frontiers | Frontiers in Systems Neuroscience

ORIGINAL RESEARCH
published: 12 May 2022
doi: 10.3389/fnsys.2022.845177

Multisensory Concept Learning Framework Based on Spiking Neural Networks

Yuwei Wang[1,2] *and Yi Zeng*[1,2,3,4]*

[1] Research Center for Brain-inspired Intelligence, Institute of Automation, Chinese Academy of Sciences, Beijing, China, [2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China, [3] Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, China, [4] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China

Concept learning highly depends on multisensory integration. In this study, we propose a multisensory concept learning framework based on brain-inspired spiking neural networks to create integrated vectors relying on the concept's perceptual strength of auditory, gustatory, haptic, olfactory, and visual. With different assumptions, two paradigms: Independent Merge (IM) and Associate Merge (AM) are designed in the framework. For testing, we employed eight distinct neural models and three multisensory representation datasets. The experiments show that integrated vectors are closer to human beings than the non-integrated ones. Furthermore, we systematically analyze the similarities and differences between IM and AM paradigms and validate the generality of our framework.

Keywords: concept learning, multisensory, spiking neural networks, brain-inspired, Independent Merge, Associate Merge

Edited by:
Yan Mark Yufik,
Virtual Structures Research Inc., United States

Reviewed by:
Laxmi R. Iyer,
*Institute for Infocomm Research (A*STAR), Singapore*
Sun Zhe,
RIKEN, Japan
Shangbin Chen,
Huazhong University of Science and Technology, China

***Correspondence:**
Yi Zeng
yi.zeng@ia.ac.cn

Received: *29 December 2021*
Accepted: *20 April 2022*
Published: *12 May 2022*

Citation:
Wang Y and Zeng Y (2022) Multisensory Concept Learning Framework Based on Spiking Neural Networks. Front. Syst. Neurosci. 16:845177. doi: 10.3389/fnsys.2022.845177

1. INTRODUCTION

Concept learning, or the ability to recognize commonalities and accentuate contrasts across a group of linked events in order to generate structured knowledge, is a crucial component of cognition (Roshan et al., 2001). Multisensory integration benefits concept learning (Shams and Seitz, 2008) and plays an important role in semantic processing (Xu et al., 2017; Wang et al., 2020). For example, when we learn the concept of "tea," acoustically, we will perceive the sound of pouring water and brewing, the sound of clashing porcelain, the sound of drinking tea; on taste, we can feel the tea is a bit bitter, astringent or sweet; in touch, tea is liquid and we can feel its temperature; on smell, we can perceive the faint scent and visually, it often appears together with the teapot or tea bowl, and the tea leaves will have different colors. Combining information from multiple senses can produce enhanced perception and learning, faster response times, and improved detection, discrimination, and recognition capabilities (Calvert and Thesen, 2004). In the brain, multisensory integration occurs mostly in the superior colliculus according to existing studies (Calvert and Thesen, 2004; Cappe et al., 2009). Multisensory integration is a field that has attracted the interest of cognitive psychologists, biologists, computational neuroscientists, and artificial intelligence researchers. The term "multisensory concept learning" is used in this work to describe the process of learning concepts using a model that mimics humans and combines information from multiple senses.

For the computational models of multisensory integration, cognitive psychologists' models are usually focused on model design and validation from the mechanism of multisensory integration. These models are highly interpretable, taking neuroimaging and behavioral studies

into consideration. The cue combination model based on Bayesian decision theory is a classical model for analyzing multisensory integration in cognitive psychology. It mainly models the stimuli of different modalities as the likelihood functions of Gaussian (Ursino et al., 2009, 2014) or Poisson (Anastasio et al., 2014) distributions with different parameters, and calculates the best combination of each modality that makes the maximum posterior distribution through the assumption of conditional independence and Bayesian rules. Anastasio et al. built a model of visual and auditory fusion that combines neuronal dynamic equations with feedback information, and this model verified that multimodal stimuli have less response time than unimodal stimuli (Anastasio et al., 2014). Parise et al. proposed multisensory correlation detector based models to describe correlation, lag, and synchrony across the senses (Parise and Ernst, 2016). A purely visual haptic prediction model is presented by Gao et al. (2016) with CNNs and LSTMs, which enables robots to "feel" without physical interaction. Gepner et al. (2015) developed a linear-nonlinear-Poisson cascade model that incorporates information from olfaction and vision to mimic Drosophila larvae navigation decisions, and the model was able to predict Drosophila larvae reaction to new stimulus patterns well.

For artificial intelligence researchers, they have proposed different types of multisensory integration models based on the available data and machine learning methods, such as direct concatenation (Kiela and Bottou, 2014; Collell et al., 2017; Wang et al., 2018b), canonical correlation analysis (Silberer et al., 2013; Hill et al., 2014), singular value decomposition of the integration matrix (Bruni et al., 2014), multisensory context (Hill and Korhonen, 2014), autoencoders (Silberer and Lapata, 2014; Wang et al., 2018a), and tensor fusion networks (Zadeh et al., 2017; Liu et al., 2018; Verma et al., 2019). These works are mostly focused on concept learning and sentiment analysis tasks and are based on modeling of speech, text, and image data, which are commonly utilized in AI.

To our knowledge, no work exists to model the five senses of vision, hearing, touch, taste, and smell together. This might be because controlling elements for experimental design is challenging for cognitive psychologists, while data for some modalities is difficult to get using perceptrons for AI researchers. Meanwhile, cognitive psychologists have published several multisensory datasets by asking volunteers how much they perceive a specific concept through their auditory, gustatory, tactile, olfactory, and visual senses in order to establish the strength of each modality. This provides a solid basis for the design of a multisensory integration model that includes these five modalities. In this article, we will model multisensory integration using brain-like spiking neural networks and merge input from five different modalities to generate integrated representations.

This paper is organized as follows: Section 2 will introduce relevant studies to our model, such as multisensory datasets and fundamental SNN models; Section 3 will describe the multisensory concept learning framework based on SNNs, which includes the Independent Merge and Associate Merge paradigms. Section 4 will exhibit the experiments, and the final section will explore the future works.

2. RELATED WORKS

2.1. Multisensory Concept Representation Datasets

Cognitive psychologists label the multisensory datasets of concepts by asking volunteers how much each concept is acquired through a specific modality and introducing statistical methods to establish the representation vector for each concept. The pioneering work in this area is by Lynott and Connell (2013), who proposed modality exclusivity norms for 423 adjective concepts (Lynott and Connell, 2009) and 400 nominal concepts on strength of association with each of the five primary sensory modalities (auditory, gustatory, haptic, olfactory, visual). In this article, we combine these two datasets of their previous works and denote them as LC823. Lancaster Sensorimotor Norms were published by Lynott et al. (2019), which included six perceptual modalities (auditory, gustatory, haptic, interoceptive, olfactory, visual) and five action effectors (foot/leg, hand/arm, head, mouth, torso). This dataset (we denote as Lancaster40k) is the largest ever, with 39,707 psycholinguistic concepts (Lynott et al., 2019). Binder et al. (2016) constructed a set of brain-based componential semantic representation (BBSR) with 65 experienced attributes, including sensory, motor, spatial, temporal, affective, social, and cognitive experiences, relying on more recent neurobiological findings. This dataset contains 535 concepts and does an excellent work of separating a priori conceptual categories and capturing semantic similarity (Binder et al., 2016). **Figure 1** shows the the concept "honey" in the multisensory concept representation datasets mentioned.

We'll concentrate on the effect of five forms of senses in this article: vision, touch, sound, smell, and taste. In BBSR, we employ the average value of the sub-dimensions corresponding to these five senses, while using the first five dimensions of Lancaster40k.

