NEUROINJECT: BRIDGING BRAIN AND BEHAVIOR WITHOUT PAIRED SUBJECTS VIA INTERMEDIATE CLINICAL-STATE INJECTION FOR DEPRESSION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal depression recognition integrates complementary behavioral and neurophysiological evidence to capture different manifestations of depression. Speech and language reflect behavioral and semantic manifestations, whereas EEG captures neurophysiological activity associated with depressive states. Existing multimodal methods typically fuse or align modality representations under naturally paired observations, whereas EEG and behavioral interview datasets are often collected from independent cohorts without subject- or sample-level correspondence. We study this unpaired setting and propose NeuroInject, which preserves the native audio–text pathway of a frozen multimodal large language model while introducing encoded EEG representations into its intermediate computation. Learnable clinical carriers retrieve modality-specific evidence through a shared cross-attention mechanism, accumulate it into a clinical state across layers, and write it back through gated low-rank mappings. EEG and behavioral samples are processed in independent forward passes, with cross-cohort transfer realized through jointly optimized shared parameters and a unified decision space. Experiments on MODMA, EDRA, CMDC, and DAIC-WOZ show that NeuroInject achieves the highest Accuracy among the compared methods on all four datasets, reaching 71.02, 86.81, 91.19, and 65.78, respectively, while AUROC gains are less consistent. Ablations show that the intermediate interaction contributes most of the improvement over a standalone EEG encoder.

1 INTRODUCTION

Multimodal depression detection aims to integrate complementary information across heterogeneous clinical modalities for more comprehensive assessment than is possible with any single modality (Wang et al., 2025b; Li et al., 2025b). Behavioral signals such as speech and language capture observable manifestations of depression, whereas electroencephalography (EEG) provides complementary neurophysiological information (Li et al., 2025b; Rehman et al., 2025). Jointly modeling externally expressed behavior and endogenous neural activity therefore provides a promising avenue for characterizing the heterogeneous manifestations of depression.

In practice, however, physiological and behavioral data are rarely collected under the paired setting assumed by conventional multimodal learning. EEG and behavioral interview datasets are often acquired from independent clinical cohorts without subject-level, temporal, or instance-level correspondence, preventing the construction of conventional patient-level multimodal samples. Beyond missing correspondence, the two modalities also enter the joint model at different representational stages. EEG records continuous multichannel neurophysiological activity, whose information is distributed across temporal, spatial, and spectral dynamics, whereas speech and language reflect externally expressed behavioral and semantic manifestations that are progressively contextualized within an audio–language model (Li et al., 2025b; Rehman et al., 2025; Chu et al., 2024). Consequently, simply concatenating an already aggregated EEG representation with behavioral tokens at the input layer may mix representations at substantially different levels of abstraction, whereas restricting

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

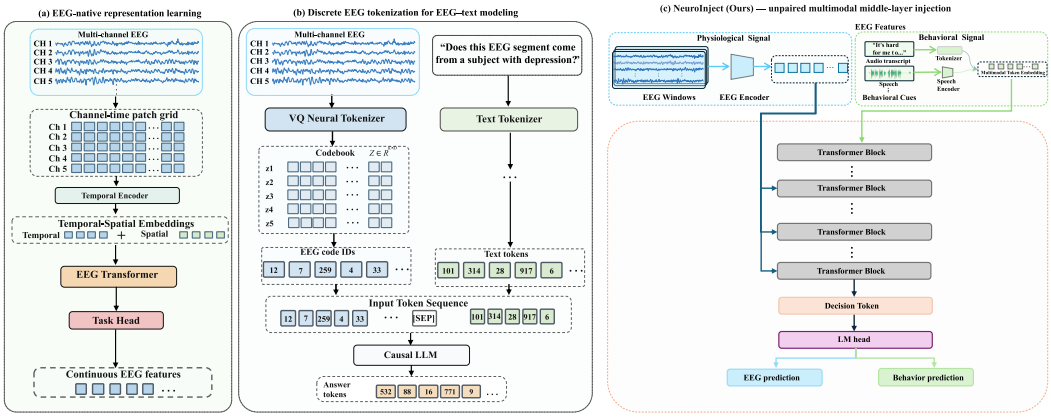


Figure 1: Schematic comparison of three EEG modeling strategies.

their interaction to the final prediction layer prevents physiological information from participating in the intermediate formation of behavioral representations.

This motivates a different formulation of physiological–behavioral multimodal learning. Rather than forcing independently collected modalities into artificial sample pairs or a single input-level representation space, we ask whether heterogeneous clinical information can interact through a shared intermediate computational interface. Such an interface should preserve the native processing structure of each modality, support variable-length and structurally different cues, and enable knowledge learned from independent clinical cohorts to influence a common prediction model without direct sample correspondence. Beyond predictive performance, it is also important to determine whether such interaction gives rise to reproducible internal computational states that are shared across physiological and behavioral tasks.

To address this setting, we propose NeuroInject, an intermediate clinical-state injection framework for unpaired physiological–behavioral learning. NeuroInject preserves the native audio–text pathway of the multimodal large language model while introducing encoded EEG information through learnable clinical carriers at intermediate blocks. EEG and behavioral samples are processed through independent forward passes and are never artificially paired. Instead, cross-cohort knowledge transfer is realized through a shared intermediate clinical interface and jointly optimized high-level adaptation parameters, with both pathways ultimately entering a unified language-model decision space.

The main contributions of this work are summarized as follows:

- We formulate depression recognition across independently collected EEG and behavioral cohorts as unpaired cross-cohort physiological–behavioral learning, explicitly accounting for both missing sample-level correspondence and representational mismatch.
- We introduce NeuroInject, which injects EEG-derived information through a shared intermediate clinical-state interface while preserving the native behavioral processing pathway of the audio–language model, enabling joint learning across unpaired heterogeneous clinical modalities.
- We evaluate NeuroInject across four independent clinical cohorts—MODMA, EDRA, CMDC, and DAIC-WOZ, demonstrating its applicability to both EEG and audio–text depression recognition without sample-level pairing.

2 Related Work

2.1 EEG Representation Learning and Brain–Language Modeling

Recent advances in EEG representation learning have improved the transferability of neural representations across datasets and downstream tasks. Large-scale pretrained models such as LaBraM,

EEGPT, and CBraMod learn transferable spatiotemporal representations from heterogeneous EEG recordings Jiang et al. (2024); Wang et al. (2024; 2025a). More recent foundation models further address the substantial variability of EEG signals across subjects, recording configurations, and temporal–spatial scales, including REVE, CSBrain, LUNA, and Uni-NTFM El Ouahidi et al. (2025); Zhou et al. (2025); Döner et al. (2025); Chen et al. (2026). Related efforts such as MMM and mdJPT also explore cross-dataset generalization under heterogeneous channel configurations and dataset distributions Yi et al. (2023); Zhang et al. (2025). Together, these developments provide increasingly generalizable neural representations for integration with broader foundation-model architectures.

A growing line of work further connects brain representations with language or foundation models. NeuroLM converts EEG into neural tokens for language-model instruction tuning Jiang et al. (2025), while BrainMosaic extracts compositional neural semantics to support natural-language reconstruction Li et al. (2026b). As summarized in Fig. 1, these approaches differ from EEG-native representation learning, whereas NeuroInject introduces encoded EEG information into the intermediate computation of an audio–language model without requiring paired EEG–behavior samples. Beyond neural decoding, recent studies increasingly examine the relationship between biological representations and the internal computation of artificial models. Bidirectional brain–model comparison has been explored to characterize their shared and distinct representations Muzellec & Kar (2026), while brain-guided intervention and interpretable neural decoding investigate how biological signals can influence or reveal model computation Xiao et al. (2026); Feng et al. (2026). Complementary studies further localize task-relevant information and functional organization across model layers Marinescu et al. (2026); Fang et al. (2026); Yu et al. (2026). Our setting extends this direction to independently collected physiological and behavioral cohorts. Instead of relying on paired neural–behavioral observations or treating EEG only as information to be mapped into a semantic space, we investigate how physiological and behavioral information can jointly contribute to shared intermediate computation without sample-level correspondence.

