Enhancing Multimodal Emotion Recognition with Swin Transformers and Dynamic Weighted Gating for Federated Learning on Edge Devices

Xuan-Phuc Phan-Nguyen[1*], Khoi Tran-Minh[2*], and Nga Ly-Tu[2,3]

***Abstract*— In this paper, we present a deep learning pipeline for multimodal emotion recognition that utilizes Swin Transformers and Dynamic Weighted Gating (SWIN-DWG), where the encoder's mechanism is designed sequentially to alternate between cross-attention and self-attention, thereby enhancing the ability to learn multimodal features from the EEG-Audio-Video (EAV) dataset and achieving a mean accuracy of 79.24%, which represents an 8.38% improvement over a recent pipeline. Next, we propose a non-independent and identically distributed (non-IID) multimodal data partitioning criteria, specifically for the EAV dataset. In addition, with our FedSAP framework of three methods, including FedAvg, FedBN, and FedPer, we illustrate metric improvements compared to the traditional federated learning framework in terms of multimodal learning through two simulation scenarios. Finally, another two deployment cases were conducted on four different NVIDIA edge devices to test the learning capabilities with both personalization and missing modality.**

I. INTRODUCTION

Swin Transformer debuted in 2021 and originally served as a general-purpose vision backbone to introduce a hierarchical representation for vision and improve scalability to high-resolution images [12]. In recent years, while Swin was not designed for multimodal tasks, researchers have adopted measures to enhance breast tumor segmentation using ultrasound, mammography, and magnetic resonance imaging [16]; integrate fingerprints, iris, and electroencephalogram (EEG) into biometric recognition systems to improve security and mitigate spoof attacks [7]; or utilize EEG in multimodal systems to aid emotion recognition tasks [18]. Similarly, federated learning (FL), developed by Google in 2016, showed the same growing trend; especially multimodal federated learning (MMFL), it demonstrated advances in various medical fields such as reconstruction and synthesis [17], segmentation [6], and recognition of human actions [8]. The most powerful federated learning frameworks, namely NVFlare [15] and FlowerAI [2], are primarily utilized for unimodal learning, while the implementation of federated learning to maintain privacy preservation and personalization on real edge devices has not yet been widely explored by existing multimodal methods. Thus, we would like to experiment and analyze the performance of federated learning frameworks in multimodal learning settings using the following contributions. *Firstly*, we design a pipeline to generate multimodal embeddings in the EAV dataset using Swin Transformer architectures and Dynamic Weighted Gating (DWG) to improve the accuracy of emotion classification compared to the existing pipeline called EEG-based multimodal representation learning (EMRL) [18]. *Secondly*, to create non-iid partitions for the EAV dataset, which has not been previously examined, a 4-step criterion was defined in Section III-B. *Thirdly*, we implemented four MMFL scenarios: two for client personalization and two for FL training with FedSAP and NVFlare on heterogeneous edge devices, including RTX 4060/4070 laptops and Jetson Orin Nano systems. *Finally*, to demonstrate the predictions of our trained model, we designed an interactive graphic user interface (GUI) shown in Fig. 5.

II. LITERATURE REVIEW

A. Multimodal Emotion Recognition

Recent advances in multimodal emotion recognition have focused primarily on exploiting the supporting relationship between audio and visual modalities. One notable example starts with a multimodal method that utilized convolutional neural networks (CNNs) to obtain an accuracy of 95.0% on the BAUM-1 dataset and up to 95.95% on the RAVDESS dataset, which had demonstrated significant improvements over previous methods [3]. In another research, the authors illustrated the efficiency of a trimodal system using the E-Branchformer network, attaining an F1-Macro score of 81.4% and an F1-Micro score of 85.3% in the CREMA-D corpus while simultaneously reducing the parameter size by 33.3% compared to the VAVL baseline [5].