2.2. Basic Neuron and Synapse Models

Spiking neural networks (SNNs) are commonly referred to be the third generation of neural network models since theyăareăinspired by current discoveries in neuroscience (Maass, 1997). Neurons are the basic processing units of the brain. They communicate with each other *via* synapses. When the membrane potential reaches a threshold, a spike is produced. External stimuli are conveyed by firing rate and the temporal pattern of spike trains (Rieke et al., 1999; Gerstner and Kistler, 2002). SNNs integrate temporal information into the model and are capable of accurately describing spike timing with dynamic changes in synaptic weights which are more biologically plausible. We will use SNNs as the foundation of our model to build a human-like multisensory integration concept learning framework. Here, we briefly outline the neural and synaptic models that will be used in this research.

2.2.1. IF Neural Model

The integrate-and-fire (IF) model is a large family of models which assumes that a membrane potential threshold controls the

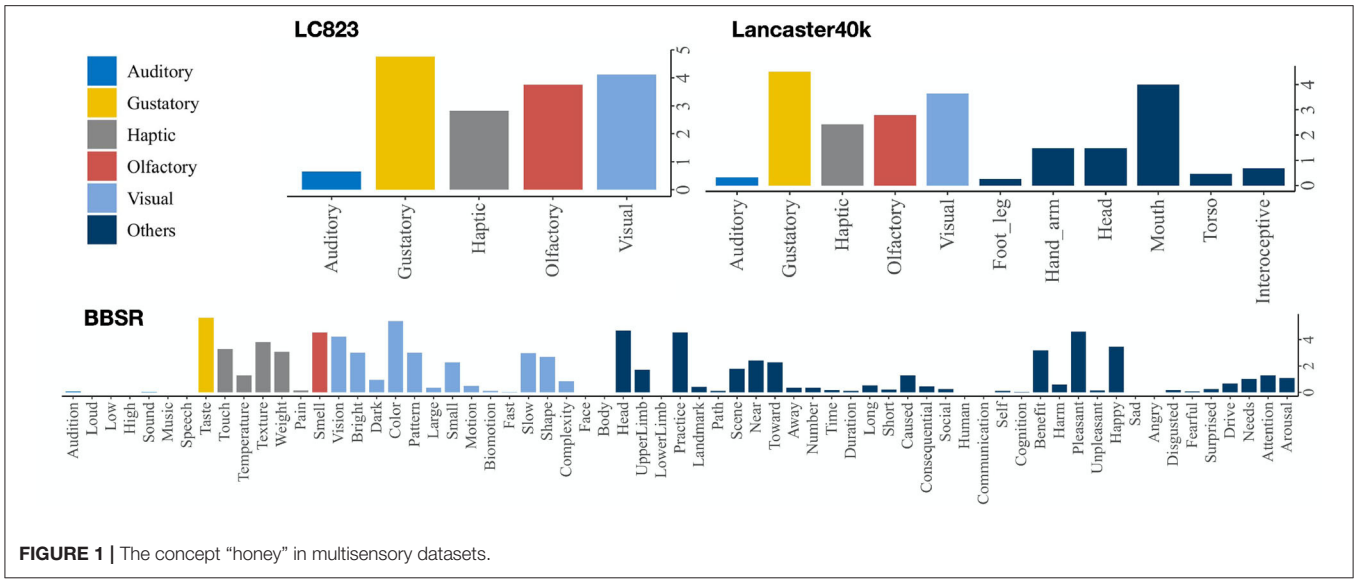

FIGURE 1 | The concept "honey" in multisensory datasets.

spikes of neurons. A spike is fired when the somatic membrane potential exceeds the threshold, and the membrane potential is resumed to reset potential (Gerstner and Kistler, 2002). The neural processing is properly formalized by the model. In this article, we follow a standard implementation (Troyer and Miller, 1997), and the membrane potential $v(t)$ obeys

$$\begin{aligned} \tau_{IF}\frac{dv(t)}{dt} &= v_{rest} - v(t) + g_e(t)(E_e - v(t)) \\ if\ \ v(t) &> v_{th},\ \ v(t) \leftarrow v_r \end{aligned} \tag{1}$$

with the membrane time constant $\tau_{IF} = 20\ ms$, the resting potential $v_{rest} = -14\ mV$, the threshold for spike firing $v_{th} = 6\ mV$, the reset potential $v_r = 0\ mV$, and excitatory potential $E_e = 0\ mV$. Synaptic inputs are modeled as conductance g_e changes with $\tau_e \frac{dg_e}{dt} = -g_e$, where $\tau_e = 5\ mV$.

2.2.2. LIF Neural Model

The leaky integrate-and-fire (LIF) neuron model is one of the most popular spiking neuron models because it is biologically realistic and computationally easy to study and mimic. The LIF neuron's subthreshold dynamics are described by the equation below:

$$\begin{aligned} \tau_{LIF}\frac{dv(t)}{dt} &= v_{rest} - v(t) + R_m I \\ if\ v(t) &> v_{th},\ v(t) \leftarrow v_r \end{aligned} \tag{2}$$

In this paper, the membrane resistance constance $R_m = 1$, $\tau_{LIF} = 20$, $v_{rest} = 1.05$, $v_{th} = 1$, and $v_r = 0$.

2.2.3. Izhikevich Neural Model

Izhikevich model was first proposed in 2003 to replicate spiking and bursting behavior of known types of cortical neurons. The model combines the biological plausibility of Hodgkin and Huxley (1952) dynamics with the computing efficiency of integrate-and-fire neurons (Izhikevich, 2003). Biophysically accurate Hodgkin-Huxley neural models are reduced to a two-dimensional system of the following dynamics ordinary with bifurcation methods:

$$\begin{aligned} \frac{dv(t)}{dt} &= 0.04v(t)^2 + 5v(t) + 140 - u(t) + I, \\ \frac{du}{dt} &= a(bv(t) - u(t)) \\ if\ v(t) &> v_{th},\ v(t) \leftarrow c\ \ and\ \ u(t) \leftarrow u(t) + d \end{aligned} \tag{3}$$

where the time scale of the recovery variable u is described by the parameter a, the sensitivity of the recovery variable u to subthreshold changes of the membrane potential v is described by the parameter b, the parameter c defines the membrane potential v's after-spike reset value, which is induced by quick high-threshold K^+ conductances and after-spike reset of the recovery variable u induced by slow high-threshold Na^+ and K^+ conductances is described by the parameter d (Izhikevich, 2003).

The model simulates the spiking and bursting activity of known kinds of cortical or thalamic neurons such as resonator (RZ), fast spiking (FS), intrinsically bursting (IB), low-threshold spiking (LTS), regular spiking (RS), chattering (CH), and thalamo-cortical (TC) based on these four parameters. These

TABLE 1 | Izhikevich models.

Neurons	Izhikevich parameters			
	a	b	c	d
RZ (resonator)	0.10	0.25	−65	2
FS (fast spiking)	0.10	0.20	−65	2
IB (intrinsically bursting)	0.02	0.20	−55	4
LTS (low-threshold spiking)	0.02	0.25	−65	2
RS (regular spiking)	0.02	0.20	−65	8
CH (chattering)	0.02	0.20	−50	2
TC (thalamo-cortical)	0.02	0.25	−65	0.05

models are employed extensively in our work and details are illustrated in **Table 1**.