2.2 Behavioral Multimodal Depression Recognition

Behavioral multimodal depression recognition commonly integrates observable cues from speech, language, facial behavior, personality, and interview context to characterize different manifestations of depression. Existing studies have explored cross-modal fusion, modality-specific representation learning, incomplete-modality learning, and individual-aware modeling Hazarika et al. (2020); Han et al. (2021); Pan et al. (2024); Fu et al. (2026). These studies demonstrate that heterogeneous behavioral cues can provide complementary information for depression assessment.

Among these behavioral sources, speech and text constitute a particularly important line of research. Speech captures nonverbal characteristics such as prosody, pauses, energy, and expressive patterns, whereas text directly conveys symptom descriptions, cognitive content, and semantic information, making the two modalities highly complementary in interview-based depression recognition. Ef-Bi-LSTM and MulT jointly model acoustic and linguistic cues through sequential modeling or cross-modal attention Zou et al. (2023); Tsai et al. (2019), while Text-Guided methods and M3L further introduce cross-modal reconstruction, feature decomposition, and multilingual modeling Chen et al. (2025); You et al. (2025). Mental-Perceiver similarly performs mental-state assessment through joint audio–text modeling Qin et al. (2025).

Large language and multimodal language models have further extended this direction. Recent studies incorporate non-textual behavioral information into language models through acoustic landmarks or projected speech representations Zhang et al. (2024); Li et al. (2025a), perform multimodal instruction tuning for depression detection Li et al. (2026a), or use multimodal behavioral cues for depression scoring and rationale generation Zhang et al. (2026). However, these approaches typically rely on naturally corresponding behavioral observations from the same participant or interview. Our setting instead considers physiological and behavioral data collected from independent clinical cohorts and investigates how the two sources of information can contribute to shared intermediate computation without EEG–behavior sample-level correspondence.

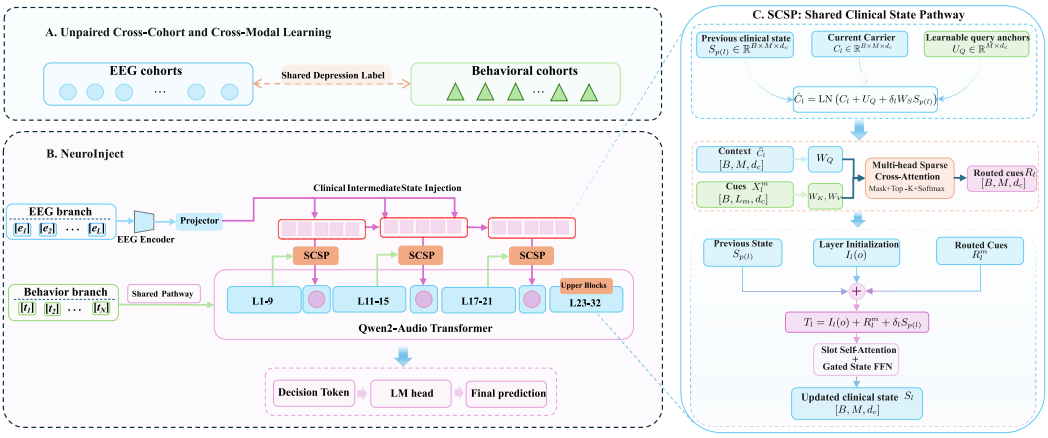


Figure 2: Overall architecture of NeuroInject.

3 METHOD

3.1 PROBLEM FORMULATION

We study an unpaired physiological–behavioral joint learning setting in which EEG and behavioral data are collected from independent clinical cohorts. The EEG and behavioral domains are denoted as

$$\mathcal{D}_E = \{(X_i^E, y_i^E, o_i^E, d_i^E)\}_{i=1}^{N_E}, \qquad \mathcal{D}_B = \{(X_j^B, y_j^B, o_j^B, d_j^B)\}_{j=1}^{N_B}, \qquad \Pi_{E\leftrightarrow B} = \varnothing, \quad (1)$$

where $y \in \{0, 1\}$ denotes the binary depression-related prediction target, o specifies the label ontology defined by clinical diagnosis, questionnaire-based screening, or risk stratification, and d indicates the dataset source. The two modalities are collected from independent clinical cohorts and therefore exhibit no subject-level, temporal, or instance-level correspondence. Consequently, conventional patient-level EEG–behavior multimodal samples cannot be constructed. Under this setting, we do not impose explicit sample-level cross-modal alignment. Instead, EEG and behavioral data are treated as distinct task domains that share a common clinical prediction objective while differing in observation form and supervision protocol. The overall architecture is shown in Fig. 2, where the EEG and behavioral pathways are processed independently and interact through the Shared Clinical State Pathway (SCSP) at multiple intermediate depths before entering a unified decision space.

3.2 HETEROGENEOUS MODALITY REPRESENTATION AND OVERALL FRAMEWORK

EEG and behavioral data exhibit substantially different representational structures and levels of abstraction when entering the joint model. For EEG, each subject is represented by a set of fixed-duration windows. A dataset-specific frozen encoder $f_{E,d}$ encodes each window, followed by a trainable projector $P_{E,d}$ that maps the resulting representation into a common clinical space:

$$Z_i^E = \left[P_{E,d}\big(f_{E,d}(x_{i,1}^E)\big), \ldots, P_{E,d}\big(f_{E,d}(x_{i,L_i^E}^E)\big)\right]. \quad (2)$$

For a given subject, Z_i^E remains fixed throughout a forward pass and serves as an external neural representation that can be repeatedly accessed at different depths of the backbone.

For the behavioral pathway, we retain the native audio–text processing pipeline of the multimodal large language model. Response audio is encoded by the frozen acoustic frontend and multimodal projector, while participant transcripts, interview context, and task instructions are introduced through the textual pathway of the language model. Let A_l denote the residual representation after self-attention at layer l. The modality representations available for intermediate retrieval are then

written uniformly as

$$X_l^m = \begin{cases} P_{E,l}^{\text{in}}(Z_i^E), & m = E, \\ P_{B,l}^{\text{in}}(A_l), & m = B. \end{cases} \tag{3}$$

On the behavioral side, only positions corresponding to the participant's response audio and transcript are exposed to the subsequent information retrieval module. Interviewer questions remain part of the native contextual computation of the backbone but are not directly treated as patient evidence.

The two modalities retain their native information organization without requiring matched sequence lengths or positional correspondence. NeuroInject introduces learnable clinical carriers at selected intermediate layers to retrieve and write back modality-specific evidence before shared high-level adaptation.

3.3 Intermediate Clinical Information Injection

The core of NeuroInject is the Shared Clinical State Pathway (SCSP), which maintains a progressively evolving clinical state across multiple backbone depths. At each selected intermediate layer, SCSP provides a shared interaction pathway for EEG and behavioral representations, enabling cross-modal information exchange through a common clinical-state space without requiring direct sample- or token-level alignment. This state enables the model to retrieve task-relevant information from heterogeneous modalities and inject the resulting clinical evidence back into the frozen language-model backbone through the clinical carriers.

