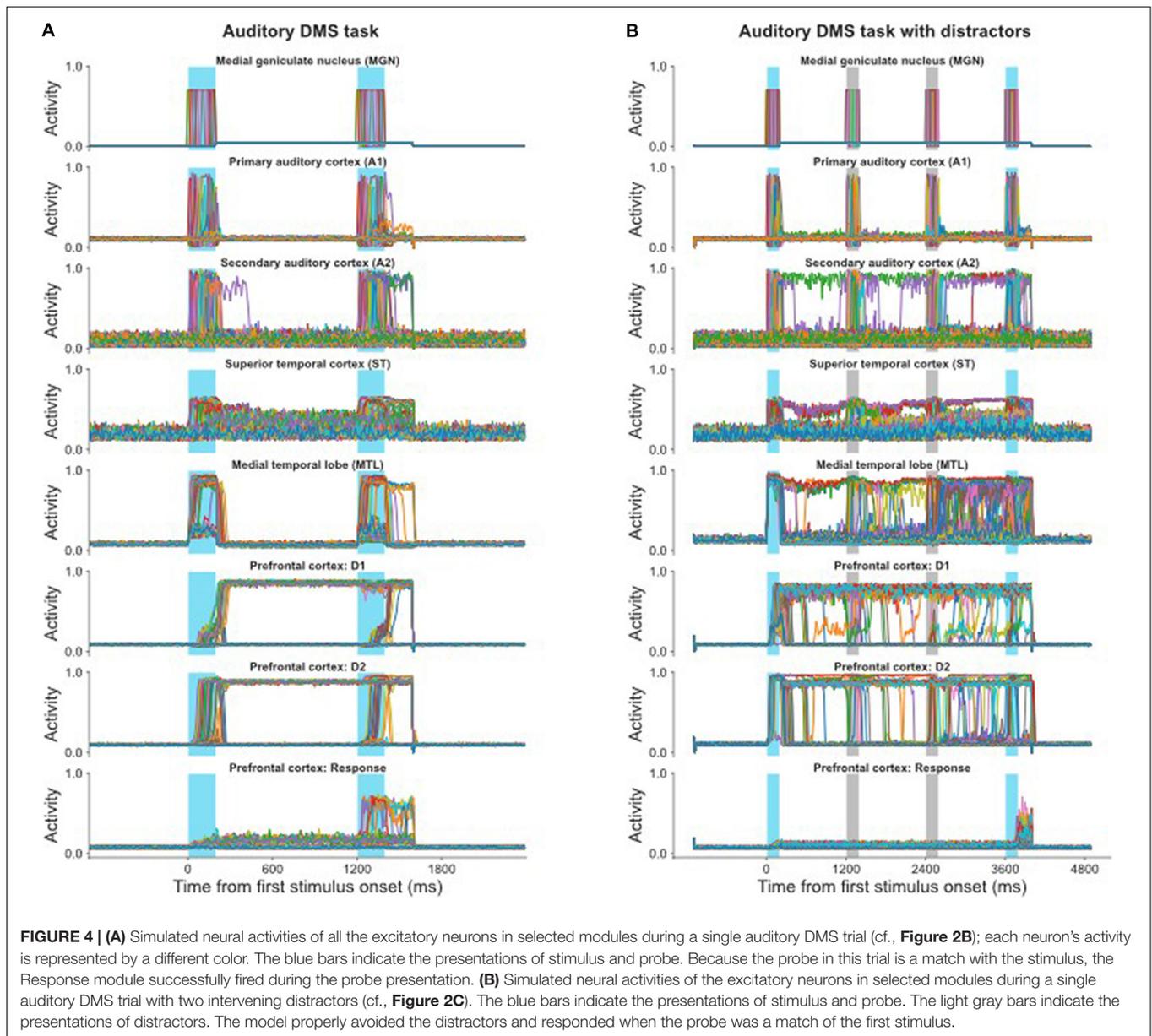| Tasks	DMS (%)	DMS w/distractors (%)	Sternberg's task (%)
Accuracy	83.9	81.8	78.7
Standard deviation	5.71	6.73	6.05

The model can successfully perform DMS task, DMS with distractors and Sternberg's task with high accuracy. Values represent means and standard deviations over all simulated subjects and trials.

In the visual model, we reported that we observed enhanced activity in the IT module during the delay period which helped short-term memory retention (Liu et al., 2017) and which was consistent with experimental findings (Fuster et al., 1982). In the current study, our modeling also displayed this type of neuronal activity in ST, as can be seen in **Figure 4**, and this enhanced activity has been reported in auditory experiments (Colombo et al., 1996; Scott et al., 2014).

Figure 5 demonstrates how the model implements the auditory version of Sternberg's recognition task and handles multiple auditory objects in short-term memory. The first three items are held in short-term memory (D1, D2), which is shown in **Figure 5B**, and when the probe matches any of the remembered three items the R module is activated (**Figure 5A**). Different groups of neurons in the gating module MTL responded to each of the stimuli and prevented the representations in working memory from overwriting one another.

Visual-Auditory Bimodality Competition Experiments

Figure 6 shows the simulated intersensory attention switch caused by input saliency changes. During the simulated experiment, the attentional inputs into working memory module D2 were both the endogenous attention and the exogenous attention (output of module aINS). In the experiment shown in **Figure 6**, five DMS trials were implemented. The model's endogenous attention was set to attend to auditory stimuli (the auditory attention/task parameter is set to "high") and not to attend to visual stimuli. Thus, the model attended to auditory stimuli and treated visual stimuli as distractors in the first three DMS trials during which the saliency of auditory stimuli was higher than that of the visual stimuli. When the saliency of visual stimuli was enhanced above a certain threshold (0.8 in our
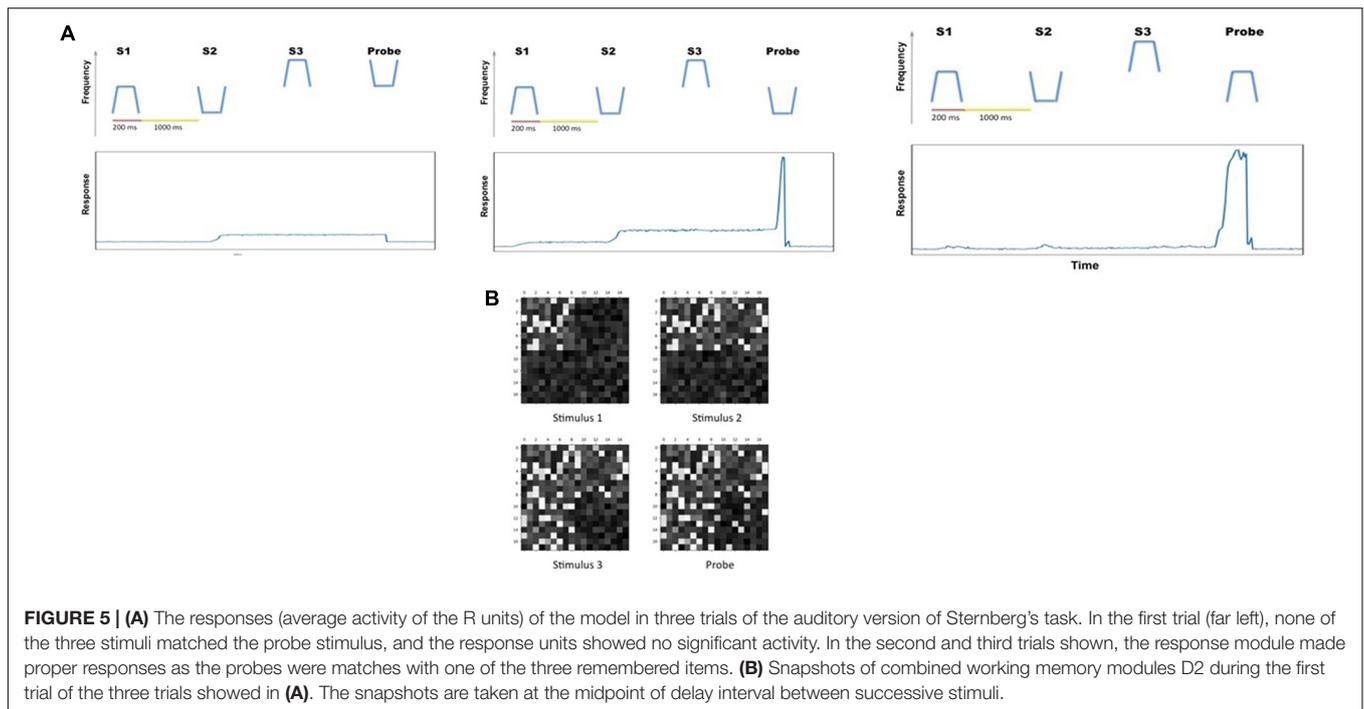
FIGURE 4 | (A) Simulated neural activities of all the excitatory neurons in selected modules during a single auditory DMS trial (cf., **Figure 2B**); each neuron's activity is represented by a different color. The blue bars indicate the presentations of stimulus and probe. Because the probe in this trial is a match with the stimulus, the Response module successfully fired during the probe presentation. **(B)** Simulated neural activities of the excitatory neurons in selected modules during a single auditory DMS trial with two intervening distractors (cf., **Figure 2C**). The blue bars indicate the presentations of stimulus and probe. The light gray bars indicate the presentations of distractors. The model properly avoided the distractors and responded when the probe was a match of the first stimulus.

modeling setting), the model started to attend to the visual stimuli and attended less to the auditory stimuli, but the model still could encode auditory stimuli into working memory.