2.2.4. STDP Synapse Models

Spike-timing-dependent plasticity (STDP) is a biological process that modifies the strength of neural connections in the brain. Learning and information storage in the brain, as well as the growth and refinement of neural circuits throughout brain development, are thought to be influenced by STDP (Bi and Poo, 2001). The typical STDP model is used in this research, and the weight change Δw of a synapse relies on the relative time of presynaptic spike arrivals and postsynaptic spike arrivals. $\Delta w = \Sigma_{t_{pre}} \Sigma_{t_{post}} W(t_{post} - t_{pre})$, where the function $W(\cdot)$ is defined as:

$$W(\Delta t) = \begin{cases} A_+ exp(\frac{\Delta t}{\tau_+}) \ \Delta t > 0 \\ -A_- exp(-\frac{\Delta t}{\tau_-}) \ \Delta t < 0 \end{cases} \tag{4}$$

When implement STDP, we follow the way of Brian2 (Stimberg et al., 2019), which defines two variables a_{pre} and a_{post} as the "traces" of of pre- and post-synaptic activity, governed by the following differential equations

$$\begin{aligned} \tau_{pre} \frac{a_{pre}}{dt} &= -a_{pre} \\ \tau_{post} \frac{a_{post}}{dt} &= -a_{post} \end{aligned} \tag{5}$$

Once a presynaptic spike occurs, the presynaptic trace is updated and the weight is modified according to the rule

$$\begin{aligned} a_{pre} &\leftarrow a_{pre} + A_{pre} \\ w &\leftarrow w + a_{post} \end{aligned} \tag{6}$$

And when a postsynaptic spike occurs:

$$\begin{aligned} a_{post} &\leftarrow a_{post} + A_{post} \\ w &\leftarrow w + a_{pre} \end{aligned} \tag{7}$$

This is proved to be equivalent for the two kinds of STDP formulations. And, in this article $\tau_{pre} = \tau_{post} = 1ms$.

3. THE FRAMEWORK OF MULTISENSORY CONCEPT LEARNING FRAMEWORK BASED ON SPIKING NEURAL NETWORKS

We present a multisensory concept learning framework based on SNNs in this part. The model's input is a multisensory vector labeled by cognitive psychologists, with an integrated vector as the output following SNNs merging. Since there is no biological study to show whether the information of multiple senses is independent or associated before integration, two different paradigms: Independent Merge (IM) and Associate Merge (AM) are designed in our framework. The types of inputs and outputs are the same for both paradigms, but the architectural design of SNNs is different. These two paradigms involve the same phase in the framework, and only oneăparadigm is chosen for concept integration, depending on the assumption that whether multiple sensory input is independent before integration.

Figure 2 illustrates the workflow: Firstly, for each modality of the concept, we employ a neural model and transform its perceptual strength in the concept's multisensory vector into external stimuli to the neuron (we work on five sensory modalities: auditory, gustatory, haptic, olfactory, visual, so the dimensions of the multisensory vector is five); Secondly, the architecture of SNN is designed according to different assumptions. We choose the IM paradigm if we assume that multiple senses are independent of each other before fusion, and we choose the AM paradigm if we assume that multiple senses are associated with each other; Thirdly, we specify the neuron model in SNN and sequentially feed concepts to the network, with STDP rules adjusting the network's connection weights. Given the running interval $[0, T]$, we record the spike trains of each neuron; Finally, we convert the spike trains of specific neurons into binarycode as the final integrated representation. The framework is described in detail with the IM and AM paradigms individually in the following sections.

3.1. The Framework

3.1.1. Independent Merge

The IM paradigm is founded on the commonly used cognitive psychology assumption that information for each modality of the concept is independent before integration. It's a two-layer spiking neural network model, with five neurons corresponding to the stimuli of the concept's five separate modal information in the second layer, and a neuron reflecting the neural state after multisensory integration in the second layer. We record the spiking train of the postsynaptic neuron and transform them into integrated vectors for the concept.

For each concept, we get its representation from human-labeled vectors, $\vec{m} = [m_A, m_G, m_H, m_O, m_V]$. The subscripts here represent the concept's perceptual strength as indicated by auditory, gustatory, haptic, olfactory, and visual senses. We min-max normalize the multisensory representation of the concept in the dataset as input to the model during the data preparation stage such that each value of the vector is between 0 and 1. In LC823, for instance, the vector for the concept "honey" is [0.13, 0.95, 0.57, 0.75, 0.80]. We employ LIF or Izhikevich as presynaptic neural models and IF as postsynaptic neural models independently for the generality of the framework. Initially, for each presynaptic neuron, we regard the current $I = m_i * I_{boost}$ as the stimuli to the neuron *where* $i \in [A, G, H, O, V]$ The the conductance g_e of the postsynaptic neuron is updated whenever the presynaptic neuron fires as $g_e \leftarrow g_e + \Delta W_{ij}$, and the postsynaptic neuron generates spikes based on the IF model. The synaptic strength between the postsynaptic neuron and the presynaptic neuron is referred to as the weight ΔW_{ij} in this case. The initial weights between presynaptic and postsynaptic neurons $W_0^i = \frac{g_i}{\Sigma_i^n g_i}$ where $g_i = \frac{1}{\sigma_i^2}$,$\sigma_i^2$ represents the variance for each kind of multisensory data. They are calculated using the Bayesian formula and the assumption that each modal is independent before to fusion (details in the Appendix). At the same time, the spike trains of presynaptic and postsynaptic neurons will dynamically adjust to the weights in accordance with the STDP law. During $[0, T]$, we record the spike train of the postsynaptic neuron $S^{post}([0, T])$ and transform them into

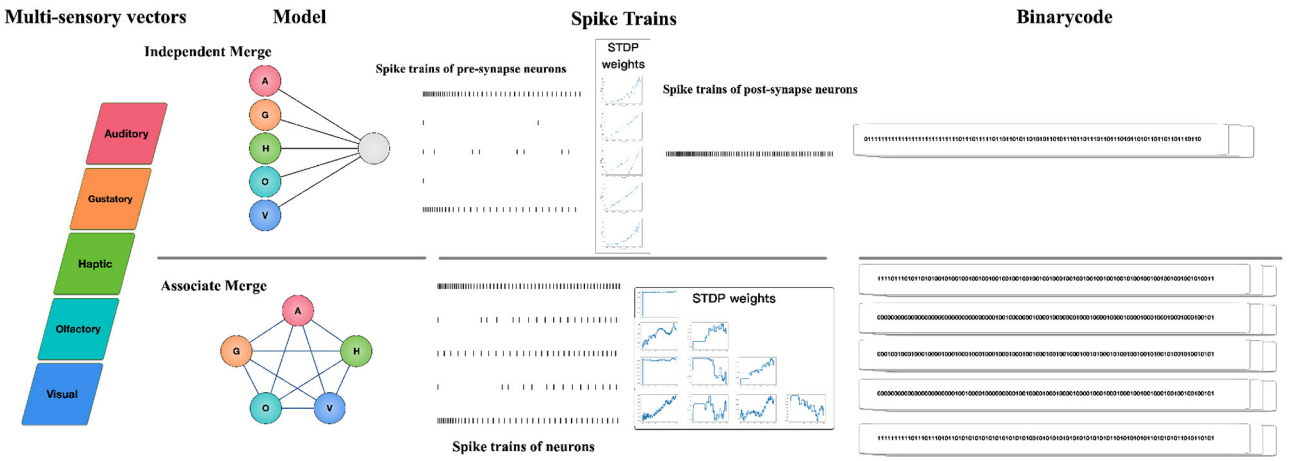

FIGURE 2 | The framework of multisensory concept learning framework based on spiking neural networks. Firstly, for each modality of the concept, we employ a neural model and transform its perceptual strength in the concept's multisensory vector into external stimuli to the neuron; Secondly, the architecture of SNN is designed according to different assumptions. We choose the IM paradigm if we assume that multiple senses are independent of each other before fusion, and we choose the AM paradigm if we assume that multiple senses are associated with each other; Thirdly, we specify the neuron model in SNN and sequentially feed concepts to the network, with STDP rules adjusting the network's connection weights. Given the running interval $[0, T]$, we record the spike trains of each neuron; Finally, we convert the spike trains of specific neurons into binarycode as the final integrated representation.

binarycode $B^{post}([0, T])$, as the final integration representation for the concept in the following manner:

$$B^{post}([0,T]) = [\mathcal{T}(S^{post}((0,tol])), \mathcal{T}(S^{post}((tol, 2*tol])), \cdots, \mathcal{T}(S^{post}(((k-1)*tol, k*tol])), \cdots, \mathcal{T}(S^{post}((\lfloor\frac{T}{tol}\rfloor * tol, T]))] \tag{8}$$

Here $\mathcal{T}(interval)$ operation means that if there is any spikes in the interval, then the bit is 1, otherwise it is 0.