For a label ontology o, the initial clinical carriers are defined as

$$C_0(o) = \mathrm{LN}(U_C + e_o), \tag{4}$$

where U_C denotes the shared carrier basis and e_o is an ontology-conditioned embedding that describes the current label definition without encoding the ground-truth class. The clinical carriers are not classification tokens; instead, they serve as persistent intermediate read–write locations within the residual stream of the language model.

At an injection layer l, the carrier representations are first gathered from the current residual state A_l and projected into the clinical state space:

$$C_l = P_{C,l}\big(\mathrm{Gather}(A_l, I_C)\big). \tag{5}$$

The retrieval query is then jointly conditioned on the current carrier context, a shared learnable query basis, and the clinical state accumulated at the preceding injection layer. Keys and values are constructed from the currently available modality-specific evidence:

$$Q_l = W_Q\,\mathrm{LN}\left(C_l + U_Q + \delta_l W_S S_{p(l)}\right), \qquad K_l^m = W_K\,\mathrm{LN}(X_l^m), \qquad V_l^m = W_V\,\mathrm{LN}(X_l^m), \tag{6}$$

where δ_l indicates whether a previous injection state is available. Modality-specific ingress projections remain separate, while the retrieval parameters are shared across modalities.

Given a fixed set of queries, cross-attention aggregates information over the variable-length modality sequence:

$$\alpha_l^m = \mathrm{softmax}\left(\mathcal{T}\left[\frac{Q_l (K_l^m)^\top}{\sqrt{d}} + \Gamma(M_l^m)\right]\right), \qquad R_l^m = W_O\left(\alpha_l^m V_l^m\right), \tag{7}$$

where M_l^m masks invalid positions and $\mathcal{T}(\cdot)$ denotes an optional sparse evidence-routing operator. Because attention weights are normalized along the input sequence dimension, each query produces a fixed-dimensional evidence representation regardless of the number of EEG windows or behavioral tokens. Consequently, the number of resulting clinical slots is determined by the number of queries rather than the input sequence length. Importantly, this mechanism does not establish one-to-one correspondence between EEG windows and behavioral tokens; instead, it extracts clinically relevant information into a common slot-based structure for each modality under the current context and accumulated clinical state.

The newly retrieved evidence R_l^m is then combined with the clinical state accumulated at the preceding injection layer:

$$T_l = I_l(o) + R_l^m + \delta_l S_{p(l)}, \qquad S_l = \mathrm{Update}_l(T_l), \tag{8}$$

where $\mathrm{Update}_l(\cdot)$ consists of slot-wise self-attention followed by a gated feed-forward network. This formulation establishes continuous propagation of clinical state across successive injection depths.

Notably, the same historical state $S_{p(l)}$ plays distinct roles in Eqs. equation 6 and equation 8. In query construction, it conditions the current retrieval process and thereby influences which evidence should be attended to at the present layer. In state updating, it acts as a residual memory that preserves previously accumulated clinical information and integrates it with newly retrieved evidence. The historical state therefore contributes to both evidence selection and state preservation, rather than being redundantly added to the same representation.

The updated clinical state is mapped back into the hidden space of the language model through a low-rank bridge shared across modalities, while a modality-specific gate controls the strength of the write-back:

$$\Delta C_l^m = \gamma_l^m B_l(S_l), \qquad \widehat{A}_l[:, I_C, :] = A_l[:, I_C, :] + \Delta C_l^m. \tag{9}$$

The write-back is restricted to the clinical carrier positions and does not directly modify the original audio, participant transcript, interview context, prompt, or decision tokens. The updated carriers re-enter the native Transformer computation, allowing information acquired at shallower injection layers to influence evolving representations and subsequent retrieval.

The cross-layer clinical state exists only within the forward pass of an individual sample and is reinitialized for every new sample. No numerical state is propagated between EEG and behavioral batches. The connection between the two modalities is instead established through the joint learning of shared retrieval, state-update, and write-back parameters.

3.4 Unified Decision and Cross-Cohort Optimization

After intermediate clinical injection, a shared parameter-efficient adaptation module further adjusts high-level representations from both modalities. For a frozen linear transformation W, the corresponding low-rank update can be written as

$$Wx + sBAx, \tag{10}$$

where A and B denote trainable low-rank matrices and s is a scaling coefficient. This stage jointly adapts high-level computations after clinical injection without introducing additional sample-level multimodal fusion.

In addition to intermediate-layer write-back, Direct State Fusion uses the clinical state accumulated at the final injection layer as a residual correction to the common decision representation. Both pathways are then classified using the corresponding binary-verbalizer rows of the frozen language-model output head:

$$h^{\mathrm{dec}} = H_L[I_{\mathrm{dec}}], \qquad z_c = w_{v_c}^\top h^{\mathrm{dec}} + b_{v_c}, \qquad c \in \{0, 1\}. \tag{11}$$

Thus, although the two modalities differ in frontend encoding and evidence organization, they ultimately share the intermediate clinical interface, high-level adaptation parameters, and the same language-model decision space.

Different data sources retain local training mechanisms consistent with their respective supervision protocols. The EEG pathway supports subject-level multi-window modeling while also allowing genuine single-window forward passes for window-level supervision and fair comparison with encoder-based baselines. The behavioral pathway performs window-level aggregation over local response or interview-unit predictions according to the structure of the original dataset. Dataset-specific projectors, calibrators, and aggregation modules are used only to accommodate protocol differences, while the primary classification signal is always derived from the shared decision interface.

The overall training objective is

$$\mathcal{L} = \sum_{d \in \mathcal{D}} \lambda_d \mathcal{L}_d + \lambda_{\mathrm{div}} \mathcal{R}_{\mathrm{div}} + \lambda_{\mathrm{use}} \mathcal{R}_{\mathrm{use}}, \tag{12}$$

Table 1: Depression detection performance on MODMA and EDRA. Best results are shown in bold and second-best results are underlined.

Method	MODMA				EDRA			
	Acc.	BAcc.	Macro-F1	AUROC	Acc.	BAcc.	Macro-F1	AUROC
EEGNet	57.39	55.52	54.41	<u>64.84</u>	<u>85.42</u>	<u>85.23</u>	<u>85.23</u>	**94.83**
BIOT	53.41	52.29	52.01	52.36	60.76	59.84	59.87	61.12
LaBraM-base	63.92	61.72	60.40	64.32	59.38	58.28	58.26	65.67
EEGPT	55.40	55.57	55.35	57.53	65.62	66.02	65.57	70.89
CBraMod	<u>67.90</u>	<u>66.46</u>	**66.28**	64.37	63.54	63.28	63.22	69.09
REVE	54.83	53.70	53.43	60.75	51.39	49.84	49.48	47.84
NeuroInject	**71.02**	**68.13**	<u>66.11</u>	**77.32**	**86.81**	**85.63**	**86.25**	<u>80.39</u>

Table 2: Window-level depression detection results on CMDC and DAIC-WOZ. All metrics are reported in percentage points.

Method	CMDC				DAIC-WOZ			
	Acc.	BAcc	Macro-F1	AUROC	Acc.	BAcc	Macro-F1	AUROC
Ef-Bi-LSTM	87.74	84.44	85.04	94.43	60.81	55.59	55.04	57.33
MulT	88.88	85.59	86.37	94.52	60.35	56.00	55.19	57.46
MISA	89.29	84.87	86.51	93.58	59.23	54.79	53.82	55.98
MMIM	<u>90.12</u>	**86.30**	<u>87.67</u>	<u>95.88</u>	<u>62.23</u>	55.46	55.17	57.06
M3L	85.63	81.83	82.43	92.58	61.19	56.38	<u>55.70</u>	57.92
DepressInstruct	83.83	80.62	80.64	89.32	61.93	56.28	54.98	<u>60.18</u>
Qwen2-Audio	83.87	78.38	79.63	88.14	60.53	<u>56.49</u>	54.72	**60.30**
NeuroInject	**91.19**	<u>86.25</u>	**88.66**	**96.19**	**65.78**	**56.90**	**57.06**	59.84

where $\mathcal{L}_d$ is the cross-entropy classification loss computed from the final predictions of the shared LM head for dataset d. $\mathcal{R}_{\text{div}}$ discourages redundant evidence across clinical slots, while $\mathcal{R}_{\text{use}}$ balances their relative magnitudes. Dataset-specific supervision units and loss coefficients are detailed in the appendix.