Nevertheless, these bimodal approaches with just vision and audio have a limitation of neglecting the rich emotional cues embedded in body posture that affect the accuracy of emotion recognition in complex scenarios such as subtle expressions or acting. To address this challenge, a trimodal dataset called EEG-Audio-Video (EAV) was built to leverage the strength of EEG signals to identify complicated interactions within associated brain activities [10]. Then a new approach to multimodal emotion recognition based on attention was introduced, which integrates EEG as a third modality in conjunction with audio and visual, later introduced to achieve an accuracy of 70.86%, averaging more than 42 subjects in the EAV dataset [18]. This shows

[1]Xuan-Phuc Phan-Nguyen is with School of Mechanical and Mining Engineering, The University of Queensland, Brisbane, Australia phuc.phan@uq.edu.au

[2]Khoi Tran-Minh is with School of Computer Science & Engineering, International University, Ho Chi Minh City, Vietnam tranminhkhoi565@gmail.com

[3]Nga Ly-Tu is with both School of Computer Science & Engineering, International University and Vietnam National University, Ho Chi Minh City, Vietnam ltnga@hcmiu.edu.vn

*Both authors share the same contribution to this research and we also sincerely appreciate Phu-Quoc Pham's initial contribution as well as the previous conference paper.

potential room for future refinement due to its simplicity in implementation [9].

B. Federated Learning (FL)

Federated learning provides a crucial layer of privacy protection for multimodal sentiment-aware systems by enabling model training without centralizing raw data. Starting with FedAvg, which has served as a baseline FL method by providing a communication-efficient framework for training deep networks on decentralized data by performing multiple local updates on clients and then averaging the parameters on a central server [13]. Later, FedBN came to address the challenge of feature shift in non-IID data by keeping batch normalization layers strictly local and unaggregated, harmonizing local data distributions and resulting in faster and more stable convergence than standard averaging methods [11], while FedPer was made to combat statistical heterogeneity in personalization tasks by leveraging a base with a personalization layer approach, where shared base layers capture general features while private personalization layers adapt to individual user preferences and data [1].

Although classical approaches remain widely used, multimodal emotion recognition in federated learning remains underexplored. Existing methods often employ resource-intensive attention mechanisms and lack reliability modeling, leading to unrealistic assumptions of homogeneous multimodal data.

C. FL Frameworks

NVFLare is an open-source FL software framework developed by NVIDIA, aimed at deploying federated learning systems in distributed environments with a high level of security [15], particularly suitable for data-sensitive fields such as healthcare and finance. Meanwhile, Flower AI is notable for its ability to simplify the deployment and scale of FL experiments on various hardware and software platforms [2]. In addition, our previous work introduced FedSAP, a multitask federated learning framework supporting face recognition, facial emotion recognition, and RFID classification across up to 20 edge devices [14]. However, its application to multimodal learning remains unexplored.

III. METHODOLOGY

A. Proposed Workflow

To fill the identified research gap, we propose a workflow illustrated in Fig. 1 for deep learning and adapt existing federated learning methods for multimodal contexts.

Starting with the multimodal EAV dataset, comprising three components, namely vision, audio, and EEG [18], which are passed through two separate pipelines EMRL in purple and our SWIN-DWG in green, for feature extraction and embedding generation. Specifically, we reused all the setups in the EMRL pipeline to first extract raw features using Vision Transformer (ViT), Audio Spectrogram Transformer (AST) and EEGFormer, which were later concatenated before being fused by a self-attention network [18].

On the other hand, we came up with a different approach by using Swin Transformers as the core of our embedding pipeline, in which vision and audio data were encoded by the pretrained Microsoft's tiny-sized Swin Transformer (Swin-T) [12] and the Hierarchical Token Semantic Audio Transformer (HTS-AT) [4], respectively, whereas we developed a custom 1D Swin-based network for EEG shown in Fig. 2. The Swin-1D architecture processes EEG signals through a convolutional embedding layer followed by two hierarchical Swin Transformer stages employing alternating window multi-head self-attention (W-MSA) and shifted window multi-head self-attention (SW-MSA) blocks. A patch-merging layer performs temporal downsampling while increasing the feature dimension. Global average pooling and a two-layer classifier generate predictions for the five emotion categories. Training is performed in two phases: head-only optimization followed by full-network fine-tuning with a reduced learning rate.

Next, instead of concatenation, the multimodal features are projected and split into two paths: one path is responsible for aggregating the global context t'_{cls} and computing dynamic weights projected to a size of 512, while the other path handles the weighted gating process with a dimension projection of 256.