Working Memory Load Reduces Intersensory Distractions

When implementing a visual DMS task with more than one item held in working memory during the delay period, the model showed a smaller response to auditory distractions (**Table 3** and **Figure 7**). As observed in **Figure 7A**, exogenous auditory attention in response to auditory distractors decreases when the visual working memory has more than one item. This was demonstrated by calculating the average auditory attention values during the presentations of auditory distractors with one, two,

and three items in visual working memory across 20 runs of each subject. The distributions of average auditory attention changes of 10 subjects due to more than one item in visual working memory are shown in the boxplots of **Figure 7A** and **Table 3**. The significance of auditory attention reductions with 2 and 3 items vs. with 1 item in visual working memory was tested against zero (2 items: degree of freedom = 9, $p = 0.0003$, $t = 4.154$; 3 items: degree of freedom = 9, $p = 0.0008$, $t = 3.727$). **Figure 7B** shows ST neuronal activity for auditory distractors is decreased as the visual working memory load is increased. Similar to auditory attention, ST activity during the presentations of auditory distractors with one, two, and three items in visual working memory was averaged across 20 runs of each subject. The distributions of average ST activity changes of 10 subjects due to more than one item in visual working memory are shown

FIGURE 5 | (A) The responses (average activity of the R units) of the model in three trials of the auditory version of Sternberg's task. In the first trial (far left), none of the three stimuli matched the probe stimulus, and the response units showed no significant activity. In the second and third trials shown, the response module made proper responses as the probes were matches with one of the three remembered items. **(B)** Snapshots of combined working memory modules D2 during the first trial of the three trials showed in **(A)**. The snapshots are taken at the midpoint of delay interval between successive stimuli.

in the boxplots of **Figure 7A** and **Table 3**. The significance of ST neuronal activity reductions with 2 and 3 items in visual working memory was tested with one-tail t-tests against zero (2 items: simulated subjects = 10, p = 0.0005, t = 3.894; 3 items: simulated subjects = 10, p = 0.0003, t = 4.160). This phenomenon shows that the working memory formed in the model is stable and is also consistent with experimental findings that higher working memory load in one modality reduces distraction from another modality (SanMiguel et al., 2008).

However, little is known about the underlying neural mechanism mediating these phenomena. Based on the structure of our model, we propose two possible pathways (**Figure 7D**) that can connect the increase in working memory load in one modality to the increase of exogenous attention in the corresponding modality, thus reducing distractions (i.e., exogenous attention) from the other competing modality: (1) D2–IT/ST–Auditory-attention. When the working memory load is high, working memory modules D1 and D2 will maintain a high activity level. Due to feedback connections from D2 to IT/ST, IT/ST and the downstream module (Auditory-attention) will exhibit increased level of activity, i.e., the exogenous attention in the corresponding modality will increase and the exogenous attention in the competing modality will decrease. (2) D2—V4/A2—IT/ST—Auditory-attention. Similar to pathway 1, the feedback connection from D2 to V4/A2, although not as strong as the connection from D2 to IT/ST, may also contribute to this working memory load effect. When the feedback connections from PFC to V4 and IT/ST in the model are removed, we no longer observed this working memory load effect. In summary, working memory load can affect intersensory neural responses through top-down feedbacks that change intersensory attention competition.
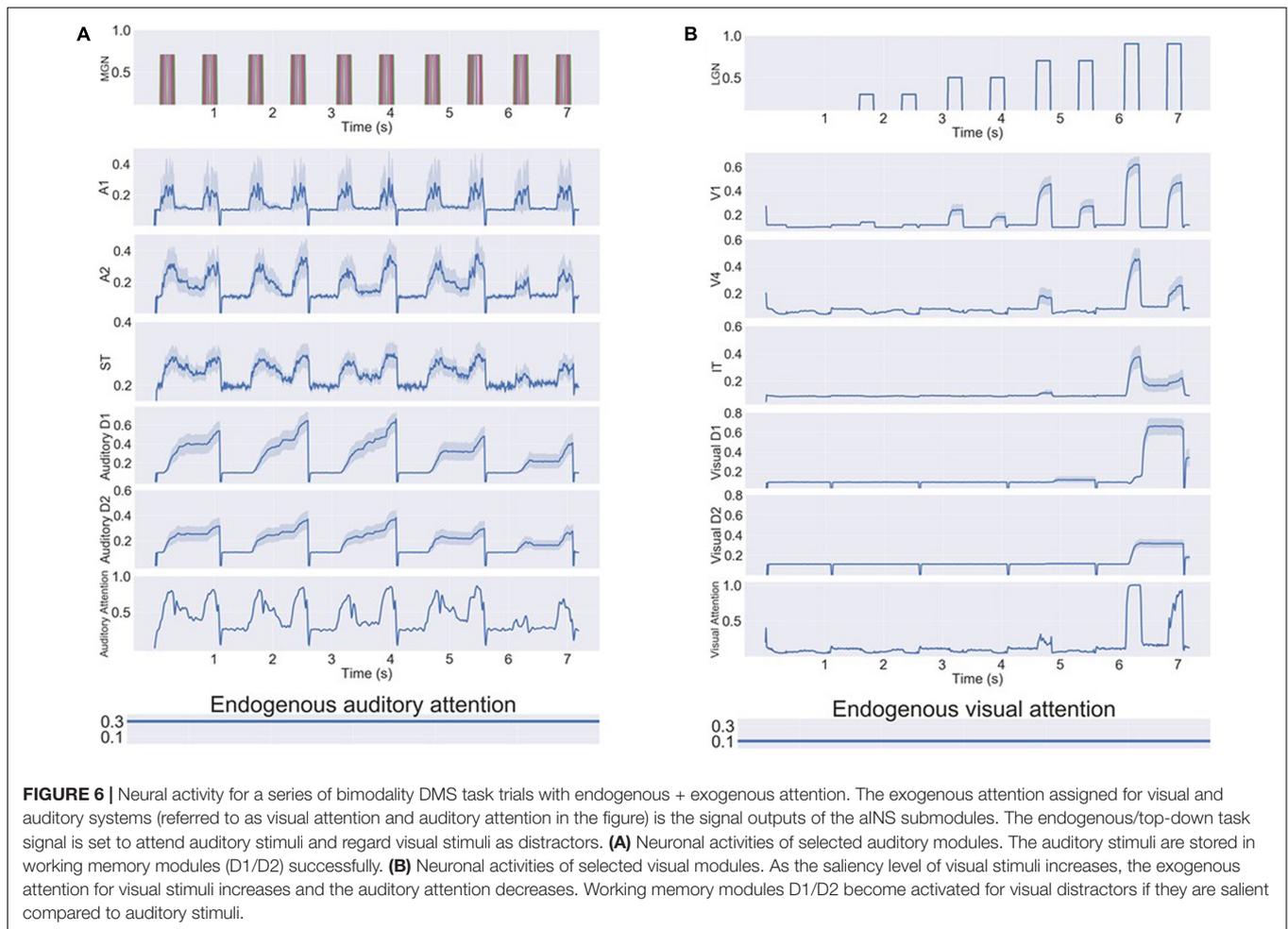
It is worth noting that the SanMiguel et al. (2008) findings hold only for intersensory tasks (i.e., tasks in which the sensory information streams are not linked). For cases where the two sensory streams are integrated (i.e., crossmodal tasks), it has been shown that increased working memory load leads to increased distraction between modalities (Michail et al., 2021).

Simulated fMRI Experiments

The results of our simulations can be tested in humans using functional neuroimaging methods. We will illustrate this using fMRI. As discussed in section "Materials and Methods," fMRI BOLD time series are generated for select regions of interests (ROIs) using integrated synaptic activity, and for each task we implemented the experiment using either a block design or an event-related design. The event-related scheme has longer duration delay periods than experiments using a block design. Thus, the event-related experiments can show a more complete BOLD response curve for each incoming stimulus.

Figure 8 shows the simulated BOLD signal for a block-design auditory DMS task, which successfully replicates the results from Husain et al. (2004). In the simulated experiment, one block of DMS trials is followed by a block of control trials in which random noise patterns are used. Modules representing early auditory areas A1 and A2 responded equally to DMS stimuli and noises. Higher order modules such as MTL and PFC, on the other hand, show much larger signal changes, as was shown empirically by Husain et al. (2004) who employed tonal contours and auditory noise patterns.