Different tasks perform independent forward and backward passes. Gradient-conflict coordination is applied only to parameters that are genuinely shared across tasks, including the clinical carriers, evidence router, state updater, shared write-back bridge, and high-level adaptation module. Modality- or dataset-specific parameters are updated only by their corresponding tasks.

4 Experiments

4.1 Experimental Settings

Datasets and protocols. We evaluate NeuroInject on MODMA (Cai et al., 2022), EDRA (Yang et al., 2023b), CMDC (Zou et al., 2023), and DAIC-WOZ (Gratch et al., 2014), using subject-disjoint splits throughout. Primary results are reported at the local window or segment level. Detailed cohort statistics, preprocessing, and split definitions are provided in Appendix A.

Implementation Details. EEG windows are normalized on a per-channel basis and encoded by dataset-specific frozen EEG encoders with the same architecture. The behavioral branch preserves the native audio–text pathway of Qwen2-Audio (Chu et al., 2024). NeuroInject is built on Qwen2-Audio-7B-Instruct, with the backbone, audio frontend, and LM head kept frozen. EEG representations are introduced through an intermediate clinical interface with multiple 256-dimensional clinical-state slots operating at three Transformer depths, where sparse information retrieval, cross-layer state updating, and gated residual write-back are performed. Shared LoRA adapters (Hu et al., 2022) are further applied to the upper Transformer layers. EEG and behavioral tasks are optimized with an alternating cross-cohort training strategy using independent forward passes, while

jointly updating the shared intermediate interface and upper-layer adaptation parameters. PCGrad (Yu et al., 2020) is applied to shared parameters to coordinate gradients across tasks. We use AdamW (Loshchilov & Hutter, 2019) with bf16 precision, with learning rates of 1×10^{-5} for shared modules and 1×10^{-4} for modality-specific parameters.

Baselines and Metrics. For EEG experiments, we compare NeuroInject with EEGNet (Lawhern et al., 2018), BIOT (Yang et al., 2023a), LaBraM-base (Jiang et al., 2024), EEGPT (Wang et al., 2024), CBraMod (Wang et al., 2025a), and REVE (El Ouahidi et al., 2025) under the same subject-disjoint protocol. For behavioral experiments, we compare against Ef-Bi-LSTM (Zou et al., 2023), MulT (Tsai et al., 2019), MISA (Hazarika et al., 2020), MMIM (Han et al., 2021), M3L (You et al., 2025), Qwen2-Audio + LoRA (Chu et al., 2024; Hu et al., 2022), and DepressInstruct (Li et al., 2026a). Conventional behavioral baselines use frozen HuBERT representations for audio (Hsu et al., 2021) and RoBERTa-family representations for text (Liu et al., 2019). All primary experiments are evaluated at the local window or segment level under subject-disjoint splits. We report Accuracy, balanced accuracy, Macro-F1, and AUROC, and select models exclusively on the validation set.

4.2 Results on EEG Depression Recognition

Table 1 reports the results on MODMA and EDRA. NeuroInject demonstrates strong performance across the two EEG datasets despite their differences in channel configuration and cohort characteristics. On MODMA, NeuroInject achieves the best Accuracy, balanced accuracy, and AUROC, reaching 71.02, 68.13, and 77.32, respectively. Compared with the strongest competing results, these correspond to improvements of 3.12, 1.67, and 12.95 percentage points. Its Macro-F1 reaches 66.11, only 0.17 points below CBraMod, indicating that the improvement in overall classification performance is not obtained at the expense of class-balanced prediction.

The advantage is also observed on EDRA. NeuroInject achieves the highest Accuracy, balanced accuracy, and Macro-F1 of 86.81, 85.63, and 86.25, improving upon the corresponding strongest baselines by 1.39, 0.40, and 1.02 points, respectively. In contrast, EEGNet obtains a substantially higher AUROC of 94.83 compared with 80.39 for NeuroInject. This discrepancy suggests that the advantage of NeuroInject on EDRA is more pronounced in threshold-dependent classification metrics than in global ranking performance. Nevertheless, its consistent gains in Accuracy, balanced accuracy, and Macro-F1 across both EEG datasets indicate that the intermediate clinical interface can effectively incorporate encoded EEG representations into the language-model computation pathway.

Notably, the improvements are obtained without directly converting EEG signals into language-like input tokens. Instead, EEG representations remain external to the original token sequence and interact with the backbone only through the intermediate interface. The results therefore support the feasibility of introducing abstract neural representations at intermediate computational depths, rather than relying exclusively on input-level tokenization or a standalone EEG classifier.

4.3 Results on Behavioral Depression Recognition

Table 2 summarizes the window-level results on CMDC and DAIC-WOZ. On CMDC, NeuroInject achieves the best Accuracy, Macro-F1, and AUROC, with scores of 91.19, 88.66, and 96.19, respectively. These results improve over the strongest baselines by 1.07, 0.99, and 0.31 percentage points. Its balanced accuracy reaches 86.25, which is nearly identical to the best result of 86.30 obtained by MMIM. Compared with the native Qwen2-Audio baseline, NeuroInject improves Accuracy and Macro-F1 by 7.32 and 9.03 points, respectively, suggesting that intermediate clinical-state interaction provides substantial task adaptation beyond directly fine-tuning the upper layers of the multimodal backbone.

The improvement is more evident on DAIC-WOZ in terms of classification performance. NeuroInject obtains the best Accuracy, balanced accuracy, and Macro-F1, reaching 65.78, 56.90, and 57.06, respectively. Compared with the strongest competing values for these metrics, the gains are 3.55, 0.41, and 1.36 points. NeuroInject also substantially improves over the native Qwen2-Audio baseline in Accuracy and Macro-F1 by 5.25 and 2.34 points. Its AUROC of 59.84 is lower than Qwen2-Audio (60.30) and DepressInstruct (60.18), indicating that the gains again primarily appear in classification performance rather than overall ranking quality.

Table 3: Ablation study on MODMA and EDRA.

Configuration	MODMA				EDRA			
	Acc.	BAcc	Macro-F1	AUROC	Acc.	BAcc	Macro-F1	AUROC
EEG encoder	65.91	62.92	59.93	71.67	80.56	78.83	79.33	78.32
w/o Direct State Fusion	70.45	67.66	65.79	77.10	86.46	85.31	85.91	79.84
Full model	**71.02**	**68.13**	**66.11**	**77.32**	**86.81**	**85.63**	**86.25**	**80.39**

Overall, NeuroInject maintains competitive performance across CMDC and DAIC-WOZ despite substantial differences in interview setting and language. Together with the EEG results, these findings indicate that the same intermediate interaction framework can accommodate both externally encoded physiological representations and audio–text representations processed along the native multimodal pathway.

4.4 Ablation Study

Table 3 evaluates the contributions of intermediate language-model interaction and Direct State Fusion on MODMA and EDRA. Compared with the standalone EEG encoder, removing Direct State Fusion while retaining intermediate interaction already improves Accuracy, balanced accuracy, Macro-F1, and AUROC by 4.54, 4.74, 5.86, and 5.43 points on MODMA, and by 5.90, 6.48, 6.58, and 1.52 points on EDRA, respectively. This demonstrates that integrating encoded EEG representations into the intermediate clinical-state computation provides the primary performance gain.

