A Brain-inspired Hierarchical Interactive In-memory Computing System and its Application in Video Sentiment Analysis

Xiaoyue Ji, *Student Member, IEEE*, Zhekang Dong, *Senior Member, IEEE*, Yifeng Han, Chun Sing Lai, *Senior Member, IEEE*, Donglian Qi, *Senior Member, IEEE*

***Abstract*—Video sentiment analysis can effectively establish the relationship between the emotion state and the multimodal information, while still suffer from intensive computation and low efficiency, due to the von Neumann computing architecture. Here, we present a brain-inspired hierarchical interactive in-memory computing (IMC) system, which can efficiently solve 'von Neumann bottleneck', enabling cross-modal interactions and semantic gap elimination. First, a 1T1M synapse array is fabricated using cost-effective, highly stable, flexible, and eco-friendly carbon materials, offering efficient analog multiply-accumulate operations. To illustrate the complexity of the proposed brain-inspired hierarchical interactive IMC system, three modules are proposed: 1) unimodal extraction module, 2) hierarchical interactive module, 3) output module. Furthermore, the proposed system is validated by applying it to video sentiment analysis. The experimental results demonstrate that the proposed system outperforms the existing state-of-the-art methods with high computational efficiency and good robustness. This work opens up a new way to achieve the deep integration of nanomaterials, deep learning, and modern electronics into IMC.**

***Index Terms*—Brain-inspired, hierarchical interactive, in-memory computing (IMC), sentiment analysis**

I. Introduction

With the fast-paced development of video acquisition technology, the amount of video data is increasing rapidly over the recent years [1]. Video has been an important information carrier, providing a new path for daily communication and socialization [2]. Emotion information in video is closely related to the human health and well-being. As a result, video sentiment analysis is now attracting increasing interest [3].

Recent researches of video sentiment analysis are mainly based on machine learning methods and deep learning methods, which have achieved state-of-the-art performance in terms of recognition accuracy [4-6]. However, these methods also have certain limitations in providing real-time information processing with relatively low energy consumption, making them unable to compete with biological neural systems especially in computational efficiency. This is because they are mostly built on von Neumann computing system with separate processing and memory units (called the 'von Neumann bottleneck') [7]. Meanwhile, these traditional systems usually rely on the complementary metal oxide semiconductors (CMOS) integrated circuit design inevitably limited by 'memory wall', leading to low efficiency and instability, particularly when in situ learning and biological interpretability are required [8-10].

In-memory computing (IMC) has recently started to revolutionize conventional methods because of its capacity to continue running computations and caching big data [11]. IMC systems are built on numerous upcoming nanotechnologies and beyond-CMOS devices, making lower power consumption and higher speed possible [12]. Memristor is a two-terminal electronic device that provides functional relationship between charge and flux. It was invented by L. O. Chua in 1971 [13] and was applied to physical devices by researchers in Hewlett Packard Labs [14]. It is a perfect choice for synapses in IMC systems because of its inherent dynamics, analog behaviors, non-volatility, high-speed, low-energy, and high-density features [15]. Memristive synapse arrays exhibit high levels of parallel computing capability, enabling multiply-accumulate (MAC) operations, the major calculation method in information processing. By using analog IMC, significant energy and time overheads incurred by data shuttling in von Neumann system can be avoided [16].

IMC systems are equipped with machine learning algorithms and the ability to perceive and process video data in different modalities, such as visual [17], tactile [18], auditory [19], olfactory [20], and gustatory data [21], drawing lessons from human sensory processing and perceptual learning mechanisms. However, the current IMC systems still have some limitations. At the device level, the cycle-to-cycle (C2C) and device-to-device (D2D) variations may cause inaccurate

Manuscript received December 31, 2022.

This work was supported in part by the National Natural Science Foundation of China under Grant 62001149, Natural Science Foundation of Zhejiang Province under Grant LQ21F010009, and Fundamental Research Funds for the Provincial University of Zhejiang under Grant GK229909299001-06. *(Corresponding authors: Zhekang Dong).*

X. Ji, Y. Han, and D. Qi are with the College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: ji.xiaoyue@zju.edu.cn; hanyf@zju.edu.cn; qidl@zju.edu.cn).

Z. Dong is with the School of Electronics and Information, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: englishp@126.com).

C. S. Lai is with the Department of Electronic and Computer Engineering, Brunel University London, London, UB8 3PH, UK and also with the School of Automation, Guangdong University of Technology, Guangzhou, China 510006 (email: chunsing.lai@brunel.ac.uk).

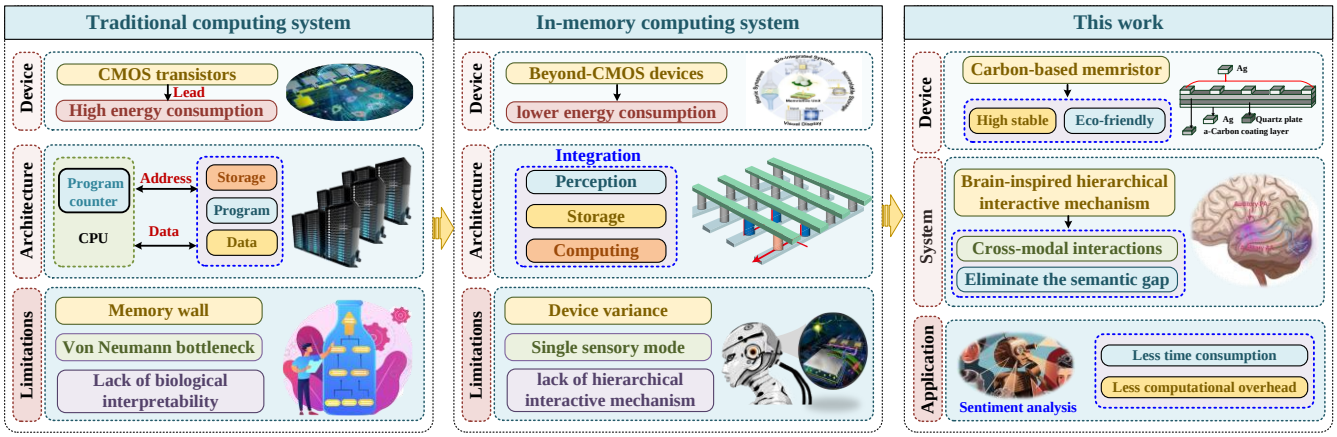


Fig. 1. The systemic comparison of different computing systems for video analysis.

encoding of network weights in IMC because of the non-uniformity of the switching function layers and electrodes [22]. Therefore, computing devices that are more reliable and environment-friendly are required. At the system level, almost all existing IMC systems concentrate on a single mode of sensory processing. A lack of hierarchical interactive systems that combine numerous senses based on cross-modal learning mechanisms make information across various sensory modalities hard to handle.

Thus, this work aims to investigate a brain-inspired hierarchical interactive IMC system. For verification purposes, the proposed system was applied to video sentiment analysis. For clarity, the systemic comparison of different computing systems for video analysis is provided in Fig. 1. The main contributions of this work are summarized as follows.

1) Different with existing IMC systems, we present a brain-inspired hierarchical interactive IMC system that can effectively capture cross-modal interactions and eliminate the semantic gap between multi-modal signals.

2) The circuit design of the entire system is developed after the fabrication of a cost-effective, highly stable, and eco-friendly carbon-based memristor, enabling a parallel-computed and highly integrated IMC system.

3) A hardware implementation of video sentiment analysis is developed, which provides the benefits of less computational overhead and time consumption, to solve computationally difficult problems with unattainable energy efficiencies for von Neumann architectures.

The rest of this paper is structured as follows. Section II presents the hierarchical architecture of the proposed brain-inspired interactive IMC system. Section III describes the fabrication of the carbon-based memristor and memristive synapse array. Section IV demonstrates the detailed circuit design of the entire system from the perspectives of unimodal extraction, hierarchical interaction, and output modules. In Section V, the proposed system is applied to sentiment analysis for verification. Finally, Section VI includes the conclusion drawn from the study.

II. Brain-inspired Hierarchical Interactive Computing System Architecture

In human brain, sensing, transmitting, and processing of information relies on distributed and hierarchical neural networks consisting of receptors, nerve pathways, and cerebral cortex [23], which are compact and efficient for solving complex and unstructured real-world problems, as shown in Fig. 2(a). Specifically, sensory receptors (the retina and cochlea) convert environmental inputs into spike trains in the cells. The nerve pathways then carry spike trains from the receptors to the brain's cerebral cortex, where information is further processed. Notably, the biological intelligence of the brain results from synapses connections, among nearly a hundred billion neurons, which enable the brain to perform intelligent tasks with high performance [24]. Therefore, we develop a brain-inspired hierarchical interactive computing system that can synthesize and process multimodal information, as shown in Fig. 2(b). Specifically, memristive synapse array is used to perform efficient MAC operation, which acts as the synapse in brain-inspired computing architectures. Unimodal extraction module and hierarchical interaction module are used to simulate the sensory processing and perceptual learning mechanism of human brain. Output module is used to simulate the cognitive analysis mechanism in the cerebral cortex.

Memristive synapse array: The huge parameters and complex calculations of brain inspired computing system necessitate basic components possess 'big data' transmission capabilities and enormous processing units. Memristors with unique qualities, such as high storage density, low power consumption, fast programming and erasing speed, have been proposed as promising artificial synaptic candidates for realizing brain-inspired computing systems. Due to the exceptional physical properties of carbon materials, including electrical tunability, sustainability, and eco-environment, carbon materials have attracted much attention from academia and industry [25]. Therefore, a carbon material-based memristor is fabricated using frame method and magnetron sputtering method. Meanwhile, a parallel connection is made by constructing a transistor (1T) and an Ag/a-Carbon/Ag memristor (1M) synapse array, lowering the system complexity and cost.

Unimodal extraction module: Different sensory perceptions are responsible for receiving and extracting various stimulus inputs, with visual and auditory perceptions being the two main types [26]. Moreover, it has been estimated that these pathways