3.1.2. Associate Merge

The AM paradigm assumes that the information for each modality of the concept is associate before integration. It's a five-neuron spiking neural network model, with five neurons corresponding to the stimuli of the concept's five separate modal information. They are connected to one another, and there are no self-connections. We record the spiking trains of all neurons and transform them into integrated vectors for the concept.

We use LIF or Izhikevich neural models to model each neuron for the generality of the framework. For each concept, we get its normalized representation from human-labeled vectors, $\vec{m} = [m_A, m_G, m_H, m_O, m_V]$. Initially, for each neuron $i \in [A, G, H, O, V]$, we consider $I = m_i * I_{boost}$ as the stimuli. The the current I of the postsynaptic neuron is updated whenever the presynaptic neuron fires as $I \leftarrow I + \Delta W_{ij}$. And the postsynaptic neuron generates spikes based on the its model. The weight W_{ij} is the synaptic strength between the presynaptic neuron and the postsynaptic neuron. The initial value for the weight is determined by the correlation each modality pair overall the representation dataset, i.e., $W_0 = Corr(i, j)$ where $i, j \in [A, G, H, O, V]$, which is different from AM paradigm. Simultaneously, presynaptic and postsynaptic neurons' spike trains will dynamically change to the weights in accordance with the STDP law. We denote $S^i([0, T])$ as the *ith* neuron's spike trains during $[0, T]$ and corresponding binary vector $B^i([0, T])$. And we record the spike trains of all neurons, transform them into binarycode $B^i([0, T])$ and concatenate them as the final integration vector $B([0, T])$ in the following way:

$$B^i([0,T]) = [\mathcal{T}(S^i((0,tol])), \mathcal{T}(S^i((tol, 2*tol])), \cdots, \mathcal{T}(S^i(((k-1)*tol, k*tol])), \cdots, \mathcal{T}(S^i((\lfloor\frac{T}{tol}\rfloor * tol, T]))] \tag{9}$$

$$B([0,T]) = [B^A([0,T]) \oplus B^H([0,T]) \oplus B^G([0,T]) \oplus B^O([0,T]) \oplus B^V([0,T])] \tag{10}$$

4. EXPERIMENTS

4.1. Concept Similarity Test

Concept similarity test is commonly used in the field of artificial intelligence to evaluate the effectiveness of system-generated representations (Agirre et al., 2009). Generally, humans score the similarity of a particular concept pair, while the concept pair corresponds to the system-generated representation to calculate the similarity score. After the two scores are ranked in the measure dataset, the Spearman's correlation coefficient is calculated to reflect how close the system-generated representations are to humans. In this article, we evaluate the closeness of the concepts' original or multisensory integration representations and human beings with WordSim353 (Agirre et al., 2009) and SCWS1994 (Huang et al., 2012).

4.1.1. The Experiment

To thoroughly test our framework, we did experiments for IM and AM paradigms with three multisensory datasets

(BBSR, LC823, Lancaster40k) respectively and analyzed the effectiveness differences between the representations after SNN integration and the original representations. In the experiments, both IM and AM paradigms involve a unique parameter in the process of conversion from spike trains to binarycode: the tolerance *tol*. It represents the size of the reducing window for converting spike trains in the time interval into binarycode, which reflects the strength of compressing the spike sequence into a integrated binarycode. In each dimension of the integrated vector, a larger *tol* signifies a higher degree of information compression and a bigger reducing window, and *vice versa*. But, if *tol* is too small, the representation vector's dimensionality will be too large, and if *tol* is too big, the diversity of all representations will be damaged. Therefore, we traverse *tol* across the range $[0, 500]$ while restricting diversity to the range $[0.05, 0.95]$, and the results indicate the present model's ideal results as well as the matching *tol*.

TABLE 2 | Concept similarity test results.

Merge way	Model	BBSR				LC823an				Lancaster40k			
		Tol	WordSim353	SCWS1994	Average	Tol	WordSim353	SCWS1994	Average	Tol	WordSim353	SCWS1994	Average
Origin	–	–	0.4182	0.5838	0.5010	–	0.1321	0.5525	0.3423	–	0.2640	**0.3974**	0.3534
AM	lzh-RZ	93	0.3455	0.6089	0.4772	165	0.3804	0.4260	0.4032	9	0.3560	0.3295	0.3427
	lzh-FS	95	0.4955	0.5659	0.5307	312	0.4223	0.3788	0.4006	9	0.3787	0.3471	0.3629
	lzh-IB	384	0.5455	0.5870	0.5662	32	0.3696	0.5277	0.4486	25	0.3388	0.3818	0.3603
	lzh-LTS	174	0.5068	0.6127	0.5598	17	0.3107	0.5390	0.4249	16	0.3557	0.3629	0.3593
	lzh-RS	366	0.4955	0.5857	0.5406	84	0.5179	0.5271	0.5225	55	0.3206	0.3708	0.3457
	lzh-CH	170	0.4273	0.5928	0.5100	7	0.1089	0.4884	0.2986	14	0.3150	0.3349	0.3249
	lzh-TC	148	0.5068	0.6103	0.5586	6	0.2214	0.5181	0.3698	7	**0.3979**	0.3364	0.3672
	LIF	187	0.5727	0.6927	0.6327	330	0.5036	**0.6330**	0.5683	86	0.1788	0.3500	0.2644
IM	lzh-RZ	17	0.4636	0.634	0.5488	10	0.5545	0.5618	0.5581	4	0.2026	0.3139	0.2583
	lzh-FS	17	0.4636	0.6388	0.5512	10	0.5545	0.5617	0.5581	21	0.3371	0.2910	0.3140
	lzh-IB	7	**0.5477**	0.5988	**0.5733**	24	0.5509	0.5491	0.5500	31	0.1597	0.3040	0.2319
	lzh-LTS	83	0.5000	**0.6417**	0.5708	18	**0.6080**	0.5361	**0.5721**	56	0.3610	0.3448	0.3529
	lzh-RS	196	0.5023	0.5530	0.5276	163	0.4830	0.4613	0.4722	68	0.0757	0.2959	0.1858
	lzh-CH	94	0.4659	0.5786	0.5222	8	0.5696	0.4746	0.5221	50	0.3843	0.3813	**0.3828**
	lzh-TC	17	0.4636	0.6125	0.5381	5	0.4509	0.5310	0.4909	20	0.3387	0.3042	0.3215
	LIF	143	0.4205	0.6167	0.5186	3	0.0643	0.5672	0.3158	324	0.0018	0.1481	0.1965

The bold values indicates the current measure dataset reflect the best results, while the underlined values imply that the multisensory integrated representation is closer to humans than the original representation.