Direct State Fusion further aggregates the clinical state from the final SCSP injection layer, projects it into the language-model hidden space, and applies a gated residual correction to the decision representation before prediction by the shared LM head. It further improves the four metrics by 0.57, 0.47, 0.32, and 0.22 points on MODMA, and by 0.35, 0.32, 0.34, and 0.55 points on EDRA. These consistent gains indicate that the accumulated clinical state provides complementary information for refining the final prediction.

5 Conclusion

We presented NeuroInject, a framework for unpaired physiological–behavioral depression recognition across independently collected clinical cohorts. Rather than constructing artificial EEG–behavior pairs or forcing heterogeneous modalities into a common input-level representation, NeuroInject preserves their native processing pathways and enables cross-cohort interaction through a shared intermediate clinical interface. A progressively updated clinical state retrieves task-relevant information from modality-specific representations and writes it back into the frozen language-model backbone, while shared upper-layer adaptation and a unified decision space support joint optimization across modalities.

Experiments on MODMA, EDRA, CMDC, and DAIC-WOZ show that NeuroInject achieves competitive performance across both EEG and audio–text depression recognition settings. The ablation results further indicate that introducing EEG representations into the multimodal backbone already provides gains over the standalone EEG encoder, while the progressive clinical-state pathway provides additional improvements. These results suggest that intermediate interaction provides a practical alternative to sample-level multimodal fusion when physiological and behavioral observations are collected from independent cohorts, and motivates further study of shared computation across heterogeneous clinical modalities.

AI use statement

In this work, we used generative AI tools to assist with language polishing and parts of the experimental implementation. We did not use generative AI tools for any other tasks requiring disclosure, and no other required-disclosure uses are applicable to this work. We reviewed all AI-assisted work. In particular, AI-assisted text was carefully reviewed and revised by the authors, and AI-assisted code was manually inspected, verified, and tested for correctness. We take responsibility for the

final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Hanshu Cai, Zhenqin Yuan, Yiwen Gao, Shuting Sun, Na Li, Fuze Tian, Han Xiao, Jianxiu Li, Zhengwu Yang, Xiaowei Li, Qinglin Zhao, Zhenyu Liu, Zhijun Yao, Minqiang Yang, Hong Peng, Jing Zhu, Xiaowei Zhang, Guoping Gao, Fang Zheng, Rui Li, Zhihua Guo, Rong Ma, Jing Yang, Lan Zhang, Xiping Hu, Yumin Li, and Bin Hu. A multi-modal open dataset for mental-disorder analysis. *Scientific Data*, 9:178, 2022. doi: 10.1038/s41597-022-01211-x.

Zhisheng Chen, Yingwei Zhang, Qizhen Lan, Tianyu Liu, Huacan Wang, Yi Ding, Ziyu Jia, Ronghao Chen, Kun Wang, and Xinliang Zhou. Uni-NTFM: A unified foundation model for EEG signal representation learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Ziqiang Chen, Dandan Wang, Liangliang Lou, Shiqing Zhang, Xiaoming Zhao, Shuqiang Jiang, Jun Yu, and Jun Xiao. Text-guided multimodal depression detection via cross-modal feature reconstruction and decomposition. *Information Fusion*, 117:102861, 2025. doi: 10.1016/j.inffus.2024.102861.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-Audio Technical Report. *arXiv preprint arXiv:2407.10759*, 2024.

Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. LUNA: Efficient and topology-agnostic foundation model for EEG signal analysis. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2377.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG – adapting to any setup with large-scale pretraining on 25,000 subjects. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0760.

Tian Fang, Gaël Guibon, and Davide Buscaldi. Some emotions run deeper: Layer-wise probing and causal intervention in large language models. *arXiv preprint arXiv:2609.01279*, 2026.

Pinyuan Feng, Hossein Adeli, Wenxuan Guo, Fan Cheng, Ethan Hwang, and Nikolaus Kriegeskorte. Towards interpretable visual decoding with attention to brain representations. In *The Fourteenth International Conference on Learning Representations*, 2026.

Changzeng Fu, Shiwen Zhao, Yunze Zhang, Zhongquan Jian, Shiqi Zhao, and Chaoran Liu. Personality-guided public-private domain disentangled hypergraph-former network for multimodal depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1801–1809, 2026. doi: 10.1609/aaai.v40i3.37159.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pp. 3123–3128. European Language Resources Association (ELRA), 2014. URL https://aclanthology.org/L14-1421/.

Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9180–9192. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.723.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1122–1131, 2020. doi: 10.1145/3394171.3413678.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QzTpTRVtrP.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. NeuroLM: A universal multitask foundation model for bridging the gap between language and EEG signals. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Io9yFt7XH7.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. doi: 10.1088/1741-2552/aace8c.

Dongdong Li, Li Ding, Zhe Wang, and Ke Zhao. Depressinstruct: Instruction tuning of large speech-language models for depression detection. *Information Fusion*, 130:104077, 2026a. doi: 10.1016/j.inffus.2025.104077.

Jiahe Li, Junru Chen, Fanqi Shen, Jialan Yang, Jada Li, Zhizhang Yuan, Baowen Cheng, Meng Li, and Yang Yang. Assembling the mind's mosaic: Towards EEG semantic intent decoding. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Yupei Li, Shuaijie Shao, Manuel Milling, and Björn W. Schuller. Large language models for depression recognition in spoken language integrating psychological knowledge. *Frontiers in Computer Science*, 7:1629725, 2025a. doi: 10.3389/fcomp.2025.1629725.

Yuxin Li, Sinchana Kumbale, Yanru Chen, Tanmay Surana, Eng Siong Chng, and Cuntai Guan. Automated depression detection from text and audio: A systematic review. *IEEE Journal of Biomedical and Health Informatics*, 29(10):7498–7513, 2025b. doi: 10.1109/JBHI.2025.3570900.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *The Seventh International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Razvan Marinescu, Victoria-Elisabeth Gruber, and Diego Fajardo Vargas. Medical interpretability and knowledge maps of large language models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Sabine Muzellec and Kohitij Kar. Reverse predictivity for bidirectional comparison of neural networks and biological brains. *Nature Machine Intelligence*, 8:474–488, 2026. doi: 10.1038/s42256-026-01204-0.

Yuchen Pan, Junjun Jiang, Kui Jiang, and Xianming Liu. Disentangled-multimodal privileged knowledge distillation for depression recognition with incomplete multimodal data. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5712–5721, 2024. doi: 10.1145/3664647.3681227.

Jinghui Qin, Changsong Liu, Tianchi Tang, Dahuang Liu, Minghao Wang, Qianying Huang, and Rumin Zhang. Mental-perceiver: Audio-textual multi-modal learning for estimating mental disorders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25029–25037, 2025. doi: 10.1609/aaai.v39i23.34687.

Hoorain Rehman, Danish M. Khan, Hafsa Amanullah, Laiba Kamran, Owais Ur Rehman, Sana Taj Siddiqui, and Komal Masroor. Advances in eeg-based detection of major depressive disorder using shallow and deep learning techniques: A systematic review. *Computers in Biology and Medicine*, 192:110154, 2025. doi: 10.1016/j.compbiomed.2025.110154.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1656.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=NPNUHgHF2w.

Luyao Wang, Chenhan Wang, Chenyang Li, Toshiya Murai, Yicai Bai, Ziyan Song, Shuoyan Zhang, Qi Zhang, Yu Huang, Xiaoying Bi, and Jiehui Jiang. Ai-assisted multi-modal information for the screening of depression: A systematic review and meta-analysis. *npj Digital Medicine*, 8:523, 2025b. doi: 10.1038/s41746-025-01933-3.