In particular, on path 1, from the features h_v, h_a, h_e, we make them act as keys K and values V for each modality. Then, we initialize the learnable parameters that serve as queries Q. Together, they went through a cross-attention network, which serves as a feature compression and filtering mechanism, condensing high-dimensional raw features into a compact and unified representation while suppressing noise:

$$z_m = \mathrm{CrossAttn}(Q_m, K_m, V_m), \quad m \in \{v, a, e\} \tag{1}$$

After that, a context token t_{cls} was also initialized as learnable parameters; the summarized feature vectors z_m are then concatenated with this token and passed through self-attention, functioning as a multimodal fusion and reasoning engine to enable deep interaction between heterogeneous tokens to aggregate a holistic global context vector t'_{cls} that guides the subsequent dynamic gating process:

$$[t'_{cls}] = \mathrm{SelfAttn}([z_v, z_a, z_e, t_{cls}]) \tag{2}$$

Finally, the refined global context vector t'_{cls} progressed through the Dynamic Weight Computation block as follows:

$$[w_v, w_a, w_e] = \mathrm{Softmax}(W_2 \cdot \mathrm{GELU}(W_1 \cdot t'_{cls} + b_1) + b_2) \tag{3}$$

where W_1, W_2 are learnable weight matrices, and $w_m \in (0, 1)$ represents the reliability of the modality m.

Meanwhile, for path 2, we implemented the element-wise product (Hadamard) using the computed weights w_m from path 1 to modulate the projected encoder features:

$$\tilde{h}_m = w_m \cdot \mathrm{Proj}(h_m) \tag{4}$$

Subsequently, we concatenate the global context vector t'_{cls} from path 1 with the weighted features $\tilde{h}_m$ from path 2, ensuring that both local specifics and global context are preserved:

$$E_{final} = \mathrm{Concat}(\tilde{h}_v, \tilde{h}_a, \tilde{h}_e, t'_{cls}) \tag{5}$$

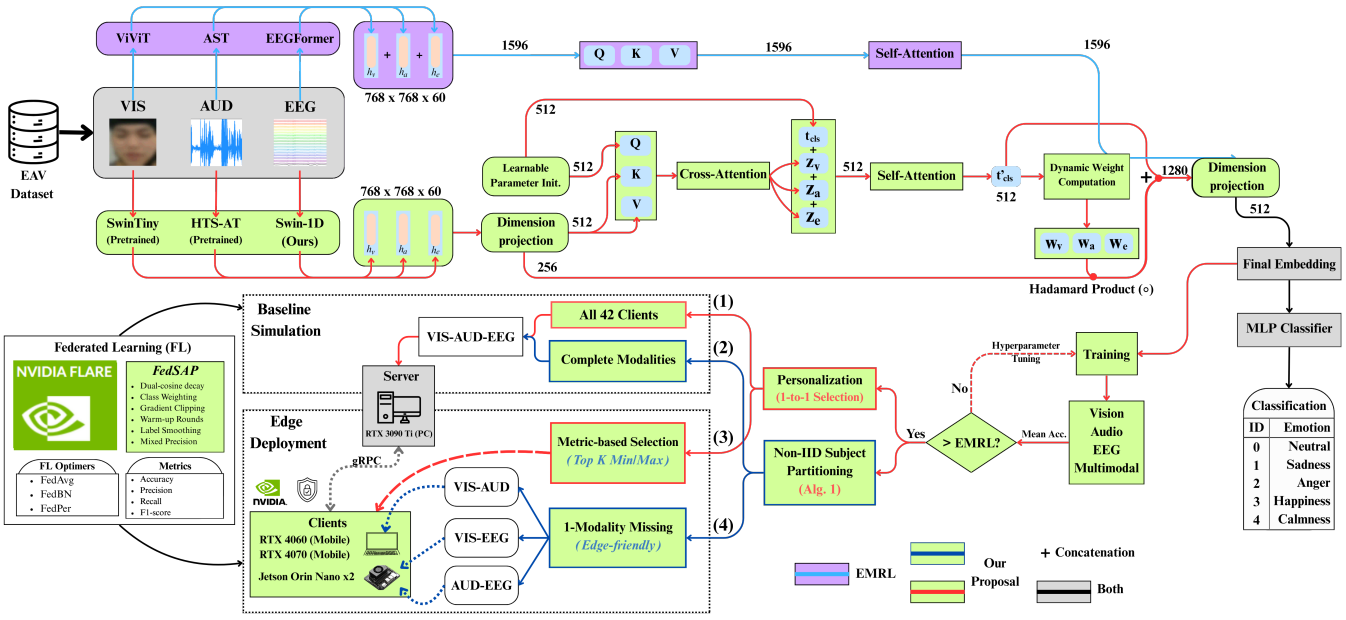

Fig. 1. A comprehensive workflow for EMRL/SWIN-DWG and federated learning pipelines