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2023.3275708

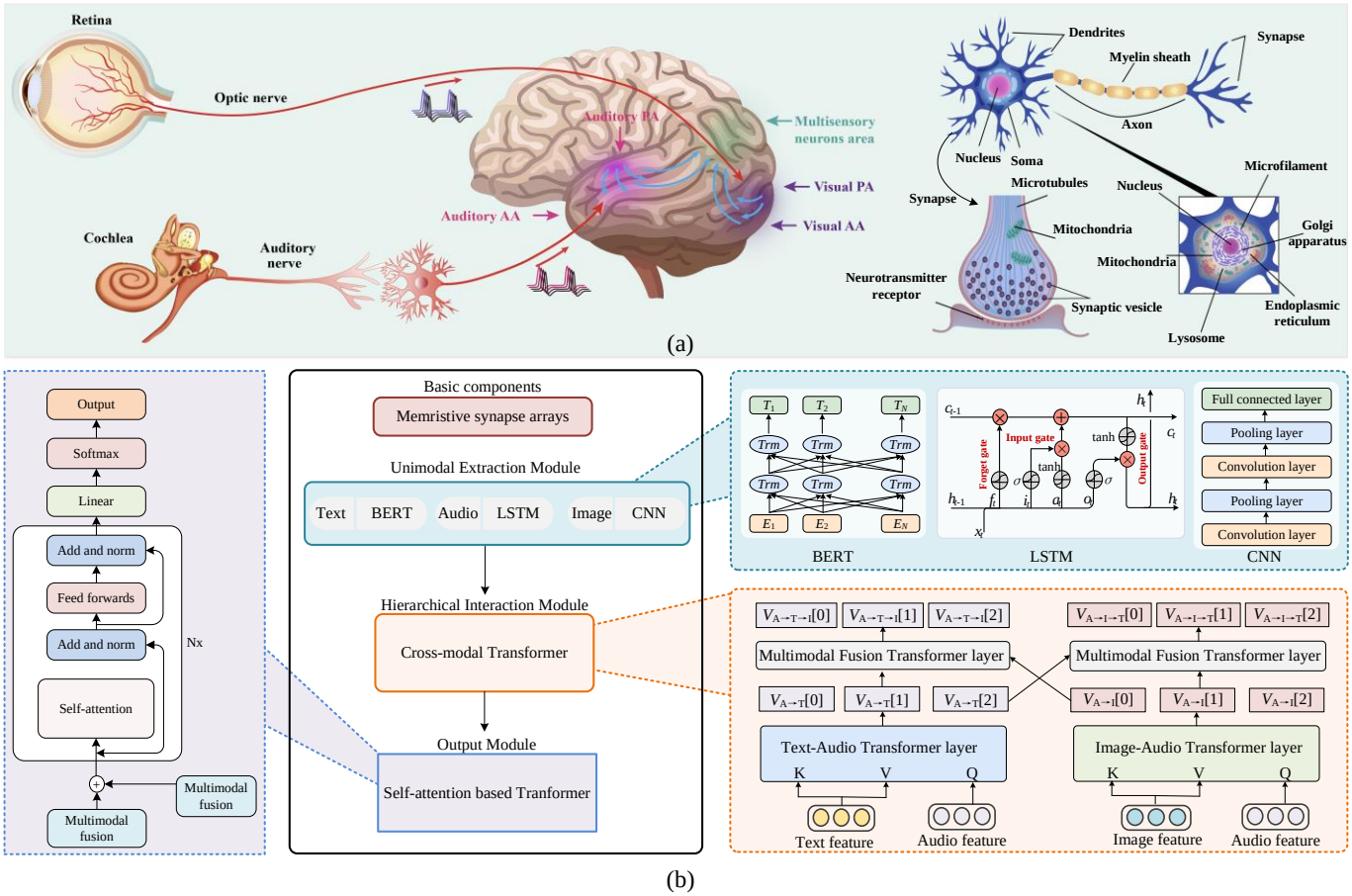


Fig. 2. (a) Schematic of information sensing, transmitting, and processing in human brain; (b) Schematic of brain-inspired hierarchical interactive IMC system.

achieve over 90% of the processed information. We propose a unimodal extraction module that fully extracts features from the input text, audio, and images to simulate the signal extraction process of sensory perceptions. Specifically, to acquire the underlying semantic and contextual meanings of the text, bidirectional encoder representations from transformers (BERT) [27] are employed for textual feature extraction, denoted by V_{FT}=BERT(v_T). To extract features that characterize the complex nature of the audio, the long short-term memory (LSTM) network [28] is adopted, which is presented as V_{FA}=LSTM(v_A). While considering the hidden representation of images, we chose a convolutional neural network (CNN) [29] for object-level visual features, which is indicated by V_{FI}=CNN(v_I).

Hierarchical interaction module: Considering human brain information interaction mechanism [23], we propose a hierarchical interaction module containing two levels of interactions: (1) low-level interactions between audio and text (or image), and (2) high-level interactions between text and image. The cross-modal transformer layer is used for the audio-text feature representation ($V_{A\to T}$) and the audio-image feature representation ($V_{A\to I}$) to model these two types of interactions. In particular, as illustrated in the right portion of Fig. 2(b), a multi-head cross-attention [30] is employed, taking audio feature representation V_{FA} as queries and text feature representation V_{FT} (or image feature representation V_{FI}) as keys and values, and then followed by a feed-forward network. Furthermore, to understand the intermodal interaction between text and image feature representations, we propose a multimodal fusion transformer above the cross-modal transformer, where $V_{A\to T}$ is treated as queries and $V_{A\to I}$ as keys and values. Moreover, $V_{A\to T\to I}$ stands for the final multimodal-fused image feature representation. Similarly, $V_{A\to I\to T}$ indicates the multimodal-fused text feature representation obtained with the same structure by treating $V_{A\to I}$ as queries and $V_{A\to T}$ as keys and values. Finally, the multimodal-fused image feature representation $V_{A\to T\to I}$ and the multimodal-fused text feature representation $V_{A\to I\to T}$ are the output of the hierarchical interaction module.

Output module: This module integrated multimodal-fused text and image feature representations using the self-attention-based transformer architecture [31], followed by a softmax layer for multimodal signal processing.

III. Memristive Synapse Array

A. Memristor Fabrication and Performance Testing

Considering the environmental protection and fabrication cost, an Ag/a-Carbon/Ag memristor is prepared using the frame method and magnetron sputtering method. The former method is used to prepare the switching functional layer, and the latter method is used to synthesize Ag electrodes. The specific preparation process (shown in Fig. 3) can be summarized as follows.

Step 1: The quartz plate substrate is cleaned with ethyl alcohol, acetone, and deionized water to remove any possible

contaminants, then dried in an oven at 65°C in an ambient atmosphere.

Step 2: Magnetron sputtering method is used to fabricate the bottom silver (Ag) electrode (thickness= 120 nm) on the cleaned quartz plate substrate.

Step 3: 0.05 g a-Carbon nano-powder obtained by frame method is added to 2 mL of n-methyl pyrrolidone solution and continuously stirred for 3 hours to fabricate the precursor solution.

Step 4: The quartz plate substrate is transferred on a spin coater. The precursor solution is spin-coated onto the quartz plate substrate at 2,000 rpm for 40 seconds.

Step 5: The quartz plate substrate is placed to an oven and annealed at 85°C. This leads to the formation of an a-Carbon-coating layer on the substrate.

Step 6: The top silver (Ag) electrodes are deposited on the surface of the a-Carbon coating layer, further developing the Ag/a-Carbon/Ag memristor.

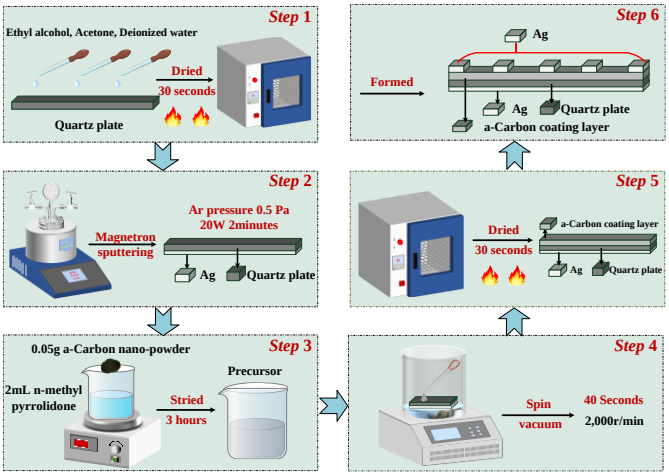


Fig. 3. Flow chart for the preparation of Ag/a-Carbon/Ag memristor.

The performance of the Ag/a-Carbon/Ag memristor is tested using an electrochemical workstation (CHI-600D). The electrical characteristics are measured with ± 2 V scanning voltages at a scan rate of 0.5 V/s, as shown in Fig. 4.

The memristor exhibits typical resistance switching (RS) behavior, as shown by the current-voltage (I-V) curves in Fig. 4(a). In the 1st phase (0 V→2 V), the current builds steadily, reaching its maximum. Meanwhile, the memristor is in a high-resistance state (HRS). When the scanning voltages sweeps from 2 V to 0 V, the memristor moves into the 2nd phase and the current decreases and the memristor remains in a low-resistance state (LRS). In the 3rd phase (0 V to −2 V), the current steadily increases until it reaches its maximum value. When a reverse bias sweep is applied, the memristor enters the 4th phase (−2 V→0 V), and the current naturally drops to a very low value. The device restitches from the LRS (3rd phase) to the HRS (4th phase). Notably, the C2C and D2D stability are crucial for hardware design and implementation [22]. The I–V curves using 120 different memristors are individually measured, and the RS behavior exhibited negligible variation, demonstrating that the Ag/a-Carbon/Ag memristors have good D2D stability (Fig. 4(b)). To investigate the stability of the memristor, the I–V curves are measured for the 1st, 10th, 50th, 200th, and 500th cycles, as shown in Fig. 4(c). The RS behavior is well maintained, which implies good C2C stability of the Ag/a-Carbon/Ag memristor. A resistance ratio between the HRS and LRS of approximately 100 can be observed and well maintained for 10^4 seconds at a reading voltage (0.5 V), indicating the good stability of the fabricated memristor (Fig. 4(d)).

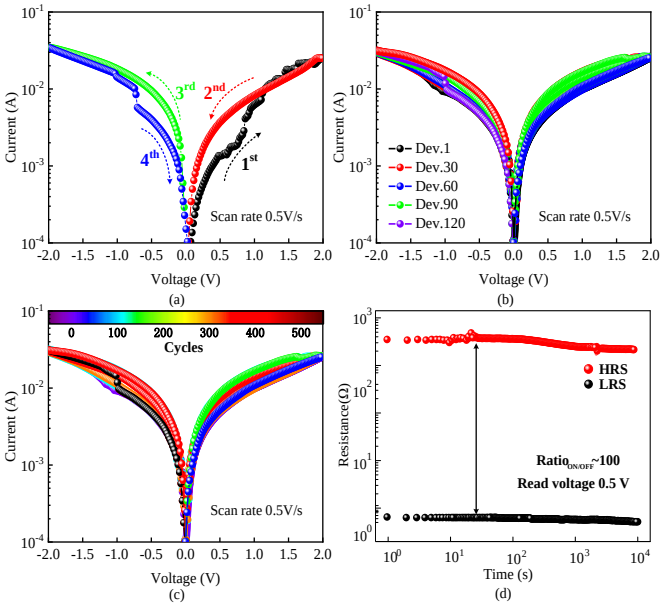


Fig. 4. (a) I-V curve of Ag/a-Carbon/Ag memristor; (b) D2D analysis; (c) C2C analysis; (d) The stability of HRS and LRS of the prepared memristor over time at 0.5V.

B. Memristive Synapse Array

Considering the proposed brain-inspired hierarchical interactive computing system has large weight parameters and matrix calculations, memristive synapse arrays are adopted to represent weights and perform an analog MAC operation based on Ohm's and Kirchhoff's laws [22]. However, the sneak path issue still affects the memristive synapse array, interfering with write or read operations and considerably increasing the power consumption.

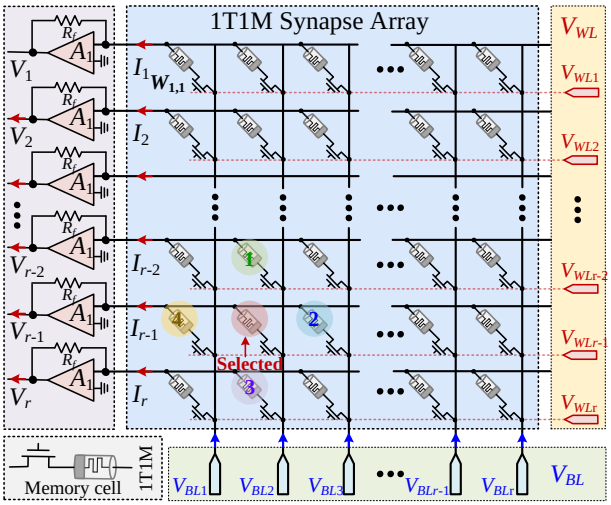
In this study, a transistor (1T) and an Ag/a-Carbon/Ag memristor (1M) connected in series consists a memory cell in the memristive synapse array, as shown in Fig. 5(a). Operational amplifier *A*1 converts current to voltage. The constructed synapse array is demonstrated as an effective manner to avoid the sneak path issue. When transistors in i_{th} row are turned off, the output voltage of the corresponding row can be denoted as $V_{out,i}$=0. In contrast, when the transistors in i_{th} row are turned on, the output voltage of the corresponding row can be expressed as:

$$V_{out,i} = R_f \cdot \sum_{i,j=1}^{row,col} W_{i,j} \cdot V_j \tag{1}$$

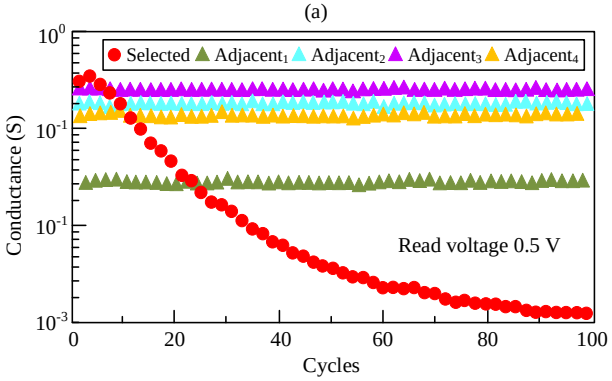
where *row* and *col* are the number of rows and columns in the memristive synapse array, respectively, R_f denotes the feedback resistance, and $W_{i,j}$ is the conductance of the memory device.