Mingqing Xiao, Kai Du, and Zhouchen Lin. Beyond representational alignment with brain-guided language models for robust reasoning. *Nature Machine Intelligence*, 8:1275–1289, 2026. doi: 10.1038/s42256-026-01278-w.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78240–78260, 2023a. doi: 10.52202/075280-3420.

Lijun Yang, Xiaoge Wei, Fengrui Liu, Xiangru Zhu, and Feng Zhou. Automatic feature learning model combining functional connectivity network and graph regularization for depression detection. *Biomedical Signal Processing and Control*, 82:104520, 2023b. doi: 10.1016/j.bspc.2022.104520.

Ke Yi, Yansen Wang, Kan Ren, and Dongsheng Li. Learning topology-agnostic EEG representations with geometry-aware modeling. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Jiajun You, Shuai Wang, Xun Gong, and Xiang Wan. M3L: A multi-modal and multi-lingual depression detection framework. In *Proceedings of Interspeech 2025*, pp. 5268–5272, 2025. doi: 10.21437/Interspeech.2025-329.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.nips.cc/paper/2020/hash/3fe78a8acf5fda99de95303940a2420c-Abstract.html.

Yanke Yu, Jin Li, Ying Sun, Ping Li, Zhefeng Wang, and Yi Zheng. Discovering decoupled functional modules in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34503–34511, 2026. doi: 10.1609/aaai.v40i41.40749.

Qingzhu Zhang, Jiani Zhong, Zongsheng Li, Xinke Shen, and Quanying Liu. Multi-dataset joint pre-training of emotional EEG enables generalizable affective computing. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5515.

Wei Zhang, Juan Chen, En Zhu, Wenhong Cheng, Yunpeng Li, Yuhan Li, and Yanbo J. Wang. MLlm-DR: Towards explainable depression recognition with multimodal large language models. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 22(4):111:1–111:23, 2026. doi: 10.1145/3796722.

Xiangyu Zhang, Hexin Liu, Kaishuai Xu, Qiquan Zhang, Daijiao Liu, Beena Ahmed, and Julien Epps. When LLMs meets acoustic landmarks: An efficient approach to integrate speech into large language models for depression detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 146–158. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.8.

Yuchen Zhou, Jiamin Wu, Zichen Ren, Zhouheng Yao, Weiheng Lu, Kunyu Peng, Qihao Zheng, Chunfeng Song, Wanli Ouyang, and Chao Gou. CSBrain: A cross-scale spatiotemporal brain foundation model for EEG decoding. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Bochao Zou, Jiali Han, Yingxue Wang, Rui Liu, Shenghui Zhao, Lei Feng, Xiangwen Lyu, and Huimin Ma. Semi-structural interview-based chinese multimodal depression corpus towards automatic preliminary screening of depressive disorders. *IEEE Transactions on Affective Computing*, 14(4):2823–2838, 2023. doi: 10.1109/TAFFC.2022.3181210.

APPENDIX

A DATASETS AND EVALUATION PROTOCOLS

Our experiments include two EEG cohorts, MODMA and EDRA, and two behavioral cohorts, CMDC and DAIC-WOZ. EEG and behavioral observations are collected from different participants. We do not construct physiological–behavioral sample pairs, nor do we share participant identity across modality-specific forward passes. All numerical labels follow the convention $y \in \{0, 1\}$, but their clinical meanings differ: MODMA and CMDC use clinical-diagnosis labels, EDRA uses depression-risk groups, and DAIC-WOZ uses the dataset-provided PHQ-8 screening labels. A unified binary encoding does not imply that these label definitions are interchangeable.

MODMA. The locally retained resting-state EEG cohort contains 53 participants, including 29 healthy controls and 24 participants in the depression group. The archived independent encoder uses 19 mapped channels in the following order: Fp1, F3, C3, P3, O1, F7, T7, P7, Fz, Fp2, F4, C4, P4, O2, F8, T8, P8, Cz, and Pz. Inputs are five-second windows represented at 256 Hz, corresponding to 1,280 samples per channel.

EDRA. The locally retained dataset contains 46 participants, including 20 in the lower-risk group and 26 in the higher-risk group; this reflects the source files actually available locally, whereas the dataset provider describes a complete cohort of 50 participants. After removing the non-scalp `IO` channel, 62 scalp channels are retained according to the checkpoint-recorded channel order and converted to the same five-second, 1,280-sample encoder input format. This task evaluates discrimination between risk groups, and a positive prediction should not be interpreted as a clinical MDD diagnosis.

CMDC. We use audio and transcript data from the Chinese interview cohort, with 78 participants retained in the current split. Local samples are segmented within each individual answer using ten-second chunks. When the remaining portion at the end of an answer is shorter than a full chunk, the final chunk is shifted backward to end at the answer boundary and may therefore overlap with the preceding chunk. Answers between two and ten seconds are retained at their original duration, whereas answers shorter than two seconds are excluded. The interview question is provided as context, and the answer transcript is retained together with its corresponding audio. Visual information is not used. The current test set contains all 18 assigned participants, including HC51; results based on the older 17-participant cache correspond to a different evaluation population.

Table 4: Retained subject splits. Counts in parentheses indicate the numbers of class-0/class-1 participants. Dataset-specific label definitions are preserved.

Dataset	Train	Validation	Test
MODMA	31 (17/14)	11 (6/5)	11 (6/5)
EDRA	28 (12/16)	9 (4/5)	9 (4/5)
CMDC	42 (28/14)	18 (12/6)	18 (12/6)
DAIC-WOZ	107 (77/30)	34 (22/12)	47 (33/14)

DAIC-WOZ. We use participant speech and interview transcripts under the official train, development, and test splits. The local development set retains 34 participants because the required transcript for participant 302 is unavailable, resulting in 107/34/47 participants in the train/development/test sets, respectively. Ten-second windows are constructed along the interview timeline, using a five-second hop during training and a ten-second hop during evaluation. Interviewer speech is masked, while conversational timing, pauses, and response delays are preserved. The current manifest requires at least 2.5 seconds of participant speech within each window. Audio from both behavioral datasets is represented as 16 kHz waveforms.

A.1 Subject Separation and Local Samples

Subjects are split before local samples are constructed. All windows and responses from the same participant inherit the same data split. The stored subject lists are used to audit missing, extra, and overlapping participants.

A.2 Metrics and Model Selection

We report Accuracy, balanced accuracy (BAcc), Macro-F1, and AUROC. Taking class 1 as the positive class,

Macro-F1 is the average of the F1 scores of the two classes, and AUROC is computed using the continuous positive-class scores.

B Independent EEG Encoder

B.1 Architecture

MODMA and EDRA use independently trained instances of the same architecture, with no weight sharing between the two datasets. For an input $x \in \mathbb{R}^{C\times 1280}$, each channel is first centered within the window and then normalized by its own root-mean-square magnitude:

$$\widetilde{x}_{c,t} = \mathrm{clip}\left(\frac{x_{c,t} - \overline{x}_c}{\max\left\{\sqrt{1280^{-1}\sum_u (x_{c,u} - \overline{x}_c)^2}, 10^{-4}\right\}}, -12, 12\right). \tag{13}$$

Three 64-dimensional representations are then constructed for each channel.

Temporal morphology branch. All channels share the same convolutional architecture: a convolution with kernel size 33 and stride 4, followed by a depthwise convolution with kernel size 17 and stride 2, and a pointwise projection. Three temporal residual blocks use kernel sizes 9, 17, and 9, respectively. Group normalization, GELU, and dropout are applied throughout the network. The temporal mean and standard deviation are concatenated and projected to a 64-dimensional channel representation.