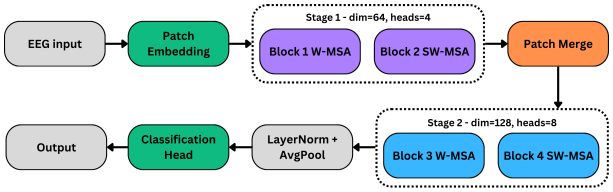

Fig. 2. Our Swin-EEG's model architecture

Then, they are projected to a dimension size of 512. Finally, for the DL workflow, the final embedding E_{final} then serves as input to a Multi-Layer Perceptron (MLP) classifier to predict the emotions, in particular neutral, sadness, anger, happiness, and calmness:

$$\hat{y} = \text{MLP}_{classifier}(Proj(E_{final})) \tag{6}$$

On the other hand, for the FL workflow, we use the final embeddings in the FL training process. The model was trained with continuous hyperparameter tuning to such an extent that the mean multimodal accuracy outperformed the ERML. Subsequently, we developed FL for personalization without modifying the data, including Case 1 – simulated on RTX 3090 Ti PC and Case 3 – deployed on edge devices with the top 4 min and top 4 max subjects in terms of accuracy shown in Table II, using a PC with an RTX 3090 Ti as a server and 4 clients, specifically RTX 4060 and 4070 laptops and 2 Jetson Orin Nanos. In addition, the data was partitioned using the methodology illustrated in Section III-B to simulate Case 2 – simulation with 6 clients with all 3 modalities and Case 4 – edge deployment with 4 clients in conditions lacking one modality. We strategically select the 2-modality data according to the device's capability; for instance, RTX 4060/4070 laptops with higher memory handle heavier tasks like vision-audio, while Jetson Orin Nanos, which have only 8GB of shared RAM, prioritize less computationally expensive ones such as vision-EEG or audio-EEG. Ultimately, we implemented FL scenarios using the NVFlare [15] and FedSAP [14] frameworks with 3 FL methods, namely FedAvg, FedBN, and FedPer with adjustments for multimodal settings. For efficient communication over the Internet during edge deployment, we use gRPC Remote Procedure Calls (gRPC) as a communication protocol.

B. Proposed Non-IID Criteria for EAV Dataset

To prepare multimodal datasets for our FL simulation and edge deployment scenarios, we propose a 4-step non-IID partitioning criterion tailored for the EAV dataset. While the Dirichlet distribution is one of the most common approaches for simulating non-IID settings in FL, it only controls label skewness across clients via a concentration parameter and does not account for the structural heterogeneity inherent in multimodal data, such as missing modalities or subject-level variation across clients. This makes it insufficient by default for multimodal federated learning scenarios. To better reflect real-world conditions, we instead build on our previous work [14] by randomly assigning a varying number of subjects per client, naturally introducing sample and subject count imbalance. We then randomly remove 1–2 emotions per subject and drop 1 modality per client, simulating practical situations where data absence may arise from privacy consent or collection difficulties. Finally, we apply a 0%–10% label scarcity to the remaining emotions in each client to further diversify the label distribution across clients.

IV. Implements and Results

A. Setup Environment

First, we configured four edge devices to perform the experiments, as shown in Table I. Similarly, the training

hyperparameters for NVFLare and FedSAP were set for both EMRL and our pipeline to achieve performance as shown in Tables IV and V. Specifically, the learning rate was set to 0.0003, the batch size to 128, training for 50 rounds with 5 epochs per round, and a weight decay coefficient of 0.0001. For FedSAP, we reused the optimizations presented in previous work, in particular dual-cosine decay, class weighting, gradient clipping, and warmup rounds [14], while adding label smoothing and mixed precision to boost generalization and memory efficiency (Fig. 1).