Figure 9 shows one event-related fMRI BOLD time-series consisting of three visual DMS trials. The probes in all three trials matched the first stimulus. During the second and the third trials, simulated auditory distractors were presented during the delay

FIGURE 6 | Neural activity for a series of bimodality DMS task trials with endogenous + exogenous attention. The exogenous attention assigned for visual and auditory systems (referred to as visual attention and auditory attention in the figure) is the signal outputs of the aINS submodules. The endogenous/top-down task signal is set to attend auditory stimuli and regard visual stimuli as distractors. **(A)** Neuronal activities of selected auditory modules. The auditory stimuli are stored in working memory modules (D1/D2) successfully. **(B)** Neuronal activities of selected visual modules. As the saliency level of visual stimuli increases, the exogenous attention for visual stimuli increases and the auditory attention decreases. Working memory modules D1/D2 become activated for visual distractors if they are salient compared to auditory stimuli.

periods. Early auditory area A1 responded to auditory distractors but did not cause large signal changes in auditory PFC regions compared with visual PFC regions. The model finished all three trials correctly, but the presence of auditory stimuli lowered the BOLD activity in visual PFC modules.

One experiment of visual-auditory intersensory attention allocation was also implemented. The BOLD signals of ROIs are displayed in **Figure 10**. No task instructions were given prior to the simulation, i.e., the endogenous attention was maintained at a low value. The model reacted to visual and auditory stimuli purely based on exogenous attention capture. The model first attended to visual stimuli as the BOLD signal for visual PFC spiked (see **Figure 10A**) and then switched to attend to salient auditory stimuli as the BOLD signal for auditory PFC module increased (see **Figure 10B**). The aINS module controls the switch by playing the role of exogenous attention.

DISCUSSION

In this paper we presented a large-scale biologically constrained neural network model that combines visual and auditory object processing streams. To construct this combined network model, we extended a previously developed auditory LSNM so it could handle distractors and added a module (aINS) that connected this extended auditory model with a previously established visual LSNM. The newly combined auditory-visual (AV) model can perform with high performance accuracy a variety of short-term memory tasks involving auditory and visual inputs that can include auditory or visual distractor stimuli. Multiple items can be retained in memory during a delay period. Our model, embedded in a whole brain connectome framework, generated

TABLE 3 | Memory load reduces intersensory distractions.

Load in visual working memory		2 items vs. 1 item (%)	3 items vs. 1 item (%)
Auditory attention changes	Mean	−6.88	−9.48
	SD	4.78	7.96
ST activity changes	Mean	−2.16	−2.46
	SD	1.54	1.84

Visual working memory tasks with auditory distractors. With the increase of visual working memory load, the auditory attention and the activity of ST neurons for auditory distractors is reduced. Values represent means and standard deviations over all simulated subjects and trials.
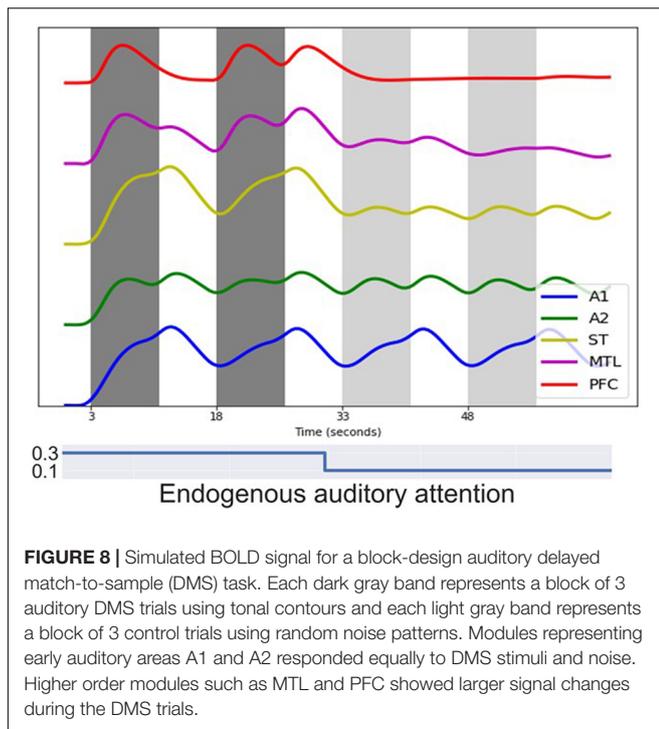
FIGURE 7 | Working memory load reduces intersensory distractions. Three visual DMS trials are implemented with 1, 2, and 3 input visual stimuli, respectively. One auditory distractor is played before the final visual probe is presented. **(A)** Exogenous auditory attention in response to auditory distractors decreases when the visual working memory has more than one item. The left panel shows the average auditory attention of one simulated subject. The right panel shows the distributions of the average attention reductions of 10 simulated subjects with more visual items in working memory. One-tail t-tests were used to determine the significance of attention reductions against zero (trial 2: simulated subjects = 10, $p = 0.0003$, $t = 4.154$; trial 3: simulated subjects = 10, $p = 0.0008$, $t = 3.727$). **(B)** ST neuronal activity for auditory distractors also decreases as the visual working memory increases. The left panel shows the average ST neuronal activity of one simulated subject. The right panel shows the distributions of the average ST activity reductions of 10 simulated subjects. One-tail t-tests were used to determine the significance of activity reductions against zero (trial 2: simulated subjects = 10, $p = 0.0005$, $t = 3.894$; trial 3: simulated subjects = 10, $p = 0.0003$, $t = 4.160$). **(C)** Exogenous visual attention responses for visual input do not change. **(D)** Neural circuits in our model that explain the working memory load effect. Working memory loads increases exogenous visual attention *via* feedback connections to V4 and IT that in turn inhibit exogenous auditory attention to auditory distractors.

simulated dynamic neural activity and fMRI BOLD signals in multiple brain regions.

This model was used to investigate the interaction between exogenous and endogenous attention on short-term memory performance. We simulated intersensory exogenous attention capture by presenting salient auditory distractors in a visual DMS task or salient visual distractors in an auditory task. Moreover, we simulated involuntary attention switching by presenting visual and auditory stimuli simultaneously with different saliency levels. We also investigated how working memory load in one modality could reduce exogenous attention capture by the other modality.

The AV network model used in this study was obtained by combining previously constructed visual (Tagamets and Horwitz, 1998; Horwitz and Tagamets, 1999; Ulloa and Horwitz, 2016; Liu et al., 2017) and auditory processing models (Husain et al., 2004). In our visual model, we assigned the entorhinal cortex to be the gating module between the inferior temporal area and PFC based on a series of experimental results (see Liu et al., 2017 for details). However, experimental evidence for the corresponding brain region that implements the auditory gating function is less conclusive. We based our MTL choice for the auditory gating module on a study of Munoz et al. (2009) that showed that ablation of MTL can result in disconnections between the rostral superior temporal gyrus and its downstream targets in thalamus

and frontal lobe. Several cognitive tasks involving short-term memory were successfully implemented with simulated auditory stimuli. These short-term memory tasks can include auditory or visual distractor stimuli or can require that multiple items be retained in mind during a delay period. Simulated neural and fMRI activity replicated the Husain et al. results (Husain et al., 2004) when no distractors were present. The simulated neural and fMRI activity in Husain et al. (2004) model themselves were consistent with empirical data.

Neural network modeling now occupies a prominent place in neuroscience research (see Bassett et al., 2018; Yang and Wang, 2020; Pulvermuller et al., 2021 for reviews). Among these network modeling efforts are many focusing on various aspects of working memory (e.g., Tartaglia et al., 2015; Ulloa and Horwitz, 2016; Masse et al., 2019; Orhan and Ma, 2019; Mejias and Wang, 2022). Some of these employ relatively small networks (Tartaglia et al., 2015; Masse et al., 2019; Orhan and Ma, 2019), whereas others use anatomically constrained large-scale networks (Ulloa and Horwitz, 2016; Mejias and Wang, 2022). Most of these models have targeted visual processing and have attempted to infer the neural mechanisms supporting the working memory tasks under study. The AV network model used in the present study was of the anatomically constrained large-scale type.

FIGURE 8 | Simulated BOLD signal for a block-design auditory delayed match-to-sample (DMS) task. Each dark gray band represents a block of 3 auditory DMS trials using tonal contours and each light gray band represents a block of 3 control trials using random noise patterns. Modules representing early auditory areas A1 and A2 responded equally to DMS stimuli and noise. Higher order modules such as MTL and PFC showed larger signal changes during the DMS trials.

Compared to visual processing tasks, there have been far fewer neural network modeling efforts directed at neocortical auditory processing of complex stimuli. Ulloa et al. (2008) extended the Husain et al. (2004) model to handle long-duration auditory objects (duration ~ 1s to a few seconds; the original model dealt with object duration of less than ~300 ms). Kumar et al. (2007) used Dynamic Causal Modeling (Friston et al., 2003) and Bayesian model selection on fMRI data to determine the effective connectivities among a network of brain regions mediating the perception of sound spectral envelope (an important feature of auditory objects). More recently, a deep learning network analysis was employed by Kell et al. (2018) to optimize a multi-layer, hierarchical network for speech and music recognition. Following early shared processing, their best performing network showed separate pathways for speech and music. Furthermore, their model performed the tasks as well as humans, and predicted fMRI voxel responses.