Spectral statistics branch. A real-valued FFT is used to estimate power in five frequency bands: $[1, 4)$, $[4, 8)$, $[8, 13)$, $[13, 30)$, and $[30, 45)$ Hz. For each band, log power and relative power are computed. Together with relative-power entropy and the log alpha/theta power ratio, these form 12

Table 5: EEG baseline implementations and their roles in the local comparison.

Method	Local implementation and comparison role
EEGNet	Temporal convolutions, spatial depthwise convolutions, and separable convolutions adapted to the local EEG window format.
Unified encoder	The dataset-specific supervised encoder described above, evaluated using its own frozen margin classification head.
BIOT	A frozen biosignal Transformer using the released 18-channel embedding vocabulary. Available electrodes are selected by name while preserving the corresponding pretrained channel indices.
LaBraM-base	A frozen EEG representation model using the local channel and window adapter, with the downstream probe selected on the validation set.
EEGPT	A frozen converted checkpoint with its learned channel IDs restored. Inputs follow the standard 62-position ordering, with unavailable electrodes zero-filled.
CBraMod	A frozen spatial-temporal token representation model whose adapter aggregates encoder outputs across the channel and temporal dimensions.
REVE-base	A frozen EEG representation model using electrode coordinates from the position table provided with the model.

features, which are then projected to 64 dimensions. These features provide numerical statistics of the normalized window.

Channel connectivity branch. The waveform is average-pooled with kernel size and stride both equal to 8, followed by centering and RMS normalization before channel-wise correlations are computed. Six statistics are constructed for each channel: mean correlation, root-mean-square correlation, mean positive correlation, mean negative-correlation magnitude, and the mean correlations with spatially neighboring and distant electrodes.

Spatial fusion and pooling. A learned softmax gate combines the three channel representations, after which a two-layer projection of the three-dimensional electrode coordinates is added. Two pre-normalized Transformer encoder layers model cross-channel relationships, with hidden width 64, four attention heads, and an FFN width of 192. Attentive channel pooling, the channel-wise mean, and the channel-wise standard deviation are concatenated to obtain the final encoder representation. A scalar margin classification head outputs m^{enc}, which is represented by two logits,

$$\left[-\frac{m^{\mathrm{enc}}}{2}, \frac{m^{\mathrm{enc}}}{2}\right].$$

C EEG Baseline Implementations

The EEG comparison includes a supervised compact convolutional baseline, the independent encoder described in Section B, and frozen pretrained encoders paired with validation-selected downstream probes. All methods follow the designated subject splits, but window IDs and preprocessing pipelines should be verified separately across different cache versions. Identical subject lists do not imply identical retained windows, channel coverage, or backbone input representations.

D Behavioral Baseline Implementations

D.1 Common Audio and Text Features

Ef-Bi-LSTM, MulT, MISA, and MMIM use a shared frozen feature cache. Audio is encoded using HuBERT-large-ll60k, while Chinese transcripts are encoded with Chinese RoBERTa-wwm-ext and English transcripts with RoBERTa-base. Audio frames and text tokens are mask-averaged within each local sample. The text representation includes the provided interviewer context and the participant transcript. Feature-wise means and standard deviations are estimated using training subjects only.

In local-window experiments, each forward pass classifies one cached audio–text feature pair. Consequently, the fusion networks receive a sequence of length one at this evaluation granularity, whereas subject-level experiments operate on multiple local feature vectors. Because the original waveform frames and text tokens have already been pooled, neither setting fully preserves fine-grained cross-attention between the original audio and textual sequences. We therefore describe these models as local reimplementations aligned to the common input protocol rather than unchanged official implementations.

D.2 Individual Baselines

Ef-Bi-LSTM. Audio and text features are concatenated and passed through a bidirectional recurrent encoder followed by attention pooling, providing a compact early-fusion baseline under the shared cached-feature interface.

MulT. Projected audio and text streams interact through bidirectional cross-modal attention, followed by stream-specific processing and pooling. In the local-window implementation, the inputs have already been pooled, which limits the sequential context available to the attention operations.

MISA. MISA constructs modality-private and modality-shared representations. In addition to the classification objective, it employs reconstruction, difference, and central moment discrepancy objectives. These auxiliary objectives and the explicit separation of representations distinguish it from simple feature concatenation.

MMIM. The recurrent modality encoders are trained using a combination of classification, variational mutual-information estimation, and contrastive predictive coding objectives. In our aligned implementation, audio and text representations are concatenated and processed by a feed-forward fusion module for binary classification.

M3L. The aligned implementation first trains an audio–text fusion backbone on the CMDC and DAIC-WOZ training subjects. The backbone consists of a projected text branch, a bidirectional recurrent audio branch, modality-channel attention, and representation regularization. For dataset-specific adaptation, the backbone is then frozen while low-rank branches and the classifier are trained. The local implementation uses the common HuBERT/RoBERTa features, which differ from the Whisper/XLM-R feature pair used in the original method.

Qwen2-Audio. This baseline retains native waveform and text processing. The base model, audio encoder, and audio projector are frozen, while rank-8 LoRA modules are applied to the query, key, value, and output projections of the decoder. No SCSP module is constructed, allowing this baseline to control for adaptation of the same backbone without the proposed clinical interface.

DepressInstruct-core. The local implementation combines Qwen2-Audio, upper-layer LoRA, a residual bottleneck following the frozen audio encoder, and a trainable audio projector. Because the available implementation does not include the original SECap generation pipeline, SECap-generated emotion descriptions are omitted.

D.3 Optimization and Reporting

For the four common-feature fusion baselines, the hidden width is 128 and dropout is 0.2. The window-level training configuration uses a batch size of 64, AdamW with a learning rate of 3×10^{-4} and weight decay of 10^{-4}, a maximum of 100 epochs, an early-stopping patience of 15, and cosine learning-rate decay. Window-level loss weights balance both classes and subjects so that longer interviews do not dominate training solely by contributing more windows. The native Qwen baseline uses a LoRA learning rate of 10^{-4}.

Table 6: Clinical interface configuration.

Setting	Value
Backbone	Qwen2-Audio-7B-Instruct
Clinical carriers / state slots	8 / 8
LM hidden width / state width	4,096 / 256
Injection layers	10, 16, 22
Routing attention heads / state FFN width	8 / 512
EEG / behavioral routing top-k	8 / 32
Write-back bridge rank	64
Interface dropout	0.1
Initial state-gate / write-back-gate parameter	0.01 / 0.05
Post-injection LoRA layers	23–32
LoRA rank / alpha / dropout	8 / 16 / 0.05

E Clinical Interface and Prediction Pathways

The Shared Clinical State Pathway (SCSP) consists of dynamic clinical evidence retrieval and a cross-layer clinical state bus. The implementation distinguishes language-model carrier tokens from clinical-state slots: the former reside in the 4,096-dimensional residual space, whereas the latter operate in a 256-dimensional interface space. The carriers provide dedicated locations for intermediate-layer read and write operations and are not the binary label tokens.

EEG and behavioral batches are processed through independent forward passes. The clinical state is reinitialized for each sample and is propagated only across the selected network layers within that sample. Shared parameters are jointly updated by both domains, but the numerical clinical state of an EEG participant is never transferred to a behavioral participant. Participant audio and transcript positions provide behavioral evidence, while interviewer questions remain part of the backbone context.