Fig. 5(b) demonstrates the conductance response of the selected memory device and the adjacent devices during the RESET operation. After 100 cycles, the conductance of the

selected device is changed from LRS to HRS, and the conductance of the adjacent devices can be maintained well.

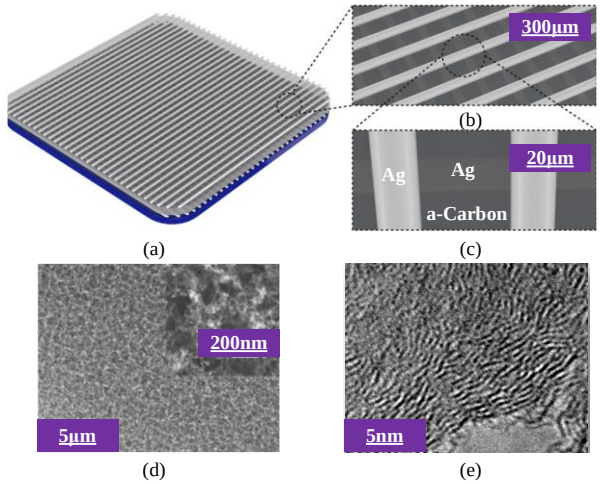


(a)



(b)

Fig. 5. (a) 1T1M synapse array; (b) conductance response of the selected memory device and the adjacent devices during the RESET operation.



(a) (b) (c) (d) (e)

Fig.6. (a) A digital image of a fabricated 1T1M synapse array (top view); optical microscopy images in (b) normal view and (c) magnified view; (d-e) FESEM image of amorphous a-Carbon switching layer at different scales.

Furthermore, for the structural and materials characterizations of the fabricated 1T1M synapse array, optical microscopy and field emission scanning electron microscopy (FESEM) are used to visualize the perfect synapse array structure and amorphous nature of the deposited thin film as a switching layer. Fig. 6(a) demonstrates the fabricated crossbar array while Fig. 6(b) and Fig. 6(c) show the optical microscopy images at different scales. As seen in Fig. 6(c), the fabricated 1T1M synapse array has a perfect crossbar structure. The FESEM results reveal the morphological analysis of the resistive switching layer of a-Carbon, which confirms the amorphous nature of the deposited thin film, as shown in Fig. 6(d) and Fig. 6(e).

IV. Circuit Design Scheme for Brain-Inspired Hierarchical Interactive IMC System

Brain-inspired computing could help us further explore neuronal functionalities and its operating mechanisms [1]. Our motivation is to design a brain-inspired hierarchical interactive IMC system via prepared memristive synapse arrays, which is capable of dealing with the computationally challenging issues involved in energy efficiencies unattainable for von Neumann architectures.

A. Unimodal Extraction Module

Various types of information can be perceived by the human sensory system despite complex surroundings. Inspired by the human brain sensory mechanism, we demonstrated a unimodal extraction module to capture features from the input text, audio, and image. Specifically, this module is consisted of three parts: memristor-based BERT, memristor-based LSTM, and memristor-based CNN.

1) Memristor-based BERT

BERT, the mainstream bidirectional pre-trained language model, can generate feature representations from unsupervised larger corpora. BERT backbone is a stack of transformer blocks [26], each of which is composed by the multi-head attention and feed-forward network, as shown in Fig. 7.

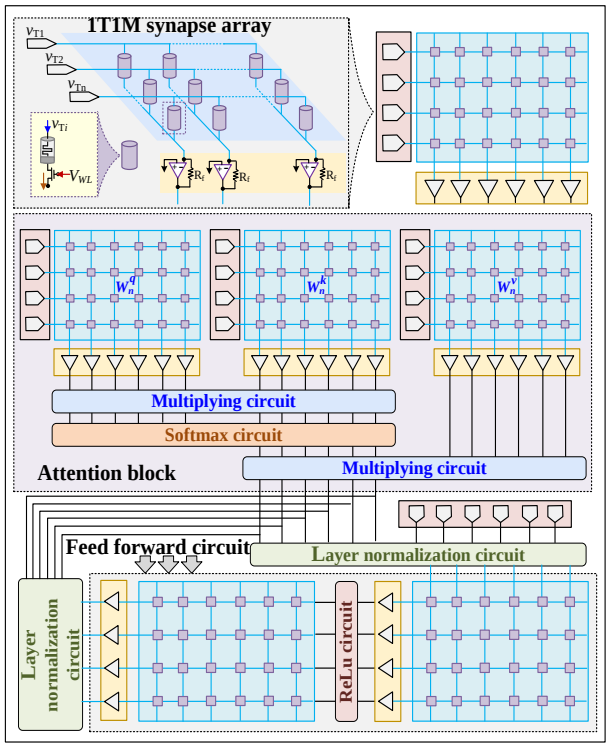


Fig. 7. The circuit design of Transformer block.

The multi-head attention can be denoted by:

$$\text{Attention}(Q,K,V)=\text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{2}$$

$$\text{Multihead}(Q,K,V)=\text{Concat}(h_1,h_2,...,h_{nh})W^O \tag{3}$$

$$h_n=\text{Attention}(QW_n^Q,KW_n^K,VW_n^V) \tag{4}$$

where $x \in R$ denotes the input matrix. T is termed as the transpose operation. The weight matrixes (W_n^Q, W_n^K, W_n^V and W^O) $\in R$ are all learnable parameters. Following the weight matrixes (W_n^Q, W_n^K, and W_n^V), the input matrix x is transformed into the attention query Q_n, attention key K_n and attention value V_n. Each head h_n is a component of complete attention.

The multi-head attention block consists of two multiplying circuits, a softmax circuit, and a number of prepared memristive synapse arrays. The multiplying circuit consists of multipliers and summation circuits that can obtain the results of the matrix weights. The softmax circuit is developed by exponential, summation, and division circuits connected in a cascaded configuration, which can convert input voltage into a probability distribution. The prepared memristive synapse array is utilized to store and compute attention query Q_n, attention key K_n and attention value V_n.

The feed-forward network is composed of two-layer normalization with a ReLU activation function in between:

$$FFN(x)=\max(0,xW_A+b_A)W_B+b_B \tag{5}$$

where weight matrixes (W_A, W_B, b_A and b_B) are learnable parameters, *FFN* refers to the feed-forward network.

The feed-forward block comprises a two-layer normalization circuit, ReLU circuit, and two prepared memristive synapse arrays. The layer normalization circuit consisted of an averaging circuit, normalization circuit, and standard deviation circuit. The ReLU circuit is used to perform nonlinear mapping operation. Using the prepared memristive synapse arrays, the learnable parameters of each row in matrices W_A and b_A and matrices W_B and b_B can be computed and stored.

Notably, the above-mentioned sub-circuits have been developed in our previous work [17].

It is noted that the input and output signals of transformer blocks are voltages, which guarantees that all these blocks can be connected in stack configuration. Based on this, the memristor-based BERT can be obtained, as shown in Fig. 8.

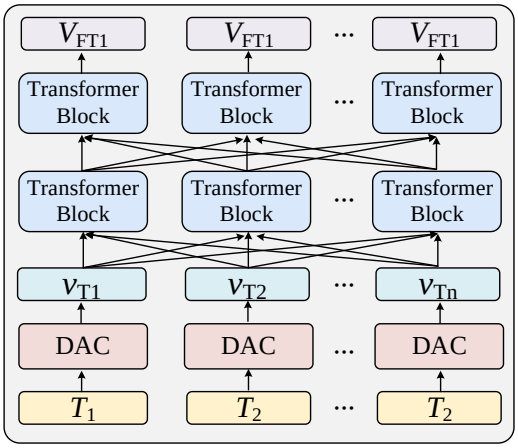


Fig. 8. The circuit design of memristor-based BERT.

2) Memristor-based LSTM

LSTM is a type of recurrent neural network for modelling long-term sequence dependencies, effectively preventing vanishing and exploding gradients [27]. The input, forget, and output gates are the three gates that constitute a typical LSTM cell. In particular, the input gate i_t controls the storage of input x_t, the forget gate f_t decides which information from the previous cell state c_{t-1} is to be abandoned, and the output gate o_t controls the cell output h_t from the current cell state c_t. The mathematical expression of LSTM is given by:

$$\begin{bmatrix} a_t \\ i_t \\ f_t \\ o_t \end{bmatrix}=\begin{bmatrix} W_a & U_a & b_a \\ W_i & U_i & b_i \\ W_f & U_f & b_f \\ W_o & U_o & b_o \end{bmatrix}\begin{bmatrix} x_t \\ h_{t-1} \\ 1 \end{bmatrix} \tag{6}$$

$$c_t=\sigma(i_t)\odot\tanh(a_t)+\sigma(f_t)\odot c_{t-1} \tag{7}$$

$$h_t=\sigma(o_t)\odot\tanh(c_t) \tag{8}$$

where the weight, recurrent weight, and bias of the LSTM are denoted as W (W_a, W_i, W_f, W_o), U (U_a, U_i, U_f, U_o), and b (b_a, b_i, b_f, b_o), respectively. Symbol σ is the logistic sigmoid, and Θ is the Hadamard product.

According to (6) ~ (8), the linear matrix operation and gated nonlinear activation are the two main stages that define an LSTM cell. Correspondingly, the linear matrix operation circuit and nonlinear activation circuit are the two parts of the specific hardware implementation of LSTM, as illustrated in Fig. 9.

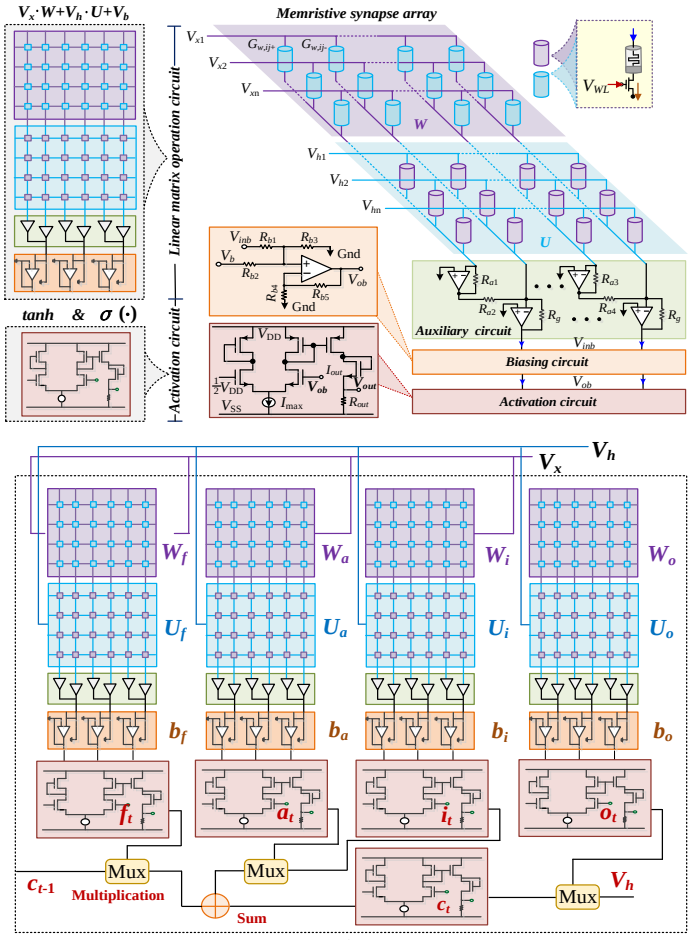


Fig. 9. The circuit design of memristor-based LSTM.