As noted in section "Materials and Methods," the auditory objects that are the input to the auditory network consist of spectrotemporal patterns. However, many auditory empirical studies utilize auditory objects that also contain sound source information. For example, among the stimuli used by Leaver and Rauschecker (2010) were human speech and musical instrument sounds. Lowe et al. (2020) employed both MEG and fMRI to study categorization of human, animal and object sounds. Depending on the exact experimental design, sound source information may require long-term memory. Our current modeling framework does not implement long-term memory. Indeed, the interaction between long-term and working memory is an active area of current research (Ranganath and Blumenfeld,

2005; Eriksson et al., 2015; Beukers et al., 2021), and our future research aims to address this issue.

Our modeling of the auditory Sternberg task used a neurofunctional architecture analogous to the one used in our visual model. This is consistent with the behavioral findings of Visscher et al. (2007) who tested explicitly the similarities between visual and auditory versions of the Sternberg task.

In the simulated experiments presented in this paper, we used salience as a way to modulate exogenous attention. The salience level of one stimulus is typically detected based on the contrast between the stimulus and its surrounding environment. However, the "contrast" can be defined on different metrics, for example, the luminance of visual objects and the loudness of auditory objects, which were used in our modeling. There are other metrics based on sensory features to define a salient object, such as bright colors, fast moving stimuli in a static background, etc. An object can also be conceptually salient. A theory has been proposed that schizophrenia may arise out of the aberrant assignment of salience to external or internal objects (Kapur, 2003). In our simulation, stimuli that resulted in high working memory load may also be considered as conceptually salient, as the effect of high working memory load is similar to high endogenous attention (Posner and Cohen, 1984).

A number of brain regions are thought to be involved in the multisensory integration process (Quak et al., 2015). The perirhinal cortex has been proposed based on monkey anatomical studies (Suzuki and Amaral, 1994; Murray and Richmond, 2001; Simmons and Barsalou, 2003), whereas the left posterior superior temporal sulcus/middle temporal gyrus is suggested to be where multisensory integration takes place based on some human functional imaging findings (Calvert, 2001; Amedi et al., 2005; Beauchamp, 2005; Gilbert et al., 2013). In our study, we focused mainly on intersensory attention competition based on bottom-up salience, as opposed to multisensory integration. The anterior insula and the anterior cingulate are major components in a network that integrates sensory information from different brain regions for salience computation (Seeley et al., 2007; Sridharan et al., 2008; Menon and Uddin, 2010; Uddin, 2015; Alexander and Brown, 2019). Recent work (Ham et al., 2013; Lamichhane et al., 2016) argues that the anterior insula accumulates sensory evidence and drives anterior cingulate and salience network to make proper responses (for a review, see Uddin, 2015). Therefore, we assigned the anterior insula (aINS) as the module responsible for the exogenous attention computation and visual-auditory competition. In our model, aINS receives its inputs from IT in the visual network and ST in the auditory network and assigns values to the visual and auditory attention/task-specific unit. This arrangement has some similarities with the conceptual model of insula functioning proposed by Menon and Uddin (2010).

The model presented in this paper represents a first step in developing a neurobiological model of multisensory processing. In the present case, stimuli from the two sensory modalities are not linked. Moreover, the visual and auditory stimuli we simulated are not located in different parts of space, and thus spatial attention is not required. Nor do these visual and auditory objects correspond to well-known objects and thus do not require a long-term semantic representation. Future work will
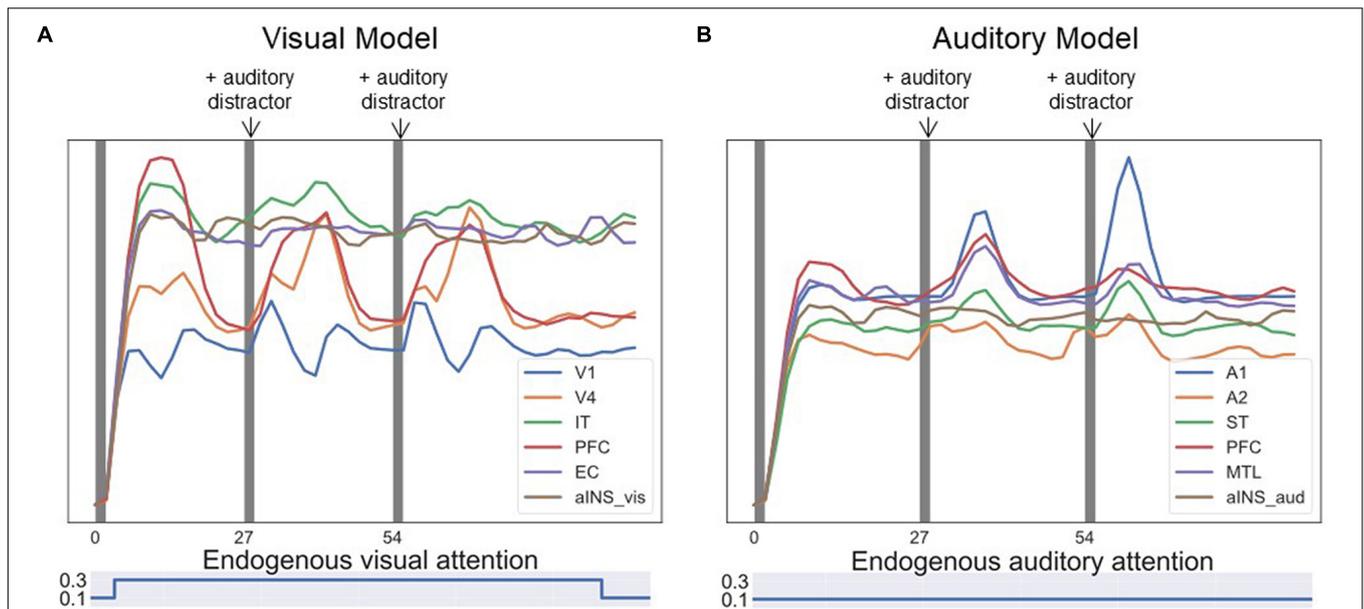
FIGURE 9 | Simulated BOLD signal for an event-related visual DMS task with auditory distractors. Each of the three dark bands represent one visual DMS trial. Each trial is 2 s and the interval between trials is 25 s. During the second and the third trials, auditory distractors were played during the delay periods. **(A)** shows the BOLD signal for ROIs in the visual model and **(B)** for ROIs in the auditory model. Early auditory area A1 responded to auditory distractors but did not cause much signal changes in auditory PFC regions compared with visual PFC regions. However, the presence of auditory stimuli lowered the BOLD activity in visual PFC modules.
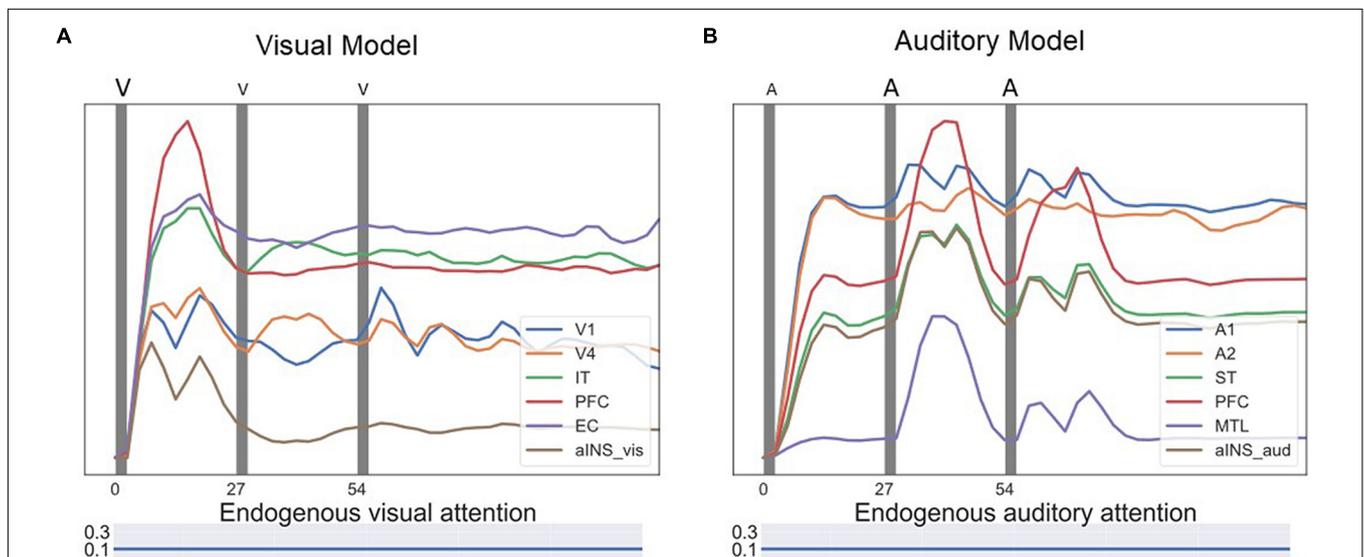
FIGURE 10 | Simulated BOLD signals for an event-related visual-auditory intersensory attention allocation experiment. In the experiment, three visual DMS trials and three auditory DMS trials were presented in parallel, represented by the dark gray bands. No task instructions were given prior to the simulation, i.e., the model reacted to visual and auditory stimuli purely based on exogenous attention capture. In the first trial, the visual stimuli were more salient while in later trials auditory stimuli were more salient. The saliency levels of visual and auditory stimuli are illustrated with different font sizes above the figure. The bigger font size indicates higher saliency level. **(A)** Shows BOLD signal for ROIs in the visual model. The model attended to visual stimuli in the beginning of the test and the BOLD signal for visual PFC spiked. **(B)** Shows BOLD signal for ROIs in the auditory model. The model, after attending to visual stimuli in the beginning, switched to attending to salient auditory stimuli and the BOLD signal for the auditory PFC module increased. Endogenous attention was maintained at a low value throughout the experiment.

entail extending the model to incorporate long-term memory representations so that cognitive tasks such as the paired-associates task can be implemented (e.g., Smith et al., 2010; Pillai et al., 2013).