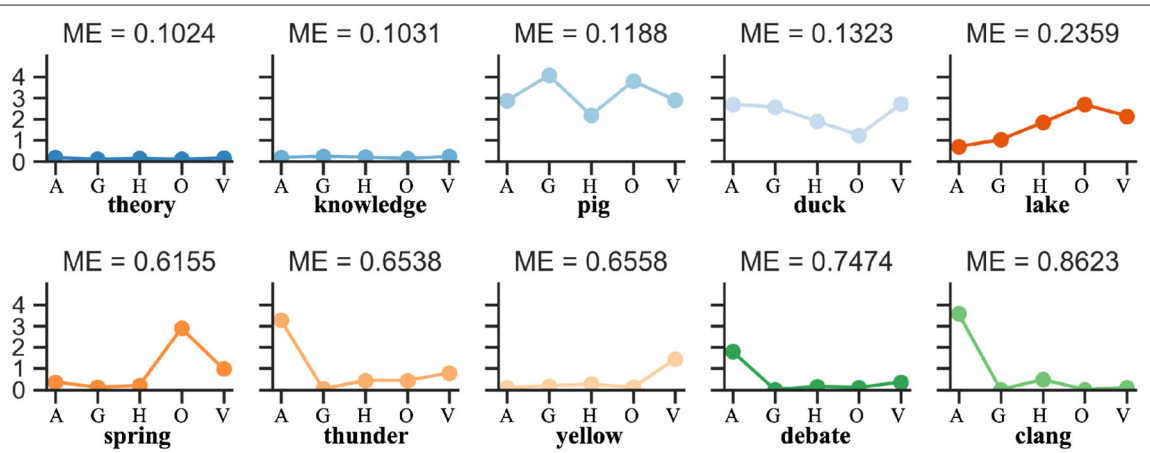

FIGURE 3 | Modality exclusivity demonstration. Modality exclusivity (ME) is a metric measuring how much of a concept is perceived through a single perceptual modality. For each concept, the value of ME is calculated as the perceptual strength range divided by the sum, and spanning from 0 to 100% for completely multimodal to completely unimodal perception.

We used the evaluation datasets WordSim353 and SCWS1994 for testing, and the inputs of the models were from different sources of multisensory representation datasets: BBSR, LC823an, Lancaster40k, and tested using two paradigms, IM and AM, respectively. For the AM paradigm, Izhikevich's seven models and LIF model were used, while for the IM paradigm, IF model were used for postsynaptic neurons and Izhikevich's seven models and LIF model were used for presynaptic neurons. The running time of all the tests is 100 ms and $I_{boost} = 100$.

4.1.2. Results and Analysis

From the overall results for both IM and AM paradigms, the integrated vectors are closer to humans than the original vectors based on our models: 37 submodels achieved better results for a total of 48 tests for both IM and AM, as **Table 2** shows. In terms of overall dataset, 15/16 tests work better for the BBSR dataset, 14/16 tests work better for LC823an, and 8/16 tests work better for Lancaster40k.

In almost all experiments, multisensory integrated representations based on our framework outperform unintegrated ones, with the exception of the instability shown in IM and AM paradigms when Lancaster40k is employed as the input. For any of the multisensory vectors, an integration way could be found to improve their representations.

4.2. Comparisons Between IM and AM Paradigms

Unlike the analysis of the macro-level above, in this section we introduce the concept feature norms to compare IM and AM paradigms from the micro-level perspective of each concept. Concept feature norms are a way of representing concepts by using standardized and systematic feature descriptions that mirror human comprehension. The similarities and differences of concepts are related to the intersection and difference of concept feature norms. McRae's concept feature norms, introduced by McRae et al. (2005), are the most prominent work in this area. They not only supplied 541 concepts with feature norms, but also proposed a methodology for generating them. For example, the feature norms of the concept "basement" are *"used for storage," "found below ground," "is cold," "found on the bottom floor," "is dark," "is damp," "made of cement," "part of a house," "has windows," "has a furnace," "has a foundation," "has stairways," "has walls," "is musty," "is scary," and "is the lowest floor."* Another semantic feature norms dataset analogous to McRae is CSLB (Centre for Speech, Language, and the Brain). They collected 866 concepts and improved the feature normalization and feature filtering procedure (Devereux et al., 2014). The McRae and CSLB criteria for human conceptual cognition are used in this research to investigate how each concept is similar to human cognition.

We compare and analyze IM and AM paradigms from two perspectives. First, we use the perceptual strength-related metric Modality Exclusivity to compare the two paradigms to explore the sensitive of them to the concepts' strength distribution of multisensory information. Then, to assess the generality of the IM and AM paradigms, we introduce nine psycholinguistic dimensions derived from the concept's nature, which are unrelated to perceptual strength.

TABLE 3 | The sensibility of IM and AM results to modality exclusivity.

Izhkevich model	AM		IM	
	McRae	CSLB	McRae	CSLB
RZ	0.0149	−0.0987	−0.1524	−0.4848
FS	0.2679	0.0901	−0.134	−0.4447
IB	−0.0559	0.0191	−0.2672	−0.4986
LTS	0.2113	0.035	−0.12	−0.0453
RS	0.1943	−0.0087	−0.006	−0.1997
CH	0.0988	0.0197	0.0294	0.0964
TC	0.2078	0.0398	−0.2115	−0.4761

4.2.1. Modality Exclusivity

Modality Exclusivity (ME) is a metric measuring how much of a concept is perceived through a single perceptual modality (Lynott and Connell, 2013). For each concept, the value of ME is calculated as the perceptual strength range divided by the sum, and spanning from 0 to 100% for completely multimodal to completely unimodal perception. **Figure 3** show some examples.

In the concept feature norms dataset, we first obtain all similar concepts $c^{similar}$ for each concept c based on the number of feature overlaps and record their rank list $R_c^{similar}$ sorted by similarity. Then, for all concepts, the corresponding multisensory integrated binary representations B^{IM} and B^{AM} are produced using the IM and AM paradigms, respectively. Next, for concept c, its k similar concepts $c_{IM}^{k\ similar}$ and $c_{AM}^{k\ similar}$ are computed based on integrated binarycodes and harming distance, respectively. We query the rank of these k similar concepts in the feature norms space $R_c^{similar}$ and take the average value, denoted as $kAR_{c_{IM}}$ and $kAR_{c_{AM}}$, which reflects the closeness of the multisensory representations to human cognition using two ways of integration in our framework. Smaller values of kAR indicate closer to human cognition at the microscopic level. Finally, we focus on all concepts in the representation dataset and calculate the correlation coefficients between the $kAR_{c_{IM}}$ or $kAR_{c_{AM}}$ arrays obtained using the above approach and the ME arrays corresponding to the concepts. This coefficient reflects the correlation between the two different multisensory concept integration paradigms and modal exclusivity. And in this experiment we only test the Izhikevich model and set k to 5.