The linear matrix operation circuit: the matrix operation circuit comprises the memristive synapse array, the biasing circuit, and the auxiliary circuit to perform the linear matrix operation. The memristive synapse array is used to realize the MAC operation in the LSTM. The conductances of the memristors ($G_{w,ij}$ and $G_{u,ij}$) are used to present the weights in LSTM.

Commonly, the input vector x_t of size $N_x \times 1$ is neither a voltage nor a current, and the necessary normalization should be performed before the injection of x_t. In this study, the min-max normalization was applied using [32]:

$$V_{xi} = \frac{x_{ti} - x_{t\min}}{x_{t\max} - x_{t\min}}, i \in [1, N_x] \tag{9}$$

where $x_{t\min}$ and $x_{t\max}$ are the minimum and maximum values of x_t, respectively. V_x is the normalized voltage of size $N_x \times 1$ within the range [0, 1].

Assuming $R_{a1}=R_{a2}=R_{a3}=R_{a4}$ and $R_{b1}=R_{b2}=R_{b3}=R_{b4}=0.5R_{b5}$, the output of the matrix operation circuit V_{ob} can be written by:

$$V_{ob} = V_x \cdot W + V_h \cdot U + V_b \tag{10}$$

where the outcomes of $W = G_{w,ij+} - G_{w,ij-}$ and $U = G_{u,ij+} - G_{u,ij-}$ are the weights and recurrent weights, respectively. V_b represents the bias parameter in LSTM.

Nonlinear activation circuit: We created a general design scheme for the nonlinear activation function in LSTM. The input voltage V_{ob} was attached to one side of an N-metal oxide semiconductors (NMOS) source-coupled pair and biased by the current $I_{\max}$. The output of this circuit I_{out} can be expressed as:

$$I_{out} = I_n \cdot I_{\max} \tag{11}$$

where I_n is the normalized current, which can be described in piecewise form:

$$I_n = \begin{cases} 1, & V_{ob} > \sqrt{\frac{2}{c}} \\ \frac{1}{2} + \frac{V_{ob}}{4}\sqrt{4c - V_{ob}^2 \cdot c^2}, & |V_{ob}| \le \sqrt{\frac{2}{c}} \\ 0, & V_{ob} < -\sqrt{\frac{2}{c}} \end{cases} \tag{12}$$

where $c=\varepsilon/I_{\max}$ represents a constant, and ε denotes a gain coefficient.

The output voltage V_{out} can be obtained by:

$$V_{out} = I_n \cdot I_{\max} \cdot R_{out} + V_{SS} \tag{13}$$

where $R_{out} = 1\Omega$ indicates a constant resistor.

Next, two representative case studies are provided to verify that the activation function circuit can be used to approximate the logistic sigmoid function and hyperbolic tangent function, respectively. The specific parameter setting and experiment results are exhibited in Fig. 10. From Fig. 10, the two brown dotted lines are the input-output curves of the proposed activation circuit under different parameter pairs, and the two navy solid lines denote the logistic sigmoid function and the hyperbolic tangent function, respectively. It is clear that the proposed activation circuit is able to approximate the logistic sigmoid function and hyperbolic tangent function with a high degree of accuracy. Actually, the proposed circuit can be regarded as a general activation circuit, it can be used to realize almost all the activation functions by tuning circuit parameters (i.e., the parameter pair V_{SS} and c).

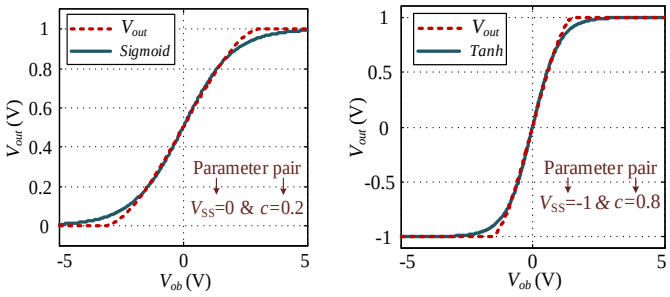


Fig. 10. Analog implementation of the activation function. (a) The logistic sigmoid function; (b) The hyperbolic tangent function

3) Memristor-based CNN

CNN, a class of deep neural networks, has become the most established class with remarkable achievements in performing tasks related to image processing, such as image recognition, image segmentation, and object detection [28]. The general architecture of LeNet-5 [33] is depicted in Fig. 11(a), it is composed of one convolutional layer and one max-pooling layer, of which the structure repeats once, followed by three fully connected layers. Accordingly, the detailed hardware implementation of the CNN can also be divided into three components: the convolutional operation circuit, max-pooling circuit, and fully connected circuit.

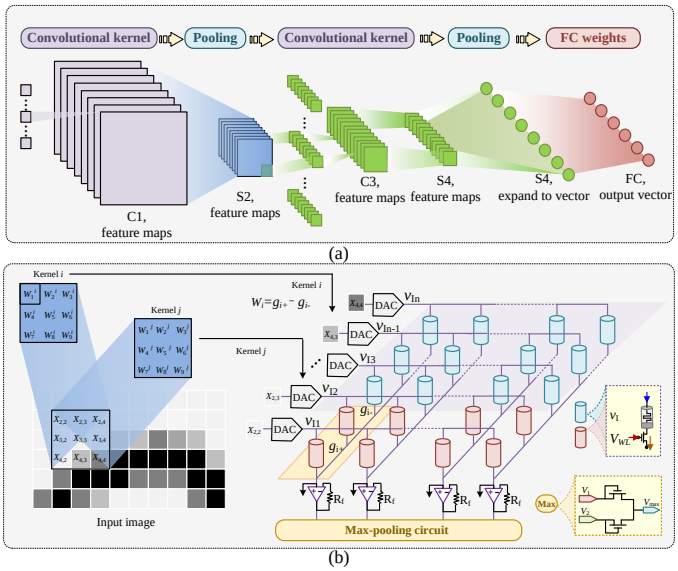


Fig. 11. (a) The general architecture of LeNet-5; (b) The circuit design of memristor-based CNN.

The convolutional operation circuit uses sliding operations with different kernels to achieve convolutional operations. Memristive synapse arrays play a significant role in realizing parallel MACs with the same input for various kernels. Fig. 11(b) demonstrates the typical convolution at a certain slipping step and the events realized by the memristive synapse arrays. Based on the conductance difference of two memristors, the signed kernel weight is first mapped, and all kernel weights are mapped to two conductance rows for positive weights g_{i+} with positive inputs x_{i+}, and for negative weights g_{i-} with negative inputs x_{i-}. While mapping different kernels to different pairs of

rows, the memristive synapse array performs MACs in parallel under the shared inputs, and in the meantime, we can get the desired weighted-sum results. The output of the convolutional operation circuit y_k is denoted as:

$$y_k = f\left[\sum_{i=1}^{m}\left(x_{i+}g_{i+} - x_{i-}g_{i-}\right) + x_{b+}g_{b+} - x_{b-}g_{b-}\right] \quad (14)$$

where k is the size of the convolution kernel. f is the ReLU function. Voltage x_b represents the additional input of the memristive synapse array. Here, the fully connected circuit could be a convolutional operation circuit with a 1×1 kernel. Because the design is the same for both types of circuits, we will not go into all specifics again.

Max-pooling circuit: max-pooling operation reduces the spatial size and the number of parameters, which can effectively prevent overfitting. As illustrated in Fig. 11(b), we adopted the voltage selector to simulate this operation and do not need consider whether the input is positive or negative.

At the beginning, all memristors installed at the intersections of the weight matrices can be set to an appropriate conductance value. Following the training voltages v_T, v_A, and v_I and based on the stochastic gradient descent approach, the conductances were updated to the desired values. When the training phase was completed, feature voltages were provided, symbolized by V_{FT}=BERT(v_T), V_{FA}=LSTM(v_A),and V_{FI}=CNN(v_I).

B. Hierarchical Interaction Module

The cross-modal semantic gap results from the fact that the text, audio, and image feature voltages are acquired in different representation spaces via the aforementioned three pre-trained unimodal extraction modules (memristor-based BERT, memristor-based LSTM, and memristor-based CNN). In this section, the hierarchical interaction module aims at eliminating the semantic gap of the three unimodal feature voltages while effectively capturing the cross-modal interactions, as shown in Fig. 12.

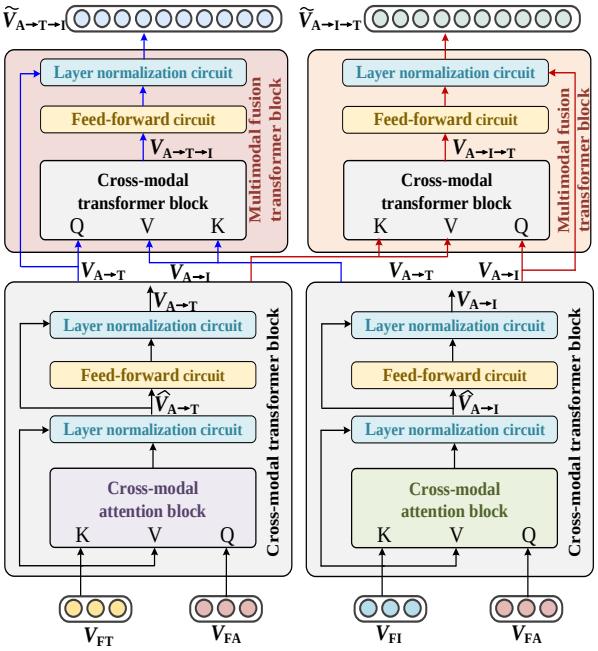


Fig. 12. The circuit design of hierarchical interaction module.

In the cross-modal transformer block, we consider two distinct modalities: V_{FA} and V_{FT} (acoustic and textual modalities, respectively). Considering W_{QA}, W_{KT}, and W_{VT} are all weight matrices, we specified the queries as $Q_A = V_{FA} \times W_{QA}$, the keys as $K_T = V_{FT} \times W_{KT}$, and the values as $V_T = V_{FT} \times W_{VT}$. The cross-modal transformer block has a multi-head version of cross-modal attention and a feed-forward network. The output of the cross-modal transformer block can be achieved using:

$$V_{A\to T} = LN\left(V_{FT} + CMA\left(V_{FT}, V_{FA}\right)\right) \quad (15)$$

$$V_{A\to T} = LN\left(FFN\left(V_{A\to T}\right) + V_{A\to T}\right) \quad (16)$$

where *CMA* is the multi-head cross-modal attention from audio to text, *LN* denotes layer normalization. $V_{A\to T}$ is the output of a layer in the cross-modal attention block. After the cross-modal transformer block, $V_{A\to T}$ is obtained as a representation of the textual modality. Similarly, the visual modality is subjected to the cross-modal transformer block, and $V_{A\to I}$ is obtained for the visual modality.

In the multimodal fusion transformer block, we firstly used a cross-modal transformer block for intermodal interactions. Queries are defined as $Q_{A\to T} = V_{A\to T} \times W_{QA\to T}$, keys as $K_{A\to I} = V_{A\to I} \times W_{KA\to I}$, and values as $V_{A\to I} = V_{A\to I} \times W_{VA\to I}$, where $W_{QA\to T}$, $W_{KA\to I}$, and $W_{VA\to I}$ are weight matrices. After the cross-modal transformer block, $V_{A\to T\to I}$ was obtained. To combine $V_{A\to T\to I}$ with $V_{A\to T}$, we fed $V_{A\to T\to I}$ to a feed-forward network first, the corresponding output and $V_{A\to T}$ are jointly injected to the layer normalization. The output of the multimodal fusion transformer block $V_{A\to T\to I}$ is denoted as:

$$V_{A\to T\to I} = CMT\left(V_{A\to I}, V_{A\to T}\right) \quad (17)$$

$$V_{A\to T\to I} = LN\left(FFN\left(V_{A\to T\to I}\right) + V_{A\to T}\right) \quad (18)$$

where *CMT* denotes the cross-modal transformer operation.