In this paper, we used our LSNM to simulate fMRI data to illustrate how our simulation model's predictions could be tested using human neuroimaging data. Our modeling framework also can simulate EEG/MEG data. Banerjee et al. (2012a) used

the visual model to simulate MEG data for a DMS task. These simulated data were employed to test the validity of a data analysis method called temporal microstructure of cortical networks (Banerjee et al., 2008, 2012b). An important future direction for our modeling effort is to enhance our framework so that high temporal resolution data such as EEG/MEG can be explored, especially in terms of frequency analysis of neural oscillations. For example, multisensory processing has been extensively investigated using such data (e.g., Keil and Senkowski, 2018).

Some caveats of our work include the following: we hypothesized that the medial temporal lobe was responsible for a gating mechanism in auditory processing, and thus the detailed location and the gating mechanism need to be confirmed by experiments. Also, the locations we chose for prefrontal nodes (D1, D2, FS, R) in the Virtual Brain are somewhat arbitrary.

The structural connectome due to Hagmann et al. (2008) that we employed here was utilized primarily so that we could compare the extended auditory model with the extended visual model of Liu et al. (2017). Its primary role was to inject neural noise into our task-based auditory-visual networks (see Ulloa and Horwitz, 2016). The Virtual Brain package allows use of other structural models, including those with higher imaging resolution (e.g., the Human Connectome Project connectome; Van Essen et al., 2013), These can be employed in future studies.

In this study, there are no explicit transmission time-delays between our model's nodes. Such delays have been shown to play an important role in modeling resting state fMRI activity (Deco et al., 2009), and in models investigating oscillatory neural network behavior (Petkoski and Jirsa, 2019). Our models show implicit delays between nodes because of the relative slow increase of the sigmoidal activation functions. Our DMS tasks also did not require explicit and detailed temporal attention unlike other studies (e.g., Zalta et al., 2020). Transmission time-delays may be needed in future studies depending on the task design.

In summary, we have performed several auditory short-term memory tasks using an auditory large-scale neural network model, and we also simulated auditory-visual bimodality competition and intersensory attention switching by combining the auditory model with a parallel visual model. We modeled short-term auditory memory with local microcircuits (D1, D2) and a large-scale recurrent network (PFC, ST) that produced neural behaviors that matched experimental findings. For generating a brain-like environment, we embedded the model into The Virtual Brain framework. In the future the model can be extended to incorporate more brain regions and functions, such as long-term memory. Our results indicate that computational modeling can be a powerful tool for interpreting and integrating nonhuman primate electrophysiological and human neuroimaging data.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: https://github.com/NIDCD.

AUTHOR CONTRIBUTIONS

QL and BH conceived, designed the study, and wrote the manuscript. QL provided software and performed data analysis. AU provided software and data analysis support. BH supervised the study, reviewed, and edited the manuscript. All authors approved the final manuscript.

FUNDING

ACKNOWLEDGMENTS

REFERENCES

Alexander, W. H., and Brown, J. W. (2019). The role of the anterior cingulate cortex in prediction error and signaling surprise. *Top. Cogn. Sci.* 11, 119–135. doi: 10.1111/tops.12307

Amedi, A., von Kriegstein, K., van Atteveldt, N. M., Beauchamp, M. S., and Naumer, M. J. (2005). Functional imaging of human crossmodal identification and object recognition. *Exp. Brain Res.* 166, 559–571. doi: 10.1007/s00221-005-2396-5

Baddeley, A. (1986). Modularity, mass-action and memory. *Q. J. Exp. Psychol. A* 38, 527–533. doi: 10.1080/14640748608401613

Baddeley, A. (1996). The fractionation of working memory. *Proc. Natl. Acad. Sci. U.S.A.* 93, 13468–13472. doi: 10.1073/pnas.93.24.13468

Banerjee, A., Pillai, A. S., and Horwitz, B. (2012a). Using large-scale neural models to interpret connectivity measures of cortico-cortical dynamics at millisecond temporal resolution. *Front. Syst. Neurosci.* 5:102. doi: 10.3389/fnsys.2011.00102

Banerjee, A., Pillai, A. S., Sperling, J. R., Smith, J. F., and Horwitz, B. (2012b). Temporal microstructure of cortical networks (TMCN) underlying task-related differences. *Neuroimage* 62, 1643–1657. doi: 10.1016/j.neuroimage.2012.06.014

Banerjee, A., Tognoli, E., Assisi, C. G., Kelso, J. A., and Jirsa, V. K. (2008). Mode level cognitive subtraction (MLCS) quantifies spatiotemporal reorganization in large-scale brain topographies. *Neuroimage* 42, 663–674. doi: 10.1016/j.neuroimage.2008.04.260

Bassett, D. S., Zurn, P., and Gold, J. I. (2018). On the nature and use of models in network neuroscience. *Nat. Rev. Neurosci.* 19, 566–578. doi: 10.1038/s41583-018-0038-8

Beauchamp, M. S. (2005). Statistical criteria in FMRI studies of multisensory integration. *Neuroinformatics* 3, 93–113. doi: 10.1385/NI:3:2:093

Berti, S., and Schroger, E. (2003). Working memory controls involuntary attention switching: evidence from an auditory distraction paradigm. *Eur. J. Neurosci.* 17, 1119–1122. doi: 10.1046/j.1460-9568.2003.02527.x

Berti, S., Roeber, U., and Schroger, E. (2004). Bottom-up influences on working memory: behavioral and electrophysiological distraction varies with distractor strength. *Exp. Psychol.* 51, 249–257. doi: 10.1027/1618-3169.51.4.249

Beukers, A. O., Buschman, T. J., Cohen, J. D., and Norman, K. A. (2021). Is activity silent working memory simply episodic memory? *Trends Cogn. Sci.* 25, 284–293. doi: 10.1016/j.tics.2021.01.003

Bichot, N. P., Heard, M. T., DeGennaro, E. M., and Desimone, R. (2015). A source for feature-based attention in the prefrontal cortex. *Neuron* 88, 832–844. doi: 10.1016/j.neuron.2015.10.001

Bieser, A. (1998). Processing of twitter-call fundamental frequencies in insula and auditory cortex of squirrel monkeys. *Exp. Brain Res.* 122, 139–148. doi: 10.1007/s002210050501

Bowling, J. T., Friston, K. J., and Hopfinger, J. B. (2020). Top-down versus bottom-up attention differentially modulate frontal-parietal connectivity. *Hum. Brain Mapp.* 41, 928–942. doi: 10.1002/hbm.24850

Calvert, G. A. (2001). Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cereb. Cortex* 11, 1110–1123. doi: 10.1093/cercor/11.12.1110

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.* 6:27755. doi: 10.1038/srep27755

Clapp, W. C., Rubens, M. T., and Gazzaley, A. (2010). Mechanisms of working memory disruption by external interference. *Cereb. Cortex* 20, 859–872. doi: 10.1093/cercor/bhp150

Colombo, M., Rodman, H. R., and Gross, C. G. (1996). The effects of superior temporal cortex lesions on the processing and retention of auditory information in monkeys (*Cebus apella*). *J. Neurosci.* 16, 4501–4517. doi: 10.1523/JNEUROSCI.16-14-04501.1996

Corchs, S., and Deco, G. (2002). Large-scale neural model for visual attention: integration of experimental single-cell and fMRI data. *Cereb. Cortex* 12, 339–348. doi: 10.1093/cercor/12.4.339

Cowan, N. (2001). The magic number 4 in short-term memory: a reconsideration of mental storage capacity. *Behav. Brain Sci.* 24, 87–114. doi: 10.1017/S0140525X01003922

Deco, G., Jirsa, V. K., Robinson, P. A., Breakspear, M., and Friston, K. (2008). The dynamic brain: from spiking neurons to neural masses and cortical fields. *PLoS Comput. Biol.* 4:e1000092. doi: 10.1371/journal.pcbi.1000092

Deco, G., Jirsa, V., McIntosh, A. R., Sporns, O., and Kotter, R. (2009). Key role of coupling, delay, and noise in resting brain fluctuations. *Proc. Natl. Acad. Sci. U.S.A.* 106, 10302–10307. doi: 10.1073/pnas.0901831106

D'Esposito, M., and Postle, B. R. (2015). The cognitive neuroscience of working memory. *Annu. Rev. Psychol.* 66, 115–142. doi: 10.1146/annurev-psych-010814-015031