The results in **Table 3** reveal the difference between IM and AM paradigms. The IM paradigm has a stronger negative correlation in both concept feature norms test sets, but the AM paradigms has a slightly positive correlation. We investigate this discrepancy further by viewing the FS model in detail, as shown in **Figure 4**. The results reveal that for concepts

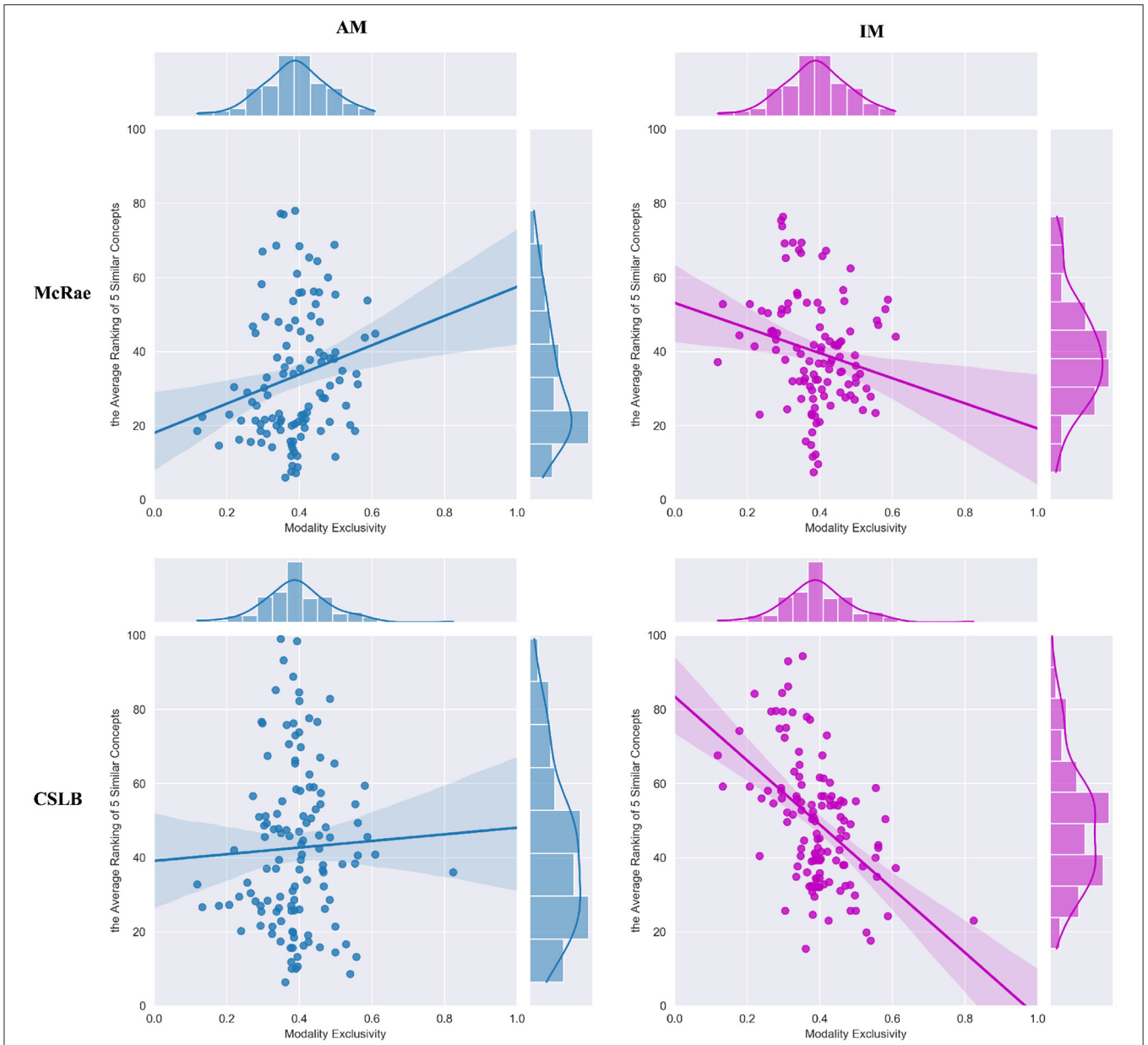

FIGURE 4 | The correlation between ME and average of five similar concept rankings.

with higher ME (such as "spring," "thunder," "yellow," "debate," "clang" in **Figure 3**), the IM paradigm is better at multisensory integration. While the AM paradigm is less input biased for each modality, it benefits the concept of uniform modal distribution (such as "theory," "knowledge," "pig," "duck," "lake" in **Figure 3**).

4.2.2. Generality Analysis

The ME metric used in the previous experiments is a perceptual strength-related indicator for the concept representation. In this part, we will test the framework from the input concept itself. And we introduce Glasgow norms which are a set of normative assessments on nine psycholinguistic dimensions: arousal (AROU), valence (VAL), dominance (DOM), concreteness (CNC), imageability (IMAG), familiarity (FAM), age of acquisition (AOA), semantic size (SIZE), and gender association (GEND) for 5,553 concepts (Scott et al., 2019).

In the same manner as the previous experiment. In concept feature norms, we first record all similar concepts for each concept, then sort them by similarity and rank them. Then, for IM and AM paradigms, we use the same concept input, get the integration vector for each concept, find their k similar, and get the mean value of their ranking in concept feature norms as $kAR_{c_{IM}}$ and $kAR_{c_{AM}}$. Finally, we determine the correlation

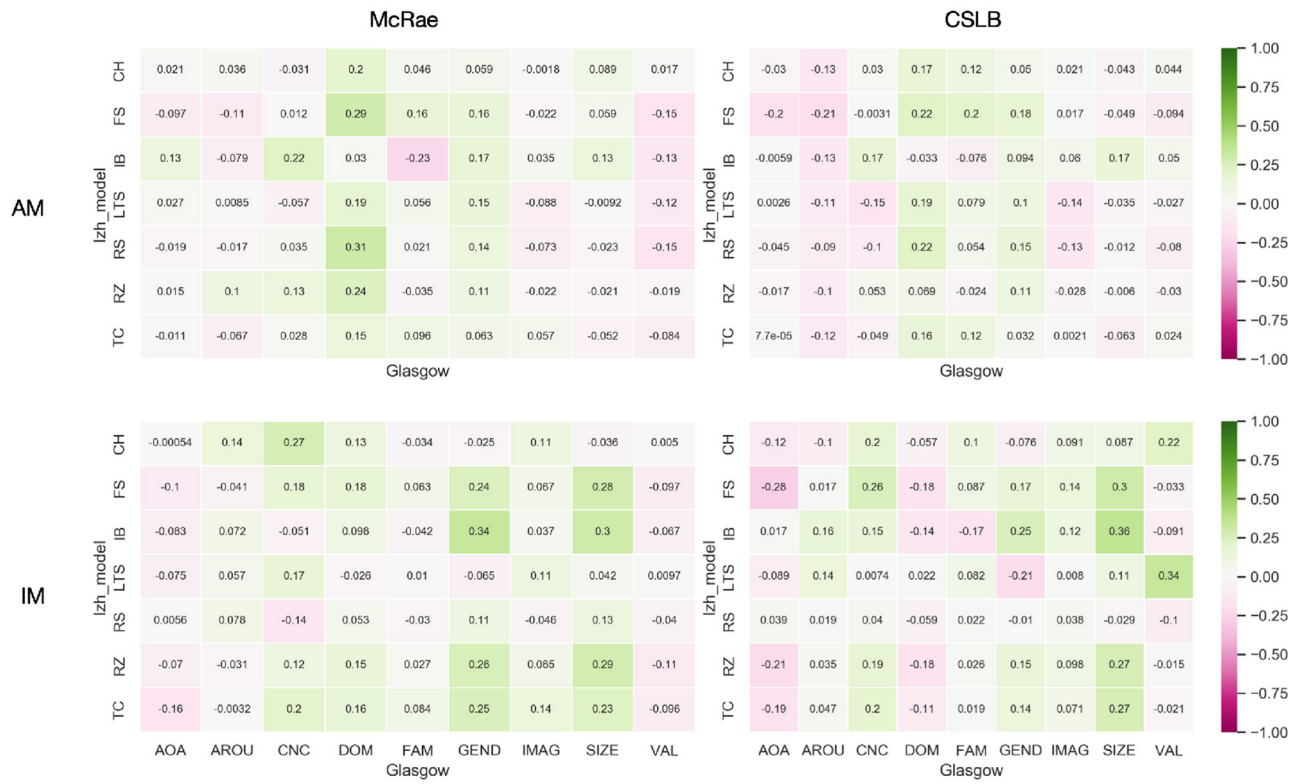

FIGURE 5 | The heatmap of generality analysis results.