Notably, the multimodal-fused text representation, represented by $V_{A\to I\to T}$, uses the same structure as the final multimodal-fused image representation $V_{A\to T\to I}$. Meanwhile, all the functional circuits embedded in the hierarchical interaction module have been described in Section IV-A, we will not go into all specifics again.

C. Output Module

Finally, we considered the hidden representation $V_{A\to T\to I}$ as the final image feature representation, and concatenated it with the multimodal-fused text feature representation $V_{A\to I\to T}$. After that the concatenated result is fed to a self-attention-based transformer layer:

$$V_M = Transformer\left(V_{A\to T\to I}, V_{A\to I\to T}\right) \quad (19)$$

where V_M is the final multimodal feature representation produced by the transformer layer.

Subsequently, we fed the final hidden representation V_M to a softmax circuit for multimodal signal processing.

V. Application in Video Sentiment Analysis

To verify the effectiveness and validity of the proposed

hierarchical interactive IMC system, a series of experiments are carried out by comparing the proposed system with the state-of-the-art methods [34-47] for video sentiment analysis. Notably, the references are selected according to the following four criteria, i.e., content relevance, total cited times, journal academic impact, and timeliness. Furthermore, the main hyperparameters used for the proposed system (including neural network and circuit parameters) are shown in Table I.

TABLE I
LIST OF THE HYPERPARAMETERS USED FOR THE PROPOSED SYSTEM

	Hyperparameters	
Neural network parameters	Learning rate	10^{-2}
	Momentum	0
	Decay	0.9
	Maximum error	10^{-4}
Circuit parameters	V_{min}	0V
	V_{max}	2.0V
	$Ratio_{HRS/LRS}$	10
	Read voltage	0.5V
	V_{WL}	1.5V
	V_{BL}	1.0V
	Scan voltage	0.5V/s

In order to ensure the classification performance, the selection of neural network hyperparameters are based on [34-47]. The circuit hyperparameters mainly rely on the prepared memristor, they are set up to keep the proposed system proper functioning and have no effect on the classification performance.

A. Database and Evaluation Metrics

Two benchmark video clip datasets (i.e., the IEMOCAP dataset and the MELD dataset), containing textual, acoustic, and visual information of each utterance, are applied to evaluate the proposed system.

IEMOCAP dataset is recorded as video clips of multimodal dyadic conversations between actors of opposite gender. Following the previous studies [34-47], 80% data is distributed in training dataset and the remaining 20% is distributed in testing dataset. Each utterance has a certain emotion label including six categories, i.e., “exited,” “happy,” “neural,” “sad,” “angry,” and “frustrated”. MELD dataset contains 1433 dialogues and 13708 utterances, in which 1039 dialogues with 9989 utterances are distributed in training dataset, 114 dialogues with 1109 utterances are distributed in validation dataset, and the remaining dialogues are distributed in testing dataset. Each utterance has a certain emotion label including seven categories, i.e., “neutral,” “surprise,” “fear,” “sad,” “angry,” “disgust” and “joy”. Specifically, the sample distribution of different emotions in IEMOCAP dataset and MELD dataset is shown in Table II and Table III.

TABLE II
SAMPLE DISTRIBUTION OF DIFFERENT EMOTIONS IN IEMOCAP DATASET (%)

Dataset	Exited	Happy	Neural	Sad	Angry	Frustrated
Training	8.8%	7.2%	13.6%	11.2%	19.2%	20%
Testing	2.2%	1.8%	3.4%	2.8%	4.8%	5%

TABLE III
SAMPLE DISTRIBUTION OF DIFFERENT EMOTIONS IN MELD DATASET (SAMPLES)

Dataset	Neutral	Surprise	Fear	Sad	Angry	Disgust	Joy
Training	4710	1205	268	683	1109	271	1743
Validation	470	150	40	111	153	22	163
Testing	1256	281	50	208	345	68	402

Then, two common performance metrics F1-score and accuracy [35] are used to evaluate the overall performance.

B. Hardware-friendly Training Method

The hardware friendly training method of the network, containing the forward pass, error backpropagation, and weight update, is implemented in the proposed hierarchical interactive IMC system and PyTorch platform, as illustrated in Fig. 13.

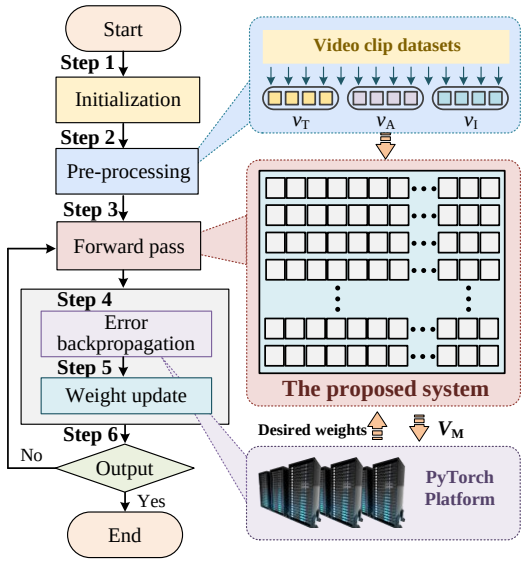


Fig. 13. The flow chart of hardware-friendly training phase.

Step 1. Initialization: At the beginning, the resistances of all the electronic devices in the memristive synapse array can be initialized to an intermediate value between LRS and HRS.

Step 2. Data pre-processing: The multimodal information in the training dataset is converted to voltage signal $\boldsymbol{v}$ (v_T, v_A, and v_I) within the range of [−2, 2] via digital to analog converter.

Step 3. Forward pass: The voltage signals v_T, v_A, and v_I are injected into the proposed circuit system, and then the corresponding output V_M can be achieved.

Step 4. Error backpropagation: The error backpropagation is implemented in the PyTorch-based GPU platform based on the stochastic gradient descent approach by a factor of 10^{-2}, and the desired weights can be obtained immediately.

Step 5. Weight update: the prepared memristors are programed to the desired weights by tuning the gate of the memory cell.

Step 6. Completion: until the entire hierarchical interactive IMC system settles, the training process is completed, otherwise, return to **Step 3**.

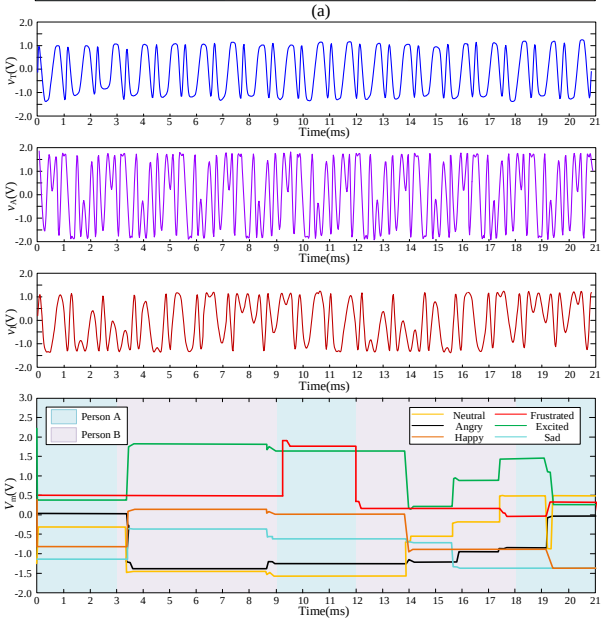
This hardware-friendly training method combines the advantages of the energy efficiency of the 1T1M synapse array in performing the analog MAC operation and digital logic for realizing the rest of the training process.

C. Classification Results

Two conversations classified by the proposed hierarchical interactive IMC system on IEMOCAP dataset and MELD dataset are illustrated in Fig. 14 and Fig. 15, respectively. In each conversation, the input multimodal information is converted to voltage signal v_T (blue solid line), v_A (purple solid line), and v_I (brown solid line). The voltage signals v_T, v_A, and v_I

Utterances from Person A	Utterances from Person B	This work	Ground Truth
Pass up what? Its little fish...		Frustrated	Frustrated
	Yeah but that moment on the beach is the highlight...	Excited	Excited
	This is a big night for these guys...	Excited	Excited
What are you so excited about? I mean it is not...		Frustrated	Frustrated
	Okay, but they are fish that do something...	Excited	Excited
	I can not understand how you cloud not be intersting.	Excited	Excited
I am just not.		Neutral	Neutral

(a)

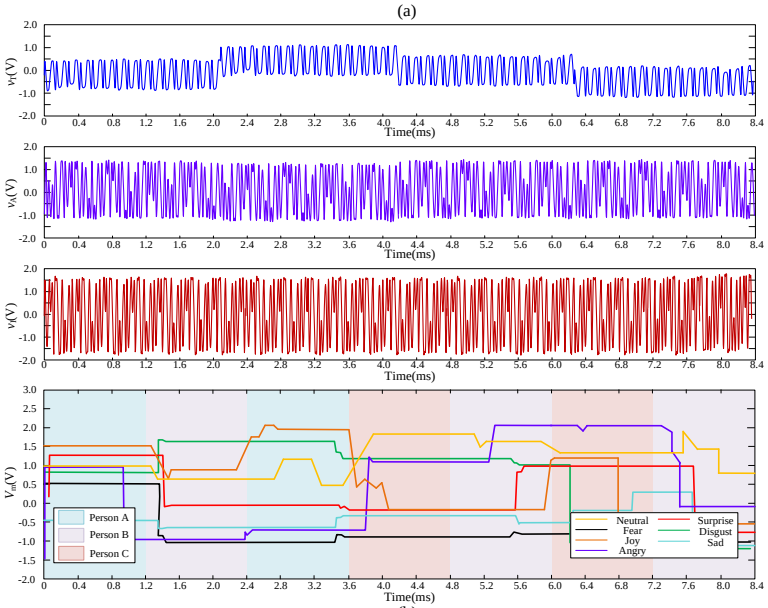


(b)

Fig. 14. (a) Illustration of a conversation from IEMOCAP dataset; (b) The corresponding results obtained by the proposed system.

Utterances from Person A	Utterances from Person B	Utterances from Person C	This work	Ground Truth
What do you think? Today is the big day			Joy	Joy
	Okay,···, do not ever tell me what you did in there.		Disgust	Disgust
Thank you!			Joy	Joy
		Where is Joey? His mom is on the phone	Neutral	Neutral
	He is in the bathroom. I do not think you want go in there		Angry	Angry
		We are roommates! My eyes!! My eyes!!	Angry	Angry
	I warned you ...		Neutral	Neutral

(a)



(b)

Fig. 15. (a) Illustration of a conversation from MELD dataset; (b) The corresponding results obtained by the proposed system.

are injected into the proposed circuit system. Following the training voltages v_T, v_A, and v_I and based on the stochastic gradient descent approach, the conductances are updated to the desired values. When the training phase is completed, the softmax circuit will output a set of voltage signals V_M with six/seven states (assigned to six/seven emotion states) representing a probability distribution. The classification result is determined by the largest output voltage in each period. Specifically, in Fig. 14, person A is in a frustrated emotion (red solid line) at the beginning. In the contrast, person B is in excited emotion (green solid line) all the time and tries to help person A out of the frustrated emotion. As a result, person A is in neutral emotion in the end. In Fig. 15, person A is in joy emotion (orange solid line), while person B is in disgust emotion (green solid line) in the initial state. As a neutral observer (yellow solid line), person C changes his emotion and becomes angry (purple solid line) after talking with person B. After that, person B gradually clams down (i.e., neutral emotion). The results demonstrate that the interaction between speakers can change the emotion state.