B. Dataset Preparation & Analysis

The preliminary dataset was pre-divided into training and testing sets in a $70:30$ ratio with 42 subjects having a balanced distribution, totaling $16,800$ samples with 400 samples per subject [10]. The features were then processed through the EMRL and Swin-DWG pipelines to create the IID FL dataset, which was used for cases 1 and 3. Subsequently, all 42 subjects leveraged our criteria presented in Section III-B to create non-IID partitions, which was an adaptation of [14] for multimodal settings, forming two scenarios: three modalities vision, audio, and EEG (Case 2); and two modalities with mixed pairs (Case 4). In Case 2 shown in Fig. 3, six clients owned training samples between 840 and 2063, representing $10.07\% - 24.83\%$, with $5-10$ subjects per client, and the emotion distribution ranged from $6.67\% - 33.33\%$. In Case 4, on the other hand, the subjects were distributed among fewer clients, resulting in a higher concentration of subjects per client, with proportions of the training sample ranging from $19.16\% - 32.94\%$ and intensity of emotions ranging from $14.55\% - 25.00\%$ demonstrated in Fig. 4.

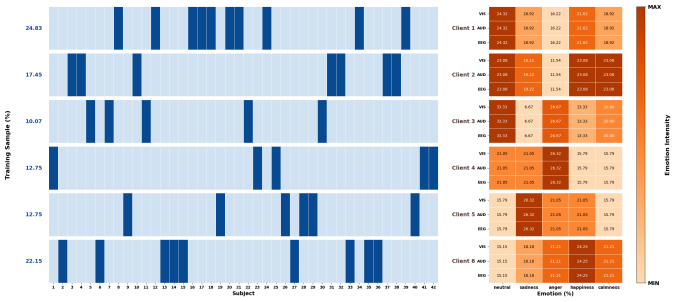

Fig. 3. Training sample and subject proportion (left) and Emotion Intensity (right) of training sample across 6 clients with 3 modalities (Case 2)

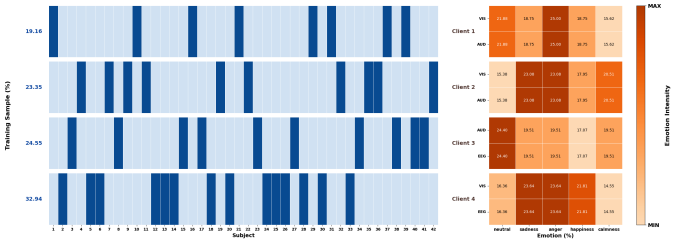

Fig. 4. Training sample and subject proportion (left) and Emotion Intensity (right) of training sample across 4 clients with 2 modalities (Case 4)

C. Performance Comparison of SWIN-DWG and EMRL

The overall multimodal mean accuracy for emotion recognition, as shown in Table II, is 79.24% on average, representing an increase of 8.38% compared to the EMRL pipeline, which had an accuracy of 70.86% [18]. The top 4 subject with lowest accuracy are subject 34, 37, 41 and 35, and the top 4 subjects with the highest accuracy are subject 17, 20, 24, and 4.

To understand what drives SWIN-DWG's performance, we ran a series of ablation experiments summarized in Table III. We first reproduced EMRL [9] as our baseline, which was originally reported 70.86% but dropped to 68.82% when we had to reduce batch sizes from 32/32/64 to 8/32/32 (audio/EEG/vision) due to GPU memory constraints. Our SWIN-DWG under its default configuration reached 70.59%, and with the optimal batch sizes of 32/64/256, it further improved to 79.24%. To see where the gains come from, we swapped encoders between the two pipelines. Using EMRL's encoders in SWIN-DWG noticeably hurt performance to 67.40%, while equipping EMRL with Swin-T, HTS-AT, and EEGFormer or our Swin-EEG raised it to 77.16% and 78.47%, confirming that Swin-based encoder choices are a key factor. We then swapped attention mechanisms: adding our cross/self-attention to EMRL gave a modest gain to 70.00%, while applying EMRL's self-attention to SWIN-DWG still held at 78.37%, largely preserved by the DWG module. Overall, these results suggest that SWIN encoders and DWG mechanisms are the two main drivers of SWIN-DWG's performance.