Devereux, B. J., Clarke, A., and Tyler, L. K. (2018). Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. *Sci. Rep.* 8:10636. doi: 10.1038/s41598-018-28865-1

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., et al. (2012). A large-scale model of the functioning brain. *Science* 338, 1202–1205. doi: 10.1126/science.1225266

Eriksson, J., Vogel, E. K., Lansner, A., Bergstrom, F., and Nyberg, L. (2015). Neurocognitive architecture of working memory. *Neuron* 88, 33–46. doi: 10.1016/j.neuron.2015.09.020

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K. J., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *Neuroimage* 19, 1273–1302. doi: 10.1016/S1053-8119(03)00202-7

Fritz, J., Mishkin, M., and Saunders, R. C. (2005). In search of an auditory engram. *Proc. Natl. Acad. Sci. U.S.A.* 102, 9359–9364. doi: 10.1073/pnas.0503998102

Funahashi, S., Bruce, C., and Goldman-Rakic, P. S. (1990). Visuospatial coding in primate prefrontal neurons revealed by oculomotor paradigms. *J. Neurophysiol.* 63, 814–831. doi: 10.1152/jn.1990.63.4.814

Fuster, J. M., Bauer, R. H., and Jervey, J. P. (1982). Cellular discharge in the dorsolateral prefrontal cortex of the monkey in cognitive tasks. *Exp. Neurol.* 77, 679–694. doi: 10.1016/0014-4886(82)90238-2

Gilbert, J. R., Pillai, A. S., and Horwitz, B. (2013). Assessing crossmodal matching of abstract auditory and visual stimuli in posterior superior temporal sulcus with MEG. *Brain Cogn.* 82, 161–170. doi: 10.1016/j.bandc.2013.03.004

Godey, B., Atencio, C. A., Bonham, B. H., Schreiner, C. E., and Cheung, S. W. (2005). Functional organization of squirrel monkey primary auditory cortex: responses to frequency-modulation sweeps. *J. Neurophysiol.* 94, 1299–1311. doi: 10.1152/jn.00950.2004

Griffiths, T. D., and Warren, J. D. (2004). What is an auditory object? *Nat. Rev. Neurosci.* 5, 887–892. doi: 10.1038/nrn1538

Gu, X., Hof, P. R., Friston, K. J., and Fan, J. (2013). Anterior insular cortex and emotional awareness. *J. Comp. Neurol.* 521, 3371–3388. doi: 10.1002/cne.23368

Hackett, T. A. (2011). Information flow in the auditory cortical network. *Hear. Res.* 271, 133–146. doi: 10.1016/j.heares.2010.01.011

Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C. J., Wedeen, V. J., et al. (2008). Mapping the structural core of human cerebral cortex. *PLoS Biol.* 6:e159. doi: 10.1371/journal.pbio.0060159

Ham, T., Leff, A., de Boissezon, X., Joffe, A., and Sharp, D. J. (2013). Cognitive control and the salience network: an investigation of error processing and effective connectivity. *J. Neurosci.* 33, 7091–7098. doi: 10.1523/JNEUROSCI.4692-12.2013

Haxby, J. V., Grady, C. L., Horwitz, B., Ungerleider, L. G., Mishkin, M., Carson, R. E., et al. (1991). Dissociation of object and spatial visual processing pathways in human extrastriate cortex. *Proc. Natl. Acad. Sci. U.S.A.* 88, 1621–1625. doi: 10.1073/pnas.88.5.1621

Haxby, J. V., Ungerleider, L. G., Horwitz, B., Rapoport, S. I., and Grady, C. L. (1995). Hemispheric differences in neural systems for face working memory: a PET-rCBF Study. *Hum. Brain Mapp.* 3, 68–82. doi: 10.1002/hbm.460030204

Hopfinger, J. B., and West, V. M. (2006). Interactions between endogenous and exogenous attention on cortical visual processing. *Neuroimage* 31, 774–789. doi: 10.1016/j.neuroimage.2005.12.049

Horwitz, B., and Tagamets, M.-A. (1999). Predicting human functional maps with neural net modeling. *Hum. Brain Mapp.* 8, 137–142. doi: 10.1002/(SICI)1097-0193199:8:2/3<137::AID-HBM11>3.0.CO;2-B

Horwitz, B., Friston, K. J., and Taylor, J. G. (2000). Neural modeling and functional brain imaging: an overview. *Neural Netw.* 13, 829–846. doi: 10.1016/S0893-6080(00)00062-9

Horwitz, B., Tagamets, M.-A., and McIntosh, A. R. (1999). Neural modeling, functional brain imaging, and cognition. *Trends Cogn. Sci.* 3, 91–98. doi: 10.1016/S1364-6613(99)01282-6

Horwitz, B., Warner, B., Fitzer, J., Tagamets, M.-A., Husain, F. T., and Long, T. W. (2005). Investigating the neural basis for functional and effective connectivity: application to fMRI. *Philos. Trans. R. Soc. Lond. B* 360, 1093–1108. doi: 10.1098/rstb.2005.1647

Hsieh, I. H., Fillmore, P., Rong, F., Hickok, G., and Saberi, K. (2012). FM-selective networks in human auditory cortex revealed using fMRI and multivariate pattern classification. *J. Cogn. Neurosci.* 24, 1896–1907. doi: 10.1162/jocn_a_00254

Husain, F. T., Tagamets, M.-A., Fromm, S. J., Braun, A. R., and Horwitz, B. (2004). Relating neuronal dynamics for auditory object processing to neuroimaging activity. *NeuroImage* 21, 1701–1720. doi: 10.1016/j.neuroimage.2003.11.012

Ito, T., Hearne, L., Mill, R., Cocuzza, C., and Cole, M. W. (2020). Discovering the computational relevance of brain network organization. *Trends Cogn. Sci.* 24, 25–38. doi: 10.1016/j.tics.2019.10.005

Jirsa, V. K., Sporns, O., Breakspear, M., Deco, G., and McIntosh, A. R. (2010). Towards the virtual brain: network modeling of the intact and the damaged brain. *Arch. Ital. Biol.* 148, 189–205. doi: 10.4449/aib.v148i3.1223

Kaas, J. H., and Hackett, T. A. (1999). 'What' and 'where' processing in auditory cortex. *Nat. Neurosci.* 2, 1045–1047. doi: 10.1038/15967

Kapur, S. (2003). Psychosis as a state of aberrant salience: a framework linking biology, phenomenology, and pharmacology in schizophrenia. *Am. J. Psychiatry* 160, 13–23. doi: 10.1176/appi.ajp.160.1.13

Kastner, S., and Ungerleider, L. G. (2000). Mechanisms of visual attention in the human cortex. *Annu. Rev. Neurosci.* 23, 315–341. doi: 10.1146/annurev.neuro.23.1.315

Kay, K. N. (2018). Principles for models of neural information processing. *Neuroimage* 180, 101–109. doi: 10.1016/j.neuroimage.2017.08.016

Keil, J., and Senkowski, D. (2018). Neural oscillations orchestrate multisensory processing. *Neuroscientist* 24, 609–626. doi: 10.1177/1073858418755352

Keil, J., Pomper, U., and Senkowski, D. (2016). Distinct patterns of local oscillatory activity and functional connectivity underlie intersensory attention and temporal prediction. *Cortex* 74, 277–288. doi: 10.1016/j.cortex.2015.10.023

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., and McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* 98, 630–644.e1. doi: 10.1016/j.neuron.2018.03.044

Kikuchi, Y., Horwitz, B., and Mishkin, M. (2010). Hierarchical auditory processing directed rostrally along the monkey's supratemporal plane. *J. Neurosci.* 30, 13021–13030. doi: 10.1523/JNEUROSCI.2267-10.2010

Koechlin, E., Ody, C., and Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science* 302, 1181–1185. doi: 10.1126/science.1088545

Kriegeskorte, N., and Diedrichsen, J. (2016). Inferring brain-computational mechanisms with models of activity measurements. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 371:20160278. doi: 10.1098/rstb.2016.0278

Kumar, S., Stephan, K. E., Warren, J. D., Friston, K. J., and Griffiths, T. D. (2007). Hierarchical processing of auditory objects in humans. *PLoS Comput. Biol.* 3:e100. doi: 10.1371/journal.pcbi.0030100

Lamichhane, B., Adhikari, B. M., and Dhamala, M. (2016). Salience network activity in perceptual decisions. *Brain Connect.* 6, 558–571. doi: 10.1089/brain.2015.0392

Leaver, A. M., and Rauschecker, J. P. (2010). Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. *J. Neurosci.* 30, 7604–7612. doi: 10.1523/JNEUROSCI.0296-10.2010

Leavitt, M. L., Mendoza-Halliday, D., and Martinez-Trujillo, J. C. (2017). Sustained activity encoding working memories: not fully distributed. *Trends Neurosci.* 40, 328–346. doi: 10.1016/j.tins.2017.04.004

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Liu, Q., Ulloa, A., and Horwitz, B. (2017). Using a large-scale neural model of cortical object processing to investigate the neural substrate for managing multiple items in short-term memory. *J. Cogn. Neurosci.* 29, 1860–1876. doi: 10.1162/jocn_a_01163