Then, the proposed IMC system is compared with the state-of-the-art methods on the IEMOCAP dataset and the MELD dataset, as shown in Table IV and Table V, respectively.

We report two common performance metrics F1-score and accuracy for each emotion category. For the IEMOCAP dataset, Table IV demonstrates that the proposed IMC system achieves the improvements on F1-score and accuracy in sad and excited emotion recognition tasks over state-of-the-art methods. Meanwhile, the classification performance of happy, neutral, and frustrated emotions also achieves the top three rankings, slightly outperforming other competitors. Notably, the average F1-score and accuracy win the second place over currently advanced approaches. For the MELD dataset, experimental results in Table V demonstrate that the proposed IMC system also outperforms other competitors (top three rankings in F1-score and accuracy), especially for the two minority classes fear (+0.6% Acc., +0.4% F1) and disgust (+0.8% Acc., +0.6% F1). Except for [44-46], the proposed method is superior to the other competitors [34-43, 47] in terms of the average F1-score and accuracy.

The main reason may be that: 1) compared with [34, 36-43, 47] there is a significant drop in classification performance when the unimodal extraction module is not considered. It is concluded that unimodal feature representation is crucial for multimodal sentiment analysis. 2) comparing with the results of removing hierarchical interaction module [34, 35, 39, 41, 42], it also leads to the reduction of classification accuracy. This observation also proves that the inter-modal interaction information is necessary for video sentiment analysis, especially for the complex fine-grained tasks. 3) Although [44-46] slightly outperforms the proposed method (within an acceptable range), these methods bring more computational

TABLE IV
COMPARISON OF DIFFERENT STATE-OF-THE-ART METHODS ON IEMOCAP

Ref.	IEMOCAP: 6-way Emotion Categories													
	Happy		Sad		Neutral		Angry		Excited		Frustrated		Average	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
[34]	24.0	34.1	65.6	70.5	55.5	52.1	**72.3$_3$**	**66.8$_3$**	64.3	62.1	**67.9$_3$**	62.5$_3$	60.1	59.9
[35]	25.7	33.2	75.1	78.8	58.6	59.2	64.7	65.3	80.3	71.9	61.2	58.9	63.4	62.8
[36]	43.1	50.6	69.4	76.8	63.0	**62.9$_2$**	63.5	56.5	**88.3$_2$**	**77.9$_3$**	53.3	55.7	64.6	64.3
[37]	44.8	40.0	**79.2$_3$**	**82.4$_3$**	60.8	61.9	**73.5$_1$**	**68.1$_1$**	63.0	69.3	64.2	59.7	65.0	64.5
[38]	47.9	51.3	78.0	79.9	**69.0$_1$**	**65.8$_1$**	**72.9$_2$**	**67.2$_2$**	**85.3$_3$**	**78.7$_2$**	52.2	58.5	**68.0$_3$**	**67.5$_3$**
[39]	**55.4$_3$**	31.6	**80.2$_2$**	**84.1$_2$**	64.7	59.7	69.1	65.3	63.2	74.3	61.1	61.5	66.1$_3$	64.6$_3$
[40]	55.1	**55.8$_1$**	70.8	73.3	**66.8$_2$**	61.9	62.1	66.0	65.3	69.5	65.7	**64.2$_3$**	65.3	65.4
[41]	/	/	/	/	/	/	/	/	/	/	/	/	65.7	64.2
[42]	**62.1$_1$**	**54.5$_2$**	66.6	72.7	63.9	59.4	58.4	61.0	58.5	66.6	64.8	61.6	62.8	63.0
[43]	43.6	40.3	72.5	70.7	52.5	52.5	66.2	61.6	69.2	65.1	55.5	61.1	59.5	59.5
[44]	/	/	/	/	/	/	/	/	/	/	/	/	**73.9$_1$**	**74.2$_1$**
[46]	/	/	/	/	/	/	/	/	/	/	/	/	69.4	69.6
[47]	24.3	30.2	64.5	74.2	57.3	59.0	61.8	62.7	81.3	72.5	**75.9$_1$**	**66.6$_1$**	64.7	64.1
This work	**55.8$_2$**	**51.4$_3$**	**80.5$_1$**	**84.4$_1$**	**64.2$_3$**	**62.0$_3$**	65.2	64.2	**88.5$_1$**	**78.9$_1$**	**68.2$_2$**	**64.5$_2$**	**70.6$_2$**	**67.9$_2$**

TABLE V
COMPARISON OF DIFFERENT STATE-OF-THE-ART METHODS ON MELD

Ref.	MELD: 7-way Emotion Categories															
	Neutral		Surprise		Fear		Sad		Joy		Disgust		Angry		Average	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
[34]	79.1	76.2	43.7	40.7	2.8	2.2	13.2	13.7	47.5	46.7	1.8	1.6	**41.4$_3$**	40.8	55.8	54.7
[35]	72.1	73.5	54.4	49.4	1.6	1.2	**23.9$_3$**	23.8	52.0	50.7	1.5	1.7	**41.9$_2$**	41.5	56.1	55.9
[36]	80.1	**78.9$_1$**	**56.9$_1$**	**55.4$_1$**	7.4	8.6	22.6	24.9	**55.3$_1$**	**57.4$_1$**	2.8	3.5	40.2	41.0	61.4	60.4
[38]	79.6	**77.4$_3$**	54.6	52.7	8.4$_3$	**10.0$_3$**	**28.3$_1$**	**32.5$_1$**	52.1	**56.0$_3$**	10.3	11.2	37.0	**44.6$_2$**	59.7	60.5
[39]	/	/	/	/	/	/	/	/	/	/	/	/	/	/	60.9	/
[40]	**83.5$_1$**	76.7	**55.4$_3$**	**53.2$_3$**	**8.6$_3$**	**11.7$_2$**	16.1	21.8	52.9	53.6	**16.5$_2$**	**21.9$_2$**	38.9	**42.6$_3$**	61.3	59.0
[41]	/	/	/	/	/	/	/	/	/	/	/	/	/	/	61.4	58.6
[42]	80.7	75.4	50.7	47.9	7.2	**10.0$_3$**	20.9	**26.3$_3$**	**53.9$_3$**	52.1	**10.4$_3$**	**12.8$_3$**	33.9	38.1	59.2	57.1
[43]	71.7	76.5	49.6	47.5	7.7	8.9	16.1	12.5	51.82	52.4	10.2	11.9	42.8	46.7	60.5	57.9
[44]	/	/	/	/	/	/	/	/	/	/	/	/	/	/	**65.6$_3$**	**64.5$_3$**
[45]	/	/	/	/	/	/	/	/	/	/	/	/	/	/	**67.4$_1$**	**67.2$_1$**
[46]	/	/	/	/	/	/	/	/	/	/	/	/	/	/	**66.6$_2$**	**66.3$_2$**
[47]	**82.3$_2$**	76.0	47.3	46.9	**9.7$_1$**	8.8	3.8	4.6	53.2	52.3	7.6	6.7	**53.0$_1$**	**47.6$_1$**	60.6	58.6
This work	**81.7$_3$**	**78.1$_2$**	**56.2$_2$**	**54.5$_2$**	**9.2$_2$**	**12.1$_1$**	**25.9$_2$**	**27.6$_2$**	**54.2$_2$**	**56.5$_2$**	**17.6$_1$**	**22.5$_1$**	40.9	43.8	61.7	60.8

Note: the subscript 1, 2, 3 represent the corresponding ranking results.

cost and energy consumption. Notably, since the forward calculation is implemented in proposed hierarchical interactive IMC system, the proposed method is faster than other competitors, indicating that the proposed system achieves a good trade-off between classification accuracy and computational efficiency. It is noted that some effective tricks and sub-modules (e.g., Tensor-based Multimodal Transformer) hardware implementations can be considered to add into the proposed scheme to increase the classification performance, which is our future work.

Furthermore, we visualize the confusion matrixes of the proposed system on the IEMOCAP testing dataset and the MELD testing dataset in Fig. 16.

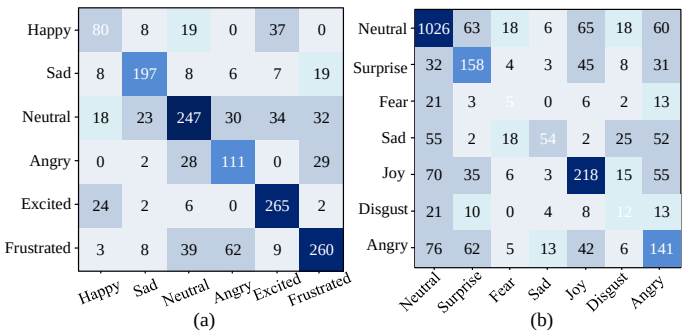


Fig. 16. Confusion matrix of the testing dataset (a) IEMOCAP dataset; (b) MELD dataset

In Fig. 16(a), we find neutral and angry emotions are confused with the frustrated emotion in some cases. The reason maybe that the majority of the utterances are labeled as the frustrated emotion in the IEMOCAP dataset. Similarly, it can be observed that the happy emotion is confused with the excited emotion in some times. The phenomenon is related to that excited and happy emotions are really close in the valence and activation domains. In Fig. 16(b), we find fear and disgust emotions are confused with other emotions in some cases, leading to lower accuracies. The main reason maybe that the number of utterances labelled as these two negative emotions is relatively less in the MELD dataset, and the proposed system inevitably tends to learn less compared with other emotion categories. As a result, how to deal with the imbalanced class distribution issue is our future work.

To explore the hierarchical interactive effect among different modalities, we use different modality combinations on the two benchmark video clip datasets, as shown in Table VI.

From Table VI, it can be observed that the textual modality is the dominant information source for video sentiment analysis. The audio and visual modalities perform poorly when used alone or in combination. The audio or visual modality is used interactively with the textual modality, the average F1-score and accuracy slightly outperforming the textual modality. The result demonstrating that the audio and visual modalities

provide supplementary information with the textual modality. Notably, the best performance is achieved when all modalities are used interactively and hierarchically.

TABLE VI
PERFORMANCE OF THE PROPOSED SYSTEM USING DIFFERENT MODALITY COMBINATIONS ON BENCHMARK VIDEO CLIP DATASETS

Modality	IEMOCAP		MELD	
	Acc.	F1	Acc.	F1
T	68.8	66.1	59.8	59.1
A	42.3	39.5	37.0	53.4
V	48.4	45.5	42.4	40.7
T+A	69.7	66.6	60.9	59.6
T+V	70.1	67.0	61.2	59.9
A+V	52.1	49.6	47.2	44.9
T+A+V	70.6	67.9	61.7	60.8

Note: T, A, V represent the textual, audio and visual modalities, respectively.

D. Computational Efficiency

The computational efficiency in terms of time, power, and area can be estimated based on the state-of-the-art components available at the 180 nm CMOS node. We analyze the time consumption of the proposed hierarchical interactive IMC system by comparing it with state-of-the-art methods on the IEMOCAP dataset and the MELD dataset. Considering the back-propagation calculation of all methods is based on software implementation, we only compare the time consumption of forward propagation with the state-of-the-art methods, as shown in Fig. 17.