In IID setting with 42 subjects (Case 1, Table IV), FedAvg, FedPer, and FedBN under the FedSAP framework achieved a higher accuracy than those implemented in NVFLare. Specifically, with FedAvg, the difference was approximately 4.93% and 6.10% for EMRL and SWIN-DWG, respectively; FedPer showed 4.90%, 4.89% respectively and FedBN exhibited the largest gap, reaching 8.61% and 6.13% for EMRL and SWIN-DWG, respectively. Among the three algorithms, FedBN created the largest gap between the FedSAP and NVFLare frameworks, while FedPer showed the smallest difference.

On the other hand, in the non-IID setting with six clients and complete modalities (Case 2, Table IV), the performance gap between the two frameworks narrows. The spread of FedAvg narrowed to approximately $3.16\% - 3.87\%$, FedPer to $1.31\% - 1.87\%$, and FedBN remained within the range of $1.66\% - 3.53\%$. Unlike case 1, among three algorithms, FedAvg created the largest gap between two frameworks, while FedBN created the smallest gap.

Overall, in both cases 1 and 2, the FedSAP framework surpassed NVFlare performance in both the EMRL and SWIN-DWG pipelines on 3 the algorithms FedAvg, FedBN and FedPer, with the gap spread from 1.31% to 8.61%. Between the 2 pipelines, our SWIN-DWG also demonstrated better learning effectiveness, convinced by the increase from 0.46% to 4.23%.

Table V compares the performance of NVFlare and Fed-

TABLE I

EDGE DEVICE CONFIGURATION

Device	RAM	VRAM	Architecture	Python	CC	CUDA	CuDNN	Driver
Jetson Orin Nano	8GB	8GB	Ampere	3.8	8.7	11.4	8.6	Jet. 5.1.2
RTX 4060 (Laptop)	16GB	8GB	Ada Lov.	3.13	8.9	13.0	9.13	581.57
RTX 4070 (Laptop)	32GB	8GB	Ada Lov.	3.13	8.9	13.0	9.13	581.57
RTX 3090Ti (PC)	64GB	24GB	Ampere	3.13	8.6	13.0	9.13	581.57

TABLE II

DL MULTIMODAL ACCURACY ACCROSS 42 SUBJECT

Subject	**Acc. (%)**	**Subject**	**Acc. (%)**	**Subject**	**Acc. (%)**
1	79.23	15	82.03	29	71.03
2	87.30	16	70.40	30	83.77
3	78.47	17	95.97	31	76.30
4	89.70	18	87.80	32	77.63
5	71.20	19	81.57	33	77.80
6	89.20	20	91.17	34	60.77
7	88.07	21	83.30	35	67.40
8	77.40	22	86.40	36	81.97
9	78.60	23	84.53	37	60.93
10	69.50	24	90.97	38	85.13
11	75.63	25	72.87	39	82.90
12	72.43	26	80.00	40	70.17
13	86.93	27	82.73	41	66.80
14	69.83	28	77.33	42	85.07
				AVG	**79.24**

TABLE III

ABLATION STUDIES FOR THIS RESEARCH

Experiments	**Acc. (%)**
EMRL (Reproduced)	68.82
SWIN-DWG (Default configuration [9])	70.59
SWIN-DWG with EMRL's encoders	67.40
EMRL with Swin-T, HTS-AT, *EEGFormer*	77.16
EMRL with Swin-T, HTS-AT, *our Swin-EEG*	78.47
EMRL with SWIN-DWG's attention mechanisms	70.00
SWIN-DWG with EMRL's self-attention	78.37

TABLE IV

SIMULATION ACCURACY ON NVFLARE AND FEDSAP (%)

Case	**Alg.**	**EMRL**		**SWIN-DWG**	
		NVFLARE	**FedSAP**	**NVFLARE**	**FedSAP**
1	FedAvg	63.95	68.88	65.03	71.13
	FedPer	65.87	70.77	69.24	74.13
	FedBN	64.60	73.21	67.39	73.52
2	FedAvg	63.33	66.49	63.79	67.66
	FedPer	61.93	63.24	65.32	67.19
	FedBN	61.76	63.42	64.12	67.65

TABLE V

EDGE DEPLOYMENTS: NVFLARE AND FEDSAP PERFORMANCE (%)