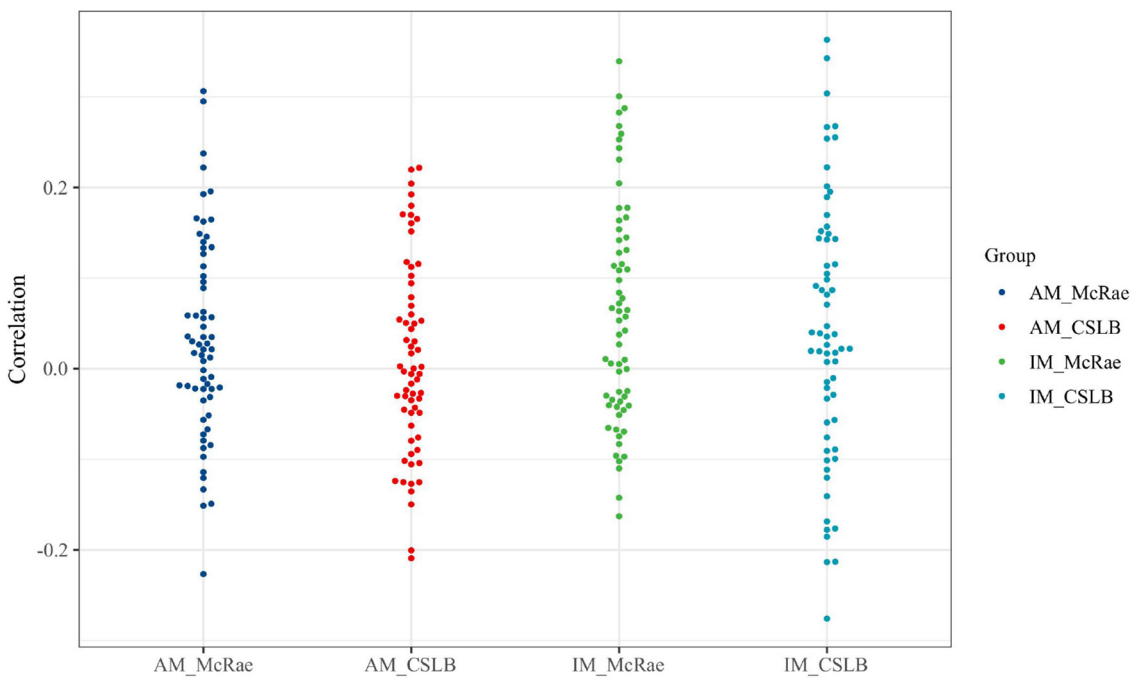

FIGURE 6 | The beeswarm of correlation distribution.

coefficient between each psychological characteristic and the concept's average ranking value *kAR* for the two paradigms. We still only test the Izhikevich model in this experiment, and the value is set to 5.

We used heatmaps (**Figure 5**) to visualize the correlation coefficients between the IM and AM paradigms' *kAR* and nine psycholinguistics in the two concept feature norms sets McRae and CSLB. Additionally, we omit the adopted Izhikevich submodels and provide the correlation coefficients using a beeswarm (**Figure 6**) to explain them more clearly.

According to the experimental results presented, the absolute values of all correlation coefficients are <0.3. The effect of vectors after integration of either IM or AM paradigms does not have any relationship with the nature of the concepts for several dimensions, including AOA, AROU, FAM, IMAG, and VAL. This indicates that both paradigms have good generality and the framework is not affected by the concepts themselves.

5. DISCUSSION

In this study, we propose a SNN-based concept learning framework for multisensory integration that can generate integration vectors based on psychologist-labeled multimodal representations. Vision, hearing, touch, smell, and taste are among the five modalities used in our research, which also includes a brain-like SNN model. We intend to add more brain-like processes in the future, such as multisensory fusion plasticity. The multisensory data we currently use are labeled by cognitive psychologists, which is relatively expensive and small, and in the future we consider expanding the relevant dataset by mapping for larger scale experiments. The current research focuses on multisensory representation of concepts, which is a subset of pattern representation in AI, and future research can be deeply integrated with downstream tasks to create AI systems that incorporate multisensory integration. At the same time, this places more demands on multisensory perceptrons. Human perception of concepts has not only multisensory perception but also more textual information based on abstract information, and it is also worth exploring how to combine these two parts to build human-like concept learning systems in the future.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: http://osf.io/7emr6/; http://www.neuro.mcw.edu/resources.html; https://link.springer.com/article/10.3758/BRM.41.2.558; https://link.springer.com/article/10.3758/s13428-012-0267-0.

AUTHOR CONTRIBUTIONS

YW and YZ designed the study, performed the experiments, and wrote the manuscript. Both authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB32070100).

ACKNOWLEDGMENTS

We thank Dr. Yanchao Bi and Dr. Xiaosha Wang for helpful discussions and generous sharing of psychology-related researches.

REFERENCES

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pacsca, M., and Soroa, A. (2009). "A study on similarity and relatedness using distributional and word net-based approaches," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (Boulder, CO: Association for Computational Linguistics), 19–27. Available online at: https://aclanthology.org/N09-1003

Anastasio, T. J., Patton, P. E., and Belkacem-Boussaid, K. (2014). Using Bayes' rule to model multisensory enhancement in the superior colliculus. *Neural Comput.* 12, 1165–1187. doi: 10.1162/089976600300015547

Bi, G.-Q., and Poo, M.-M. (2001). Synaptic modification by correlated activity: Hebb's postulate revisited. *Annu. Rev. Neurosci.* 24, 139–166. doi: 10.1146/annurev.neuro.24.1.139

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., et al. (2016). Toward a brain-based componential semantic representation. *Cogn. Neuropsychol.* 33, 130–174. doi: 10.1080/02643294.2016.1147426

Bruni, E., Tran, N.-K., and Baroni, M. (2014). Multimodal distributional semantics. *J. Artif. Intell. Res.* 49, 1–47. doi: 10.1613/jair.4135

Calvert, G. A., and Thesen, T. (2004). Multisensory integration: methodological approaches and emerging principles in the human brain. *J. Physiol. Paris* 98, 191–205. doi: 10.1016/j.jphysparis.2004.03.018

Cappe, C., Rouiller, E. M., and Barone, P. (2009). Multisensory anatomical pathways. *Hear. Res.* 258, 28–36. doi: 10.1016/j.heares.2009.04.017

Collell, G., Zhang, T., and Moens, M. -F. (2017). "Imagined visual representations as multimodal embeddings," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31 (San Francisco, CA: AAAI).