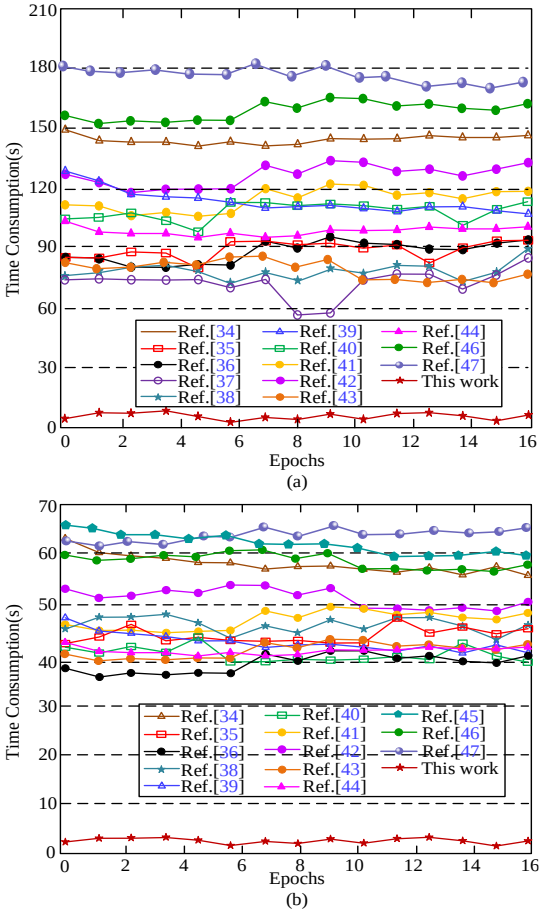


Fig. 17. The time consumption of forward propagation (a)IEMOCAP dataset; (b)MELD dataset

The results demonstrate that the proposed system takes less time (approximately 8~15 times) than other competitors. The reason is related to the efficient analog MAC operations via memristive synapse array.

Table VII shows the power consumption of each circuit module and the entire proposed hierarchical interactive IMC system. Commonly, the dimensions of memory devices, transistors, and resistors are all micron order. The energy and the power consumption for 1-bit computing are 2231.34 pJ and 44.63mW with 0.5V, 50ns read voltage, respectively. The circuit is implemented using a 180-nm CMOS technology. The total area of the proposed system is about 64.73μm^2. The proposed system has advantages in terms of time, power, and area, which indicates that the proposed system is cost saving and energy-efficient.

TABLE VII
THE ENERGY CONSUMPTION OF THE PROPOSED SYSTEM

Module	Power consumption/pJ
1T1M Synapse Array	30.32
DAC	326.4
Multiplying circuit	8.22
ReLU circuit	0.45
Softmax circuit	0.66
Layer normalization circuit	1.62
Max-pooling circuit	0.27
Total	2231.34

E. Non-idealities Analysis

Here, the non-idealities analysis mainly refers to the anti-noise analysis and the device failure analysis, the specific description is provided below:

1) Anti-noise analysis: considering the multimodal information perception and transmission may be inevitably affected by the noise, we added the read noise to these three multimodal signals (Text signal, audio signal, and image signal), and the classification accuracy on the IEMOCAP dataset and the MELD dataset are demonstrated in Fig. 18(a) and Fig. 18(b), respectively. It can be seen that the classification accuracy can be maintained at a high level even though all the three multimodal signals are polluted by read noise. In other words, the read noise has little influence on the classification accuracy, indicating that the proposed system has good anti-noise ability (especially for the textual modality).

2) Device failure analysis: considering the device failure phenomenon may occur, this work sets a ratio range (0-50%) of failed memristors. As shown in Fig. 18(c), when the memristor is in LRS and the failure ratio reaches about 20%, the classification accuracy can be kept over 70% and 60% on the IEMOCAP dataset and the MELD dataset, respectively. Once the memristor failure ratio exceeds 20%, the accuracy decreases sharply to about 20% on the two benchmark video clip datasets. When the memristor is in HRS, the accuracy deceases smoothly with the increase of device failure ratio. When the device failure ratio reaches about 25%, the classification accuracy can be maintained in an acceptable level (~67% on the IEMOCAP dataset and ~58% the MELD dataset). Based on this, we can conclude that the proposed system has a better tolerance to the failed memristor in HRS. The reason related to this phenomenon may be that LRS failure always produce large error current, and the output of the weighted

TABLE VIII
COMPARISON OF EXISTING IN-MEMORY COMPUTING SYSTEMS

Reference	Cross-modal Interaction	Computing devices			Network structure	Application
		Materials	R_{on}/R_{off}	Stability		
[48]	No	TiN/TaO_x/HfO_x/TiN	~10^3	High	CNN	Handwritten recognition
[49]	No	RRAM device	/	/	CNN/LSTM	Handwritten recognition
[50]	No	HP memristor	/	High	LSTM	Sentiment analysis
[51]	No	Pt/TaO_x/Ta	/	/	ConvLSTM	Handwritten recognition
[52]	No	AIST-based model	/	/	Transformer	Handwritten recognition
[17]	No	2D material-based	~10^2	High	Transformer/SNN	Behavior recognition
[53]	No	MRAM device	~10^3	/	BNN	Face recognition
This work	Yes	Carbon-based	~10^2	High	Transformer/CNN/LSTM…	Sentiment analysis

summation may be affected significantly. While HRS failure does not generate large error current, thus keeping a good classification performance despite the large failure rate of the memristors.

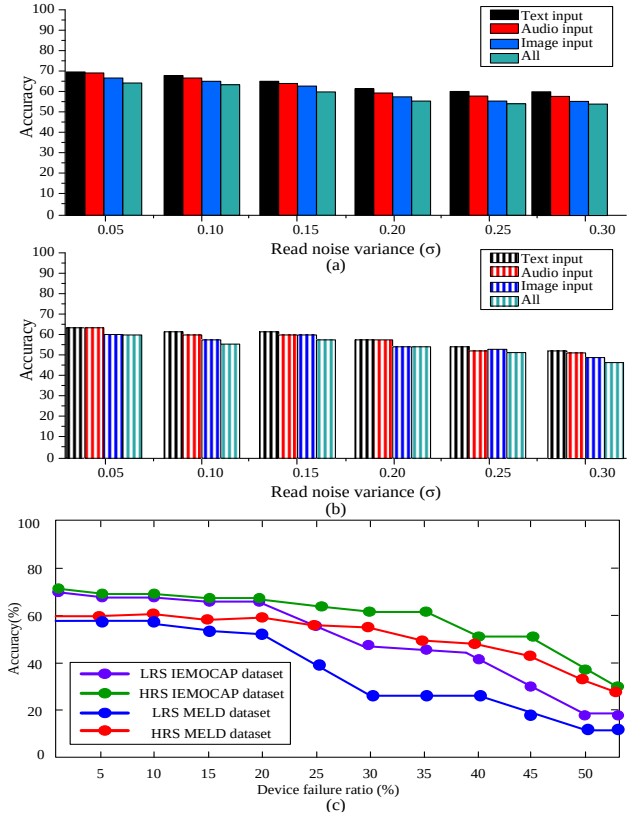


Fig. 18. Non-idealises analysis of the proposed system (a) anti-noise analysis on IEMOCAP dataset; (b) anti-noise analysis on MELD dataset; (c) device failure analysis

A comparison between different IMC computing systems is provided in Table VIII. From Table VIII, much of the work in IMC has focused on hardware development of neural networks based on specific structures (e.g., convolutional neural networks (CNNs) [48, 49], long short-term memory (LSTM) networks [50, 51], transformer networks [52, 17], and binary neural network (BNN) [53])). Almost all existing IMC systems concentrate on a single mode of sensory processing. A lack of hierarchical interactive systems that combine numerous senses based on cross-modal learning mechanisms make information across various sensory modalities hard to handle. In this paper, a brain-inspired hierarchical interactive IMC system is proposed, which can efficiently solve 'von Neumann bottleneck', enabling cross-modal interactions and semantic gap elimination.

VI. CONCLUSION

This paper mainly focuses on the investigation of a brain-inspired hierarchical interactive IMC system for video sentiment analysis. Specifically, the Ag/a-Carbon/Ag memristor is prepared based on frame method and magnetron sputtering method, and the corresponding performance testing demonstrates its high stability. Then, the Ag/a-Carbon/Ag memristor circuit with the 1T1M configuration is realized, which enables parallel MAC operations. Meanwhile, a brain-inspired hierarchical interactive IMC system mainly consisted of unimodal extraction module, hierarchical interactive module, and output module is designed. For verification, the brain-inspired hierarchical interactive IMC system is applied to realize the video sentiment analysis and the experimental results demonstrate the proposed system has good performance in terms of classification accuracy, computational efficiency, and robustness. The future direction of research includes the development of brain-inspired IMC systems and investigation of new techniques for the deep integration of nanotechnology and energy-efficient integrated circuits.

VII. REFERENCES

[1] Y. Ou, Z. Chen and F. Wu, "Multimodal Local-Global Attention Network for Affective Video Content Analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 1901-1914, 2021, doi: 10.1109/TCSVT.2020.3014889.

[2] Y. Zhu, Z. Chen and F. Wu, "Affective Video Content Analysis via Multimodal Deep Quality Embedding Network," in *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1401-1415, 1 July-Sept. 2022, doi: 10.1109/TAFFC.2020.3004114.

[3] J. Tang et al., "BAFN: Bi-direction Attention based Fusion Network for Multimodal Sentiment Analysis," *IEEE Trans. Circuits Syst. Video Technol.*, 2022, doi: 10.1109/TCSVT.2022.3218018.

[4] W. Seo, N. Kim, C. Park, and S. M. Park, "Deep learning approach for detecting work-related stress using multimodal signals," *IEEE Sens. J.*, vol. 22, no. 2, pp. 11892-11902, 2022, doi:10.1109/JSEN.2022.3170915.

[5] R. Chen, W. Zhou, Y. Li and H. Zhou, "Video-Based Cross-Modal Auxiliary Network for Multimodal Sentiment Analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 12, pp. 8703-8716, Dec. 2022, doi: 10.1109/TCSVT.2022.3197420.

[6] R. Yang, W. Zhang, N. Tiwari, H. Yan, T. Li, and H. Cheng, "Multimodal sensors with decoupled sensing mechanisms," *Adv. Sci.*, pp. 2202470, 2022, doi:10.1002/advs.202202470.

[7] P. R. Genssler, and H. Amrouch, "Brain-inspired computing for circuit reliability characterization," *IEEE Trans. Comput.*, 2022, doi:10.1109/TC.2022.3151857.

[8] T. J. Park, S. Deng, S. Manna, A. N. M. N. Islam, H. Yu, Y. Yuan, D. D. Fong, A. A. Chubykin, A. Sengupta, S. K. R. S. Sankaranarayanan, and S. Ramanathan, "Complex oxides for brain-inspired computing: A review," *Adv. Mater.*, pp. 2203352, 2022, doi:10.1002/adma.202203352.

[9] A. Mehonic, and A. J. Kenyon, "Brain-inspired computing needs a master plan," *Nature*, vol. 604, no. 7905, pp. 255-260, 2022, doi:10.1038/s41586-021-04362-w.

[10] J. B. Aimone, "A roadmap for reaching the potential of brain-derived computing," *Adv. Intell. Syst.*, vol. 3, no. 1, pp. 2000191, 2021, doi:10.1002/aisy.202000191.

[11] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, "Memory devices and applications for in-memory computing," *Nat. Nanotechnol.*, vol. 15, no. 7, pp. 529-544, 2020, doi:10.1038/s41565-020-0655-z.

[12] N. Verma, H. Jia, H. Valavi, Y. Tang, M. Ozatay, L. Y. Chen, B. Zhang, and P. Deaville, "In-memorycomputing advances and prospects," *IEEE Solid-State Circuits Mag.*, vol. 11, no. 3, pp. 43-55, 2019, doi:10.1109/MSSC.2019.2922889.

[13] L. Chua, "Memristor-The missing circuit element," *IEEE Trans. Circuit Theory.*, vol. 18, no. 5, pp. 507-519, 1971, doi:10.1109/TCT.1971.1083337.

[14] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, "The missing memristor found," *Nature*, vol. 453, no. 7191, pp. 80-83, 2008, doi:10.1038/nature06932.

[15] C. Wan, P. Cai, M. Wang, Y. Qian, W. Huang, and X. Chen, "Artificial sensory memory," *Adv. Mater.*, vol. 32, no. 15, pp. 1902434, 2020, doi:10.1002/adma.201902434.