Case	**Alg.**	**NVFLARE**		**FedSAP**	
		Acc.	**F1**	**Acc.**	**F1**
3 ($\text{Min}_{k=4}$)	FedAvg	64.05	63.45	64.30	64.54
	FedPer	64.72	63.54	64.73	65.64
	FedBN	64.80	63.63	65.28	66.17
3 ($\text{Max}_{k=4}$)	FedAvg	67.12	67.34	67.27	68.50
	FedPer	67.75	67.13	67.55	69.69
	FedBN	69.04	68.71	67.82	70.57
4	FedAvg	61.58	61.48	67.27	67.27
	FedPer	67.27	67.14	67.67	67.95
	FedBN	67.23	67.07	67.88	68.37

SAP across Cases 3 and 4 using FedAvg, FedPer, and FedBN. Overall, FedSAP consistently achieved higher F1-scores than NVFlare in all experiments, while also providing competitive or superior accuracy. In Case 3 ($\text{Min}_k = 4$), FedSAP improved the accuracy of FedAvg, FedPer, and FedBN by 0.25%, 0.01%, and 0.48%, respectively, with corresponding F1-score gains of 1.09%, 2.10%, and 2.54%. Similarly, in Case 3 ($\text{Max}_k = 4$), FedSAP increased the F1-score of FedAvg, FedPer, and FedBN from 67.34%, 67.13%, and 68.71% to 68.50%, 69.69%, and 70.57%, respectively.

The largest improvement was observed in Case 4, where FedSAP significantly outperformed NVFlare across all algorithms. For FedAvg, accuracy increased from 61.58% to 67.27%, while F1-score improved from 61.48% to 67.27%. FedPer and FedBN also showed consistent gains, reaching 67.67% and 67.88% accuracy, respectively. Among all configurations, FedSAP with FedBN achieved the highest overall accuracy of 67.88% and the highest F1-score of 70.57%. These results indicate that FedSAP provides stronger and more balanced performance than NVFlare, particularly under heterogeneous multimodal federated learning conditions.

D. Emotion Recognition GUI

Fig. 5 shows the GUI of the multimodal emotion recognition system. Users can visualize image, audio, and EEG inputs, play audio samples, and compare individual model predictions with the final fused result. Prediction errors may stem from limitations of the EAV dataset, such as inconsistent emotional expression, non-native English speech, and facial occlusion from EEG equipment [10].

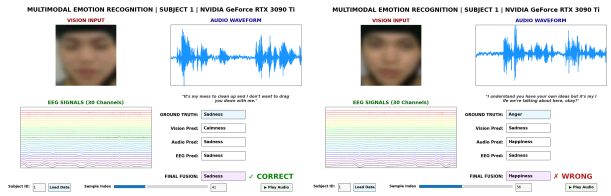

Fig. 5. Subject 1's Correct (L) & Wrong (R) predictions via our GUI

V. CONCLUSIONS

We proposed SWIN-DWG, a multimodal framework that improved mean accuracy by 8.38% over EMRL on the EAV dataset. Our optimized FedSAP framework also improved NVFLare, achieving accuracy gains of 1.31%–8.61% across four federated learning scenarios, including IID, non-IID, and missing-modality settings. Despite these promising results, challenges remain, including overfitting, limited deployment on four edge devices, a basic GUI implementation,

and evaluation of only three FL algorithms (FedAvg, FedPer, and FedBN). Future work will focus on addressing these limitations and expanding the framework's scalability and robustness.

REFERENCES

[1] Arivazhagan, M.G., Aggarwal, V., Singh, A., Choudhary, S.: Federated learning with personalization layers (12 2019). https://doi.org/10.48550/arXiv.1912.00818

[2] Beutel, D.J., Topal, T., Mathur, A., Qiu, X., Fernandez-Marques, J., Gao, Y., Sani, L., Li, K.H., Parcollet, T., de Gusmão, P.P.B., Lane, N.D.: Flower: A friendly federated learning research framework (2020). https://doi.org/10.48550/ARXIV.2007.14390, https://arxiv.org/abs/2007.14390

[3] Bilotti, U., Bisogni, C., De Marsico, M., Tramonte, S.: Multimodal emotion recognition via convolutional neural networks: Comparison of different strategies on two multimodal datasets. Engineering Applications of Artificial Intelligence **130**, 107708 (Apr 2024). https://doi.org/10.1016/j.engappai.2023.107708, http://dx.doi.org/10.1016/j.engappai.2023.107708