Lorenc, E. S., Mallett, R., and Lewis-Peacock, J. A. (2021). Distraction in visual working memory: resistance is not futile. *Trends Cogn. Sci.* 25, 228–239. doi: 10.1016/j.tics.2020.12.004

Lowe, M. X., Mohsenzadeh, Y., Lahner, B., Charest, I., Oliva, A., and Teng, S. (2020). Spatiotemporal dynamics of sound representations reveal a hierarchical progression of category selectivity. *bioRxiv* [Preprint] doi: 10.1101/2020.06.12.149120

Lynn, C. W., and Bassett, D. S. (2019). The physics of brain network structure, function, and control. *Nat. Rev. Phys.* 1, 318–332. doi: 10.1038/s42254-019-0040-8

Ma, W. J., Husain, M., and Bays, P. M. (2014). Changing concepts of working memory. *Nat. Neurosci.* 17, 347–356. doi: 10.1038/nn.3655

Masse, N. Y., Yang, G. R., Song, H. F., Wang, X. J., and Freedman, D. J. (2019). Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nat. Neurosci.* 22, 1159–1167. doi: 10.1038/s41593-019-0414-3

Mejias, J. F., and Wang, X.-J. (2022). Mechanisms of distributed working memory in a large-scale network of macaque neocortex. *Elife* 11 doi: 10.7554/eLife.72136

Mendelson, J. R., and Cynader, M. S. (1985). Sensitivity of cat primary auditory cortex (A1) neurons to the direction and rate of frequency modulation. *Brain Res.* 327, 331–335. doi: 10.1016/0006-8993(85)91530-6

Mendelson, J. R., Schreiner, C. E., Sutter, M. L., and Grasse, K. L. (1993). Functional topography of cat primary auditory cortex: responses to frequency-modulated sweeps. *Exp. Brain Res.* 94, 65–87. doi: 10.1007/BF00230471

Mendoza-Halliday, D., and Martinez-Trujillo, J. C. (2017). Neuronal population coding of perceived and memorized visual features in the lateral prefrontal cortex. *Nat. Commun.* 8:15471. doi: 10.1038/ncomms15471

Menon, V., and Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain Struct. Funct.* 214, 655–667. doi: 10.1007/s00429-010-0262-0

Michail, G., Senkowski, D., Niedeggen, M., and Keil, J. (2021). Memory load alters perception-related neural oscillations during multisensory integration. *J. Neurosci.* 41, 1505–1515. doi: 10.1523/JNEUROSCI.1397-20.2020

Miller, G. A. (1956). The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* 63, 81–97. doi: 10.1037/h0043158

Mishkin, M., Ungerleider, L. G., and Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. *Trends Neurosci.* 6, 414–417. doi: 10.1016/0166-2236(83)90190-X

Munoz, M., Mishkin, M., and Saunders, R. C. (2009). Resection of the medial temporal lobe disconnects the rostral superior temporal gyrus from some of its projection targets in the frontal lobe and thalamus. *Cereb. Cortex* 19, 2114–2130. doi: 10.1093/cercor/bhn236

Murray, E. A., and Richmond, B. J. (2001). Role of perirhinal cortex in object perception, memory, and associations. *Curr. Opin. Neurobiol.* 11, 188–193. doi: 10.1016/S0959-4388(00)00195-1

Naselaris, T., Bassett, D. S., Fletcher, A. K., Kording, K., Kriegeskorte, N., Nienborg, H., et al. (2018). Cognitive computational neuroscience: a new conference for an emerging discipline. *Trends Cogn. Sci.* 22, 365–367. doi: 10.1016/j.tics.2018.02.008

Nourski, K. V. (2017). Auditory processing in the human cortex: an intracranial electrophysiology perspective. *Laryngoscope Investig. Otolaryngol.* 2, 147–156. doi: 10.1002/lio2.73

Orhan, A. E., and Ma, W. J. (2019). A diverse range of factors affect the nature of neural representations underlying short-term memory. *Nat. Neurosci.* 22, 275–282. doi: 10.1038/s41593-018-0314-y

Petkoski, S., and Jirsa, V. K. (2019). Transmission time delays organize the brain network synchronization. *Philos. Trans. A Math. Phys. Eng. Sci.* 377:20180132. doi: 10.1098/rsta.2018.0132

Pillai, A. S., Gilbert, J. R., and Horwitz, B. (2013). Early sensory cortex is activated in the absence of explicit input during crossmodal item retrieval: evidence from MEG. *Behav. Brain Res.* 238, 265–272. doi: 10.1016/j.bbr.2012.10.011

Posner, M. I., and Cohen, Y. (1984). Components of visual orientating. *Atten. Perform. X* 32, 531–556.

Pulvermuller, F., Tomasello, R., Henningsen-Schomers, M. R., and Wennekers, T. (2021). Biological constraints on neural network models of cognitive function. *Nat. Rev. Neurosci.* 22, 488–502. doi: 10.1038/s41583-021-00473-5

Quak, M., London, R. E., and Talsma, D. (2015). A multisensory perspective of working memory. *Front. Hum. Neurosci.* 9:197. doi: 10.3389/fnhum.2015.00197

Ranganath, C., and Blumenfeld, R. S. (2005). Doubts about double dissociations between short- and long-term memory. *Trends Cogn. Sci.* 9, 374–380. doi: 10.1016/j.tics.2005.06.009

Rauschecker, J. P. (1997). Processing of complex sounds in the auditory cortex of cat, monkey and man. *Acta Otolaryngol. (Stockh.)* 532(Suppl.), 34–38. doi: 10.3109/00016489709126142

Rauschecker, J. P. (1998). Parallel processing in the auditory cortex of primates. *Audiol. Neurootol.* 3, 86–103. doi: 10.1159/000013784

SanMiguel, I., Corral, M. J., and Escera, C. (2008). When loading working memory reduces distraction: behavioral and electrophysiological evidence from an auditory-visual distraction paradigm. *J. Cogn. Neurosci.* 20, 1131–1145. doi: 10.1162/jocn.2008.20078

Sanz Leon, P., Knock, S. A., Woodman, M. M., Domide, L., Mersmann, J., McIntosh, A. R., et al. (2013). The virtual brain: a simulator of primate brain network dynamics. *Front. Neuoinform.* 7:10. doi: 10.3389/fninf.2013.00010

Saxe, A., Nelli, S., and Summerfield, C. (2021). If deep learning is the answer, what is the question? *Nat. Rev. Neurosci.* 22, 55–67. doi: 10.1038/s41583-020-00395-8

Schreiner, C. E., Read, H. L., and Sutter, M. L. (2000). Modular organization of frequency integration in primary auditory cortex. *Annu. Rev. Neurosci.* 23, 501–529. doi: 10.1146/annurev.neuro.23.1.501

Scott, B. H., Mishkin, M., and Yin, P. (2014). Neural correlates of auditory short-term memory in rostral superior temporal cortex. *Curr. Biol.* 24, 2767–2775. doi: 10.1016/j.cub.2014.10.004

Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., et al. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *J. Neurosci.* 27, 2349–2356. doi: 10.1523/JNEUROSCI.5587-06.2007

Shamma, S. (2001). On the role of space and time in auditory processing. *Trends Cogn. Sci.* 5, 340–348. doi: 10.1016/S1364-6613(00)01704-6

Shamma, S. A., Fleshman, J. W., Wiser, P. R., and Versnel, H. (1993). Organization of response areas in ferret primary auditory cortex. *J. Neurophysiol.* 69, 367–383. doi: 10.1152/jn.1993.69.2.367

Simmons, W. K., and Barsalou, L. W. (2003). The similarity-in-topography principle: reconciling theories of conceptual deficits. *Cogn. Neuropsychol.* 20, 451–486. doi: 10.1080/02643290342000032

Smith, J. F., Alexander, G. E., Chen, K., Husain, F. T., Kim, J., Pajor, N., et al. (2010). Imaging systems level consolidation of novel associate memories: a longitudinal neuroimaging study. *Neuroimage* 50, 826–836. doi: 10.1016/j.neuroimage.2009.11.053

Song, H. F., Yang, G. R., and Wang, X. J. (2016). Training excitatory-inhibitory recurrent neural networks for cognitive tasks: a simple and flexible framework. *PLoS Comput. Biol.* 12:e1004792. doi: 10.1371/journal.pcbi.1004792

Spinks, J. A., Zhang, J. X., Fox, P. T., Gao, J. H., and Tan, L. H. (2004). More workload on the central executive of working memory, less attention capture by novel visual distractors: evidence from an fMRI study. *Neuroimage* 23, 517–524. doi: 10.1016/j.neuroimage.2004.06.025

Sridharan, D., Levitin, D. J., and Menon, V. (2008). A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. *Proc. Natl. Acad. Sci. U.S.A.* 105, 12569–12574. doi: 10.1073/pnas.0800005105

Stephan, K. E., Weiskopf, N., Drysdale, P. M., Robinson, P. A., and Friston, K. J. (2007). Comparing hemodynamic models with DCM. *Neuroimage* 38, 387–401. doi: 10.1016/j.neuroimage.2007.07.040

Sternberg, S. (1966). High-speed scanning in human memory. *Science* 153, 652–654. doi: 10.1126/science.153.3736.652

Sternberg, S. (1969). Memory-scanning: mental processes revealed by reaction-time experiments. *Am. Sci.* 57, 421–457. Available online at: https://www.jstor.org/stable/27828738

Suzuki, W. A., and Amaral, D. G. (1994). Perirhinal and parahippocampal cortices of the macaque monkey: cortical afferents. *J. Comp. Neurol.* 350, 497–533. doi: 10.1002/cne.903500402

Tagamets, M.-A., and Horwitz, B. (1998). Integrating electrophysiological and anatomical experimental data to create a large-scale model that simulates a delayed match-to-sample human brain imaging study. *Cereb. Cortex* 8, 310–320. doi: 10.1093/cercor/8.4.310

Talairach, J., and Tournoux, P. (1988). *Co-Planar Stereotaxic Atlas of the Human Brain.* New York, NY: Thieme Medical Publishers, Inc.