[16] L. Yin, R. Cheng, Y. Wen, C. Liu, and J. He, "Emerging 2D memory devices for in-memory computing," *Adv. Mater.*, vol. 33, no. 29, pp. 2007081, 2021, doi:10.1002/adma.202007081.

[17] Z. Dong, X. Ji, G. Zhou, M. Gao and D. Qi, "Multimodal Neuromorphic Sensory-Processing System with Memristor Circuits for Smart Home Applications," *IEEE Trans. Ind. Appl.*, 2022, doi: 10.1109/TIA.2022.3188749.

[18] M. Wang, J. Tu, Z. Huang, T. Wang, Z. Liu, F. Zhang, W. Li, K. He, L. Pan, X. Zhang, X. Feng, Q. Liu, M. Liu, and X. Chen, "Tactile near-sensor analogue computing for ultrafast responsive artificial skin," *Adv. Mater.*, pp. 2201962, 2022, doi:10.1002/adma.202201962.

[19] Y. Liu, E. Li, X. Wang, Q. Chen, Y. Zhou, Y. Hu, G. Chen, H. Chen, and T. Guo, "Self-powered artificial auditory pathway for intelligent neuromorphic computing and sound detection," *Nano Energy*, vol. 78, pp. 105403, 2020, doi:10.1016/j.nanoen.2020.105403.

[20] Z. Gao, S. Chen, R. Li, Z. Lou, W. Han, K. Jiang, F. Qu, and G. Shen, "An artificial olfactory system with sensing, memory and self-protection capabilities," *Nano Energy*, vol. 86, pp. 106078, 2021, doi:10.1016/j.nanoen.2021.106078.

[21] L. C. Antaket, and Y. Kashimori, "Neural mechanisms of the maintenance and manipulation of gustatory working memory in orbitofrontal cortex," *Cognit. Comput.*, pp. 1-9, 2022, doi:10.1007/s12559-022-10035-1.

[22] C. Paniagua, J. Eliasson, and J. Delsing, "Efficient device-to-device service invocation using arrowhead orchestration," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 429-439, 2020, doi:10.1109/JIOT.2019.2952697.

[23] D. Ielmini, and G. Pedretti, "Device and circuit architectures for in-memory computing," *Adv. Intell. Syst.*, vol. 2, no. 7, pp. 2000040, 2020, doi:10.1002/aisy.202000040.

[24] X. Ji, Z. Dong, C. S. Lai, and D. Qi, "A brain-inspired in-memory computing system for neuronal communication via memristive circuits," *IEEE Commun. Mag.*, vol. 60, no. 1, pp. 100-106, 2022, doi:10.1109/MCOM.001.21664.

[25] G. Lan, J. Yang, R. P. Ye, Y. Boyjoo, J. Liang, X. Liu, Y. Li, J. Liu, and K. Qian, "Sustainable carbon materials toward emerging applications," *Small Methods*, vol. 5, no. 5, pp. 2001250, 2021, doi:10.1002/smtd.202001250.

[26] S. Zhang, X. Xia, F. Li, C. Chen, and L. Zhao, "Study on visual and auditory perception characteristics of children with different type of mathematics learning disability," *Int. J. Disabil. Dev. Educ.*, vol. 68, no. 1, pp. 78-94, 2021, doi:10.1080/1034912X.2019.1634248.

[27] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint*, pp. arXiv:1810.04805, 2019.

[28] G. Van Houdt, C. Mosquera, and G. Nápoles, "A review on the long short-term memory model," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5929-5955, 2020, doi:10.1007/s10462-020-09838-1.

[29] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on Convolutional Neural Networks (CNN) in vegetation remote sensing," *ISPRS J. Photogramm. Remote Sens.*, vol. 173, pp. 24-49, 2021, doi:10.1016/j.isprsjprs.2020.12.010.

[30] Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: multi-head cross attention network for facial expression recognition," *arXiv preprint*, pp. arXiv:2109.07270, 2021.

[31] S. Sharda, M. Singh, and K. Sharma, "RSAM: Robust self-attention based multi-horizon model for solar irradiance forecasting," *IEEE Trans. Sustain. Energy*, vol. 12, no. 2, pp. 1394-1405, 2021, doi:10.1109/TSTE.2020.3046098.

[32] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognit.*, vol. 38, no. 12, pp. 2270-2285, 2005, doi:10.1016/j.patcog.2005.01.012.

[33] Z. Hao, T. Zhang, M. Chen, and Z. Kaixu, "RRL: Regional rotate layer in convolutional neural networks," *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 1, pp. 826-833, 2022, doi:10.1609/aaai.v36i1.19964.

[34] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, pp. 5634–5641, 2018, doi:10.1609/aaai.v32i1.12021.

[35] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An attentive RNN for emotion detection in conversations," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, pp. 6818-6825, 2019, doi:10.1609/aaai.v33i01.33016818.

[36] S. Xing, S. Mai, and H. Hu, "Adapted dynamic memory network for emotion recognition in conversation," *IEEE Trans. Affect. Comput.*, 2020, doi:10.1109/TAFFC.2020.3005660.

[37] M. Ren, X. Huang, X. Shi, and W. Nie, "Interactive multimodal attention network for emotion recognition in conversation," *ISPL*, vol. 28, pp. 1046-1050, 2021, doi:10.1109/LSP.2021.3078698.

[38] Z. Lian, B. Liu, and J. Tao, "CTNet: Conversational transformer network for emotion recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 985-1000, 2021, doi:10.1109/TASLP.2021.3049898.

[39] W. Li, W. Shao, S. Ji, and E. Cambria, "BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis," *Neurocomputing*, vol. 467, pp. 73-82, 2022, doi:10.1016/j.neucom.2021.09.057.

[40] H. Ma, J. Wang, H. Lin, X. Pan, Y. Zhang, and Z. Yang, "A multi-view network for real-time emotion recognition in conversations," *Knowledge-Based Syst.*, vol. 236, pp. 107751, 2022, doi:10.1016/j.knosys.2021.107751.

[41] F. Chen, J. Shao, A. Zhu, D. Ouyang, X. Liu, and H. T. Shen, "Modeling hierarchical uncertainty for multimodal emotion recognition in conversation," *IEEE Trans. Cybern.*, pp. 1-12, 2022, doi:10.1109/TCYB.2022.3185119.

[42] W. Jiao, M. Lyu, and I. King, "Real-time emotion recognition via attention gated hierarchical memory network," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 5, pp. 8002-8009, 2020, doi:10.1609/aaai.v34i05.6309.

[43] Y. H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L. P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," *Proc. Conf. Assoc. Comput. Linguist. Mtg.*, vol. 2019, pp. 6558-6569, 2019, doi: 10.18653/v1/p19-1656.

[44] Z. Lian, B. Liu and J. Tao, "SMIN: Semi-supervised Multi-modal Interaction Network for Conversational Emotion Recognition," *IEEE Trans. Affect. Comput.*, 2022, doi: 10.1109/TAFFC.2022.3141237.

[45] Y. Zhang, J. Wang, Y. Liu, L. Rong, Q. Zheng, D. Song, P. Tiwari,and, J. Qin, "A multitask learning model for multimodal sarcasm, sentiment and emotion recognition in conversations," *Inf. Fusion*, vol. 93, pp. 282-301, 2023, doi: 10.1016/j.inffus.2023.01.005.

[46] F. Chen, Z. Sun, D. Ouyang, X. Liu, and J. Shao, "Learning what and when to drop: Adaptive multimodal and contextual dynamics for emotion recognition in conversation," *Proc. ACM Int. Conf. Multimedia*, pp. 1064-1073, 2021, doi: 10.1145/3474085.3475661.

[47] J. Wen, D. Jiang, G. Tu, C. Liu, and E. Cambria, "Dynamic interactive multiview memory network for emotion recognition in conversation," *Inf. Fusion*, vol. 91, pp. 123-133, 2023, doi:10.1016/j.inffus.2022.10.009.

[48] P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang, and, H. Qian, "Fully hardware-implemented memristor convolutional neural network," *Nature*, vol. 577, pp. 641-646, 2020, doi: 10.1038/s41586-020-1942-4.

[49] S. Jung, H. Lee, S. Myung, H. Kim, S. K. Yoon, S. W. Kwon, ... and, S. J. Kim, "A crossbar array of magnetoresistive memory devices for in-memory computing," *Nature*, vol. 601, pp. 211-216, 2022, doi: 10.1038/s41586-021-04196-6.

[50] S. Wen, H. Wei, Y. Yang, Z. Guo, Z. Zeng, T. Huang, and, Y. Chen, "Memristive LSTM network for sentiment analysis," *IEEE Trans. Syst.*

Man Cybern. Syst., vol. 51, pp. 1794-1804, 2019, doi: 10.1109/TSMC.2019.2906098.

[51] Z. Wang, C. Li, P. Lin, M. Rao, Y. Nie, W. Song, ... and, J. J. Yang, "In situ training of feed-forward and recurrent convolutional memristor networks," *Nat. Mach. Intell.*, vol. 1, pp. 434-442, 2019, doi: 10.1038/s42256-019-0089-1.

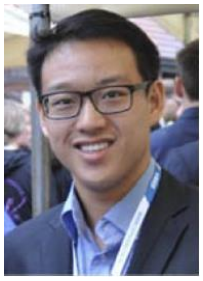
[52] C. Yang, X. Wang, and, Z. Zeng, "Full-circuit implementation of transformer network based on memristor," *IEEE Trans. Circuits Syst. I Regul. Pap.*, vol. 64, pp. 1395-1407, 2022, doi: 10.1109/TCSI.2021.3136355.

[53] W. Wan, R. Kubendran, C. Schaefer, S. B. Eryilmaz, W. Zhang, D. Wu, ... and, G. Cauwenberghs, "A compute-in-memory chip based on resistive random-access memory," *Nature*, vol. 608, pp. 504-512, 2022, doi: 10.1038/s41586-022-04992-8.

Xiaoyue Ji (Student Member, IEEE) received the B.E. degree in electronics and information engineering in 2016 degree from the School of Electrical Engineering, Harbin Engineering University, China. She is currently working toward the Ph.D. degree in control theory and control engineering from the School of Electrical Engineering, Zhejiang University, China. Her research interests cover memristor and memristive system, artificial neural network, the design and analysis of nonlinear systems based on memristor and computer simulation.

Zhekang Dong (Senior Member, IEEE) received the B.E. and M.E. degrees in electronics and information engineering in 2012 and 2015, respectively, from Southwest University, Chongqing, China. He received the Ph.D. degree from the School of Electrical Engineering, Zhejiang University, China, in 2019. Currently, he is an associate professor in Hangzhou Dianzi University, Hangzhou, China. He is also a Research Assistant (Joint-Supervision) at The Hong Kong Polytechnic University. His research interests cover memristor and memristive system, artificial neural network, the design and analysis of nonlinear systems based on memristor and computer simulation.

Yifeng Han received the B.E. degrees in automation in 2017 from Zhejiang University. He received the M.E. degrees in control system from Imperial College London in 2018. Currently, he is studying for the Ph.D. degree in Electrical engineering from Zhejiang University. His current research interests cover image processing and data analysis.



Chun Sing Lai (Senior Member, IEEE) received the BEng in electronic and electrical engineering from Brunel University London, UK, and DPhil in engineering science from the University of Oxford, UK in 2013 and 2019, respectively. Dr Lai is currently a Lecturer at the Department of Electronic and Electrical Engineering, Brunel University London, UK. His current interests are in data analytics, power system optimization, energy system modelling, and energy economics for low carbon energy networks and energy storage systems.

Donglian Qi (Senior Member, IEEE) received the Ph.D. degree in control theory and control engineering from the School of Electrical Engineering, Zhejiang University, China, in 2002. She is currently a Full Professor and a Ph.D. Advisor with Zhejiang University. Her recent research interest covers intelligent information processing, chaos system, and nonlinear theory and application.