Devereux, B. J., Tyler, L. K., Geertzen, J., and Randall, B. (2014). The centre for speech, language and the brain (CSLB) concept property norms. *Behav. Res. Methods* 46, 1119–1127. doi: 10.3758/s13428-013-0420-4

Gao, Y., Hendricks, L. A., Kuchenbecker, K. J., and Darrell, T. (2016). Deep learning for tactile understanding from visual and haptic data. *arXiv:1511.06065*. doi: 10.1109/ICRA.2016.7487176

Gepner, R., Skanata, M. M., Bernat, N. M., Kaplow, M., and Gershow, M. (2015). Computations underlying drosophila photo-taxis, odor-taxis, and multi-sensory integration. *eLife* 4:e6229. doi: 10.7554/eLife.06229

Gerstner, W., and Kistler, W. M. (2002). *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge University Press. doi: 10.1017/CBO9780511815706

Hill, F., and Korhonen, A. (2014). "Learning abstract concept embeddings from multi-modal data: since you probably can't see what I mean," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (Cambridge, MA: EMNLP), 255–265. doi: 10.3115/v1/D14-1032

Hill, F., Reichart, R., and Korhonen, A. (2014). Multi-modal models for concrete and abstract concept meaning. *Trans. Assoc. Comput. Linguist.* 2, 285–296. doi: 10.1162/tacl_a_00183

Hodgkin, A. L., and Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* 117, 500–544. doi: 10.1113/jphysiol.1952.sp004764

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012). "Improving word representations via global context and multiple word prototypes," in *Proceedings of the 50th Annual Meeting of the Association for ComputationalLinguistics, Vol. 1* (Jeju Island: Association for Computational Linguistics), 873–882.

Izhikevich, E. M. (2003). Simple model of spiking neurons. *IEEE Trans. Neural Netw.* 14, 1569–1572. doi: 10.1109/TNN.2003.820440

Kiela, D., and Bottou, L. (2014). "Learning image embeddings using convolutional neural networks for improved multi-modal semantics," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (Doha: EMNLP). doi: 10.3115/v1/D14-1005

Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., and Morency, L.-P. (2018). Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*. doi: 10.18653/v1/P18-1209

Lynott, D., and Connell, L. (2009). Modality exclusivity norms for 423 object properties. *Behav. Res. Methods* 41, 558–564. doi: 10.3758/BRM.41.2.558

Lynott, D., and Connell, L. (2013). Modality exclusivity norms for 400 nouns: the relationship between perceptual experience and surface word form. *Behav. Res. Methods* 45, 516–526. doi: 10.3758/s13428-012-0267-0

Lynott, D., Connell, L., Brysbaert, M., Brand, J., and Carney, J. (2019). The Lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 English words. *Behav. Res. Methods* 1–21. doi: 10.31234/osf.io/ktjwp

Maass, W. (1997). Networks of spiking neurons: the third generation of neural network models. *Neural Netw.* 10, 1659–1671. doi: 10.1016/S0893-6080(97)00011-7

McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behav. Res. Methods* 37, 547–559. doi: 10.3758/BF03192726

Parise, C. V., and Ernst, M. O. (2016). Correlation detection as a general mechanism for multisensory integration. *Nat. Commun.* 7:11543. doi: 10.1038/ncomms11543

Rieke, F., Warland, D., Van Steveninck, R. d. R., and Bialek, W. (1999). *Spikes: Exploring the Neural Code*. MIT Press.

Roshan, C., Barker, R. A., Sahakian, B. J., and Robbins, T. W. (2001). Mechanisms of cognitive set flexibility in Parkinson's disease. *Brain A J. Neurol.* 124, 2503–2512. doi: 10.1093/brain/124.12.2503

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., and Sereno, S. C. (2019). The glasgow norms: ratings of 5,500 words on nine scales. *Behav. Res. Methods* 51, 1258–1270. doi: 10.3758/s13428-018-1099-3

Shams, L., and Seitz, A. R. (2008). Benefits of multisensory learning. *Trends Cogn.* 12, 411–417. doi: 10.1016/j.tics.2008.07.006

Silberer, C., Ferrari, V., and Lapata, M. (2013). "Models of semantic representation with visual attributes," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Vol. 1* (Sofia: Association for Computational Linguistics), 572–582.

Silberer, C., and Lapata, M. (2014). "Learning grounded meaning representations with autoencoders," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Vol. 1* (Baltimore: Association for Computational Linguistics), 721–732. doi: 10.3115/v1/P14-1068

Stimberg, M., Brette, R., and Goodman, D. F. (2019). Brian 2, an intuitive and efficient neural simulator. *Elife* 8:e47314. doi: 10.7554/eLife.47314

Troyer, T. W., and Miller, K. D. (1997). Physiological gain leads to high isi variability in a simple model of a cortical regular spiking cell. *Neural Comput.* 9, 971–983. doi: 10.1162/neco.1997.9.5.971

Ursino, M., Cuppini, C., and Magosso, E. (2014). Neurocomputational approaches to modelling multisensory integration in the brain: a review. *Neural Netw.* 60, 141–165. doi: 10.1016/j.neunet.2014.08.003

Ursino, M., Cuppini, C., Magosso, E., Serino, A., and Pellegrino, G. D. (2009). Multisensory integration in the superior colliculus: a neural network model. *J. Comput. Neurosci.* 26, 55–73. doi: 10.1007/s10827-008-0096-4

Verma, S., Wang, C., Zhu, L., and Liu, W. (2019). "Deepcu: Integrating both common and unique latent information for multimodal sentiment analysis," in *International Joint Conference on Artificial Intelligence* (Macao). doi: 10.24963/ijcai.2019/503

Wang, S., Zhang, J., and Zong, C. (2018a). "Associative multichannel autoencoder for multimodal word representation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels), 115–124. doi: 10.18653/v1/D18-1011

Wang, S., Zhang, J., and Zong, C. (2018b). "Learning multimodal word representation via dynamic fusion methods," in *Thirty-Second AAAI Conference on Artificial Intelligence* (New Orelans, LA).

Wang, X., Men, W., Gao, J., Caramazza, A., and Bi, Y. (2020). Two forms of knowledge representations in the human brain. *Neuron* 107, 383–393.e5. doi: 10.1016/j.neuron.2020.04.010

Xu, Y., Yong, H., and Bi, Y. (2017). A tri-network model of human semantic processing. *Front. Psychol.* 8:1538. doi: 10.3389/fpsyg.2017.01538

Zadeh, A., Chen, M., Poria, S., Cambria, E., and Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*. doi: 10.18653/v1/D17-1115

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

APPENDIX

The Initial Weights in IM

Similar to what cognitive psychologists (Ursino et al., 2014) have done before, we assume that for the concept s and its each modality $i \in [A, G, H, O, V]$ representations, $p(x_i|s) \sim N(x_i; s, \sigma_i)$, where $N(x; \mu, \sigma)$ stands for the normal distribution over x with mean μ and standard deviation σ. They are conditionally independent from each other and by Bayes' rule,

$$\begin{aligned} p(s|x_A, x_G, x_H, x_O, x_V) &\propto p(x_A, x_G, x_H, x_O, x_V|s) \\ &\propto \prod_i p(x_i|s) = \frac{1}{\prod_i(\sqrt{2\pi}\sigma_i)} e^{-\Sigma_i \frac{(x_i-s)^2}{2\sigma_i^2}} \\ &\propto -\Sigma_i \frac{(x_i-s)^2}{2\sigma_i^2} \end{aligned} \tag{11}$$

The maximum-a-posteriori estimation for s is $\widehat{s} = \Sigma_i \frac{\frac{1}{\sigma_i^2}}{\Sigma_i \frac{1}{\sigma_i^2}} x_i$, where $\frac{1}{\sigma_i^2}$ reflects the reliability of each modality for the same concept s. In our IM schema, we regard normalized reliability as the initial weights between pre-synaptic neurons (describing each modality) and the post-synaptic neuron(for integration), i.e.,

$$w_i^0 = \frac{\frac{1}{\sigma_i^2}}{\Sigma_i \frac{1}{\sigma_i^2}} \tag{12}$$

where we can get each σ_i *via* psychologist-labeled multisensory datasets.