[4] Chen, K., Du, X., Zhu, B., Ma, Z., Berg-Kirkpatrick, T., Dubnov, S.: Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). p. 646–650. IEEE (May 2022). https://doi.org/10.1109/icassp43922.2022.9746312, http://dx.doi.org/10.1109/ICASSP43922.2022.9746312

[5] Fan, J., Wang, Y.: Audio-visual learning for multimodal emotion recognition. Symmetry **17**(3), 418 (2025). https://doi.org/10.3390/sym17030418

[6] Gao, Y., Gong, M., Ong, Y.S., Qin, A.K., Wu, Y., Xie, F.: A collaborative multimodal learning-based framework for covid-19 diagnosis. IEEE Transactions on Neural Networks and Learning Systems **35**(11), 15883–15895 (Nov 2024). https://doi.org/10.1109/tnnls.2023.3290188, http://dx.doi.org/10.1109/TNNLS.2023.3290188

[7] Garg, R., Pathak, P., Singh, M.P.: A multimodal biometric recognition system based on fingerprints, iris and ecg via swin transformer and cnn model. Systems and Soft Computing **7**, 200369 (Dec 2025). https://doi.org/10.1016/j.sasc.2025.200369, http://dx.doi.org/10.1016/j.sasc.2025.200369

[8] Hwang, T.H., Shi, J., Lee, K.: Enhancing privacy-preserving personal identification through federated learning with multimodal vital signs data. IEEE Access **11**, 121556–121566 (2023). https://doi.org/10.1109/access.2023.3328641, http://dx.doi.org/10.1109/access.2023.3328641

[9] Kang: bci-winter-2025. https://github.com/Kang1121/bci-winter-2025 (2025)

[10] Lee, M.H., Shomanov, A., Begim, B., Kabidenova, Z., Nyssanbay, A., Yazici, A., Lee, S.W.: Eav: Eeg-audio-video dataset for emotion recognition in conversational contexts. Scientific data **11**(1), 1026 (2024)

[11] Li, X., Jiang, M., Zhang, X., Kamp, M., Dou, Q.: Fedbn: Federated learning on non-iid features via local batch normalization (02 2021). https://doi.org/10.48550/arXiv.2102.07623

[12] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)

[13] McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)

[14] Phan-Nguyen, X.P., Pham, P.Q., Nguyen-Thi-Thanh, T., Ly-Tu, N.: Performance analysis and optimization for a multi-tasking federated learning framework. In: International Conference on Computational Collective Intelligence. pp. 107–121. Springer (2025)

[15] Roth, H.R., Cheng, Y., Wen, Y., Yang, I., Xu, Z., Hsieh, Y.T., Kersten, K., Harouni, A., Zhao, C., Lu, K., Zhang, Z., Li, W., Myronenko, A., Yang, D., Yang, S., Rieke, N., Quraini, A., Chen, C., Xu, D., Ma, N., Dogra, P., Flores, M., Feng, A.: Nvidia flare: Federated learning from simulation to real-world (2022). https://doi.org/10.48550/ARXIV.2210.13291, https://arxiv.org/abs/2210.13291

[16] Wang, J., Ye, W., He, J., Zhang, L., Huang, G., Yu, Z., Liang, Z.: Integrating biological and machine intelligence: Attention mechanisms in brain–computer interfaces. Information Fusion **125**, 103417 (Jan 2026). https://doi.org/10.1016/j.inffus.2025.103417, http://dx.doi.org/10.1016/j.inffus.2025.103417

[17] Yan, Y., Wang, H., Huang, Y., He, N., Zhu, L., Xu, Y., Li, Y., Zheng, Y.: Cross-modal vertical federated learning for mri reconstruction. IEEE Journal of Biomedical and Health Informatics **28**(11), 6384–6394 (Nov 2024). https://doi.org/10.1109/jbhi.2024.3360720, http://dx.doi.org/10.1109/JBHI.2024.3360720

[18] Yin, K., Shin, H.B., Li, D., Lee, S.W.: Eeg-based multimodal representation learning for emotion recognition. In: 2025 13th International Conference on Brain-Computer Interface (BCI). p. 1–4. IEEE (Feb 2025). https://doi.org/10.1109/bci65088.2025.10931743, http://dx.doi.org/10.1109/BCI65088.2025.10931743