Tanabe, H. C., Honda, M., and Sadato, N. (2005). Functionally segregated neural substrates for arbitrary audiovisual paired-association learning. *J. Neurosci.* 25, 6409–6418. doi: 10.1523/JNEUROSCI.0636-05.2005

Tartaglia, E. M., Brunel, N., and Mongillo, G. (2015). Modulation of network excitability by persistent activity: how working memory affects the response to incoming stimuli. *PLoS Comput. Biol.* 11:e1004059. doi: 10.1371/journal.pcbi.1004059

Tian, B., and Rauschecker, J. P. (2004). Processing of frequency-modulated sounds in the lateral auditory belt cortex of the rhesus monkey. *J. Neurophysiol.* 92, 2993–3013. doi: 10.1152/jn.00472.2003

Uddin, L. Q. (2015). Salience processing and insular cortical function and dysfunction. *Nat. Rev. Neurosci.* 16, 55–61. doi: 10.1038/nrn3857

Ulloa, A., and Horwitz, B. (2016). Embedding task-based neural models into a connectome-based model of the cerebral cortex. *Front. Neuroinform.* 10:32. doi: 10.3389/fninf.2016.00032

Ulloa, A., Husain, F. T., Kemeny, S., Xu, J., Braun, A. R., and Horwitz, B. (2008). Neural mechanisms of auditory discrimination of long-duration tonal patterns: a neural modeling and FMRI study. *J. Integr. Neurosci.* 7, 501–527. doi: 10.1142/S021963520800199X

Ungerleider, L. G., and Haxby, J. V. (1994). 'What' and 'where' in the human brain. *Curr. Opin. Neurobiol.* 4, 157–165. doi: 10.1016/0959-4388(94)90066-3

Ungerleider, L. G., and Mishkin, M. (1982). "Two cortical visual systems," in *Analysis of Visual Behavior*, eds D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield (Cambridge: MIT Press), 549–586.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., et al. (2013). The WU-minn human connectome project: an overview. *Neuroimage* 80, 62–79. doi: 10.1016/j.neuroimage.2013.05.041

Visscher, K. M., Kaplan, E., Kahana, M. J., and Sekuler, R. (2007). Auditory short-term memory behaves like visual short-term memory. *PLoS Biol.* 5:e56. doi: 10.1371/journal.pbio.0050056

Wilson, H. R., and Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys J.* 12, 1–24. doi: 10.1016/S0006-3495(72)86068-5

Yang, G. R., and Wang, X. J. (2020). Artificial neural networks for neuroscientists: a primer. *Neuron* 107, 1048–1070. doi: 10.1016/j.neuron.2020.09.005

Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., and Wang, X. J. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nat. Neurosci.* 22, 297–306. doi: 10.1038/s41593-018-0310-2

Yantis, S., and Jonides, J. (1990). Abrupt visual onsets and selective attention: voluntary versus automatic allocation. *J. Exp. Psychol. Hum. Percept. Perform.* 16, 121–134. doi: 10.1037/0096-1523.16.1.121

Zalta, A., Petkoski, S., and Morillon, B. (2020). Natural rhythms of periodic temporal attention. *Nat. Commun.* 11:1051. doi: 10.1038/s41467-020-14888-8

APPENDIX

Model Formulation

We simulated the electrical activity and fMRI activity of the model while implementing the tasks discussed in the main text using the large-scale neural modeling (LSNM) framework developed by Tagamets and Horwitz (1998), Horwitz and Tagamets (1999), Husain et al. (2004), and Liu et al. (2017). Our modeling framework is based on several critical assumptions. First, the models are constructed to perform specific tasks (e.g., the DMS task). Each submodel is designed to represent the visual/auditory object processing stream running from primary sensory cortex to prefrontal cortex. The idea is to employ neural data obtained in primate experiments, when available, to constrain the model parameters (e.g., the connection weights). For example, we used the fact that the spatial receptive field of neurons increases as one moves along the pathway from posterior to anterior cortex in the visual model; likewise, the corresponding assumption in the auditory model is that the spectrotemporal window of integration increases. The neural data generated by our model can be used to simulate neuroimaging data that in turn can be directly compared to human fMRI or PET or MEG data.

The simulated electrical activity of each E and I element of the basic unit, a modified Wilson and Cowan (1972) configuration, is determined by a sigmoidal function of the summed synaptic inputs that arrive at the unit, which corresponds to average spiking rates from single-cell recordings. This electrical activity is given by the following equations:

$$\frac{dE_i(t)}{dt} = \Delta \left(\frac{1}{1 + e^{-K_E[w_{EE}E_i(t) + w_{IE}I_i(t) + in_{iE}(t) - \phi_E + N(t)]}} \right) - \delta E_i(t)$$

and

$$\frac{dI_i(t)}{dt} = \Delta \left(\frac{1}{1 + e^{-K_I[w_{EI}E_i(t) + in_{iI}(t) - \phi_I + N(t)]}} \right) - \delta I_i(t)$$

where, Δ is the rate of change, δ is the decay rate, K_E, K_I are gain constants, w_{EE}, w_{IE}, w_{EI} are the connectivity weights within one neuronal unit, ϕ_E, ϕ_I are the input thresholds, N(t) is the noise. $in_{iE}(t)$, $in_{iI}(t)$, are the incoming inputs from other nodes. $in_{iE}(t)$ is given by:

$$in_{iE}(t) = \sum_j w_{ji}^E E_j(t) + \sum_j w_{ji}^I I_j(t) + \sum_j c_{ji} z_{ji}^c C_j(t)$$

where, w_{ji}^E and w_{ji}^I are the weights for connections from the excitatory (E) and inhibitory (I) elements of jth LSNM unit to the excitatory element of ith LSNM unit, C_j is the electrical activity of the connectome excitatory unit j connected to LSNM unit i, and z_{ji}^C is the connection weight. c_{ji} is a coupling term obtained by the Gaussian pseudo-random number generator of Python. $in_{iI}(t)$ is given by:

$$in_{iI}(t) = \sum_k w_{ki}^E E_k(t) + \sum_k w_{ki}^I I_k(t)$$

where, w_{ki}^E and w_{ki}^I are the weights for connections from the excitatory (E) and inhibitory (I) elements of kth LSNM unit to the excitatory element of ith LSNM unit. Each timestep corresponds to ∼5 ms.

Ulloa and Horwitz (2016) embedded the visual LSNM into a whole-brain framework using The Virtual Brain (TVB) software package (Sanz Leon et al., 2013). TVB is a simulator that combines (i) white matter structural connections among brain regions to simulate long-range connections, (ii) a neuronal population model to simulate local brain activity, and (iii) forward models that convert simulated neural activity into simulated functional neuroimaging data (i.e., fMRI or EEG/MEG). In the current study, TVB neurons provide neural noise to the embedded LSNM. The structural connectome used is due to Hagmann et al. (2008), which is composed of 998 nodes (regions of interest), and the simulated neuronal microcircuits at each TVB node are Wilson–Cowan units. The integrated synaptic activity is computed prior to computing fMRI BOLD activity by spatially integrating over each LSNM module and temporally integrating over 50 ms (Horwitz and Tagamets, 1999).

$$rSYN = \sum IN_i(t)$$

where, $IN_i(t)$ is the sum of absolute values of inputs to the excitatory and inhibitory elements of unit i, at time t:

$$IN_i(t) = w_{EE}E_i(t) + w_{EI}E_i(t) + |w_{IE}I_i(t)| + \sum_{k,i} w_{ki}E_k(t)$$

The last term is the sum of synaptic connections from all other LSNM units and connectome nodes to the ith unit in LSNM.

Details concerning the connectivity and the connection weights in the previous versions of the model can be found in the Tagamets and Horwitz, Husain et al., and Liu et al. papers. In brief, the connection weights are generally taken from previous papers. The connections for the new modules are hand-wired as a proof-of-hypothesis study.

Modeling the aINS

The anterior insula is modeled as two attention modules, one for vision and one for audition, competitively inhibiting each other. The inhibition weights between the two modules are equally chosen in a way such that the two modules can be suppressed during the resting-state when only visual and auditory noises are provided and can be excited by salient visual/auditory stimuli. The choice of a specific value depends on the definition of salient stimuli.