E.1 Training Objective and Slot Regularization

For dataset d, let $z_{di} \in \mathbb{R}^2$ denote the final binary-classification logits obtained after DSR and the frozen LM head, and let

$$p_{di} = \mathrm{softmax}(z_{di}).$$

The classification loss is defined as

$$\mathcal{L}_d = -\frac{\sum_{i=1}^{n_d} \omega_{di} \log p_{di,y_{di}}}{\sum_{i=1}^{n_d} \omega_{di}}, \tag{14}$$

where n_d denotes the number of valid prediction units and $\omega_{di} > 0$ denotes the class-level weight used under the corresponding configuration, with $\omega_{di} = 1$ in the unweighted case. For window-level supervision, each window is treated directly as a prediction unit and inherits the label of its corresponding subject.

Let R_{lj} denote the j-th retrieved evidence slot at injection layer l, where M is the number of slots and $\mathcal{I}$ denotes the set of injection layers. We define

$$\widehat{R}_{lj} = \frac{R_{lj}}{\max\left(\|R_{lj}\|_2, \epsilon\right)}, \qquad \epsilon = 10^{-6}. \tag{15}$$

The diversity regularization term is

$$\mathcal{R}_{\mathrm{div}} = \frac{1}{|\mathcal{I}|M(M-1)} \sum_{l\in\mathcal{I}} \sum_{j\neq k} \left(\widehat{R}_{lj}^{\top} \widehat{R}_{lk}\right)^2. \tag{16}$$

This term suppresses excessive correlation among retrieved evidence slots. When $M = 1$, we define $\mathcal{R}_{\mathrm{div}} = 0$.

The slot-utilization regularization is defined as

$$u_{lj} = \frac{\|R_{lj}\|_2}{\max\left\{\sum_{k=1}^{M} \|R_{lk}\|_2, \epsilon\right\}}, \tag{17}$$

and

$$\mathcal{R}_{\text{use}} = \frac{1}{|\mathcal{I}|M} \sum_{l \in \mathcal{I}} \sum_{j=1}^{M} \left(u_{lj} - \frac{1}{M} \right)^2 . \tag{18}$$

Both regularization terms are averaged over valid samples. The latter constrains the relative magnitude of the retrieved evidence across slots rather than the frequency with which individual slots are selected by the attention mechanism.

The overall objective is

$$\mathcal{L} = \sum_{d} \lambda_d \mathcal{L}_d + \lambda_{\text{div}} \mathcal{R}_{\text{div}} + \lambda_{\text{use}} \mathcal{R}_{\text{use}}. \tag{19}$$

F Cross-Modal Shared Representation Analysis

To examine class-related correspondence between physiological and behavioral representations, we analyze FFN activations from two separately trained models based on the same pretrained Qwen backbone. The EEG model is trained on MODMA and EDRA, and the behavioral model on CMDC and DAIC-WOZ. Using their validation cohorts, comprising 72 participants in total, we evaluate whether a class direction estimated from one cohort is associated with class ordering in a cohort from the other modality. The native FFN weights in layers L1–L22 are identical and frozen, while layers L23–L32 contain separately trained LoRA parameters.

Representation extraction. We extract the gated FFN activations at each Qwen layer from the final valid input-token position, immediately before the down-projection layer. This position has a corresponding intermediate representation at every layer and is not an already generated class token. Window-level representations are first averaged within each participant, followed by cohort-wise centering and unsupervised scale normalization. To stabilize the scale, the variance of each unit is shrunk by 10% toward the median nonzero variance of the corresponding layer. The resulting representation is denoted by

$$x_i^{m,\ell} \in \mathbb{R}^{11008},$$

where m, ℓ, and i denote the cohort, network layer, and participant, respectively. Class labels retain their dataset-specific definitions. The target cohort is normalized using its own unlabeled statistics; therefore, this analysis is transductive with respect to the current validation cohorts.

Class directions and residual decomposition. For a source cohort s, we define the class direction as the difference between the mean representations of class 1 and class 0:

$$d_s^{\ell} = \frac{1}{n_s^{+}} \sum_{i:y_i^s=1} x_i^{s,\ell} - \frac{1}{n_s^{-}} \sum_{i:y_i^s=0} x_i^{s,\ell}, \tag{20}$$

where $y_i^s \in \{0,1\}$ is the label of source participant i, and n_s^{+} and n_s^{-} denote the numbers of participants in the two classes. This direction describes the average representational change from class 0 to class 1 in the source cohort.

We then perform unlabeled principal component analysis on the source-cohort representations. Let p_s^{ℓ} denote the unit vector of the first principal component, i.e., the direction of maximum source-side representational variance. The class direction is decomposed as

$$\begin{aligned} d_{s,\text{PC1}}^{\ell} &= p_s^{\ell} (p_s^{\ell})^{\top} d_s^{\ell}, \\ d_{s,\text{res}}^{\ell} &= d_s^{\ell} - d_{s,\text{PC1}}^{\ell}. \end{aligned} \tag{21}$$

Here, $d_{s,\text{PC1}}^{\ell}$ is the projection of the class direction onto the first principal component, whereas $d_{s,\text{res}}^{\ell}$ satisfies

$$(p_s^{\ell})^{\top} d_{s,\text{res}}^{\ell} = 0$$

and represents the component orthogonal to PC1. Thus, the residual here refers to the class direction remaining after removing the dominant source-side variance component; it is neither a classification residual nor a residual connection in the network. This post-hoc decomposition is used to examine whether class-ordering associations remain in non-dominant variance components, without assuming that PC1 corresponds to noise or assigning clinical meaning to the residual component.

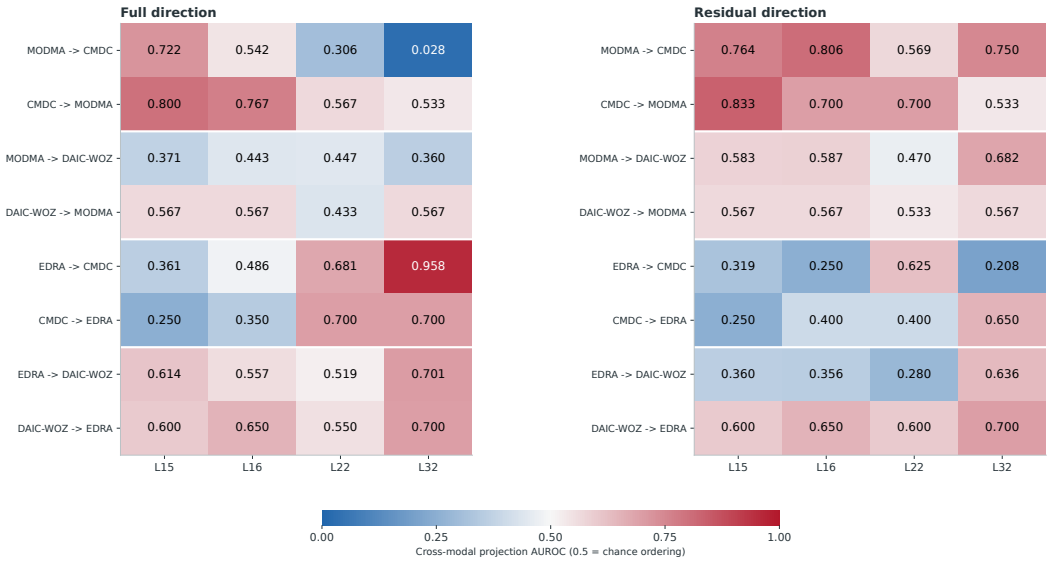


Figure 3: Cross-modal class-direction projection analysis across selected Qwen layers. Projection AUROC is reported for both the full source class direction and the residual direction after removing its PC1 component.

Cross-modal projection and AUROC. For a target cohort t from the other modality, we compute a continuous projection score for each participant using either the full class direction or its residual component:

$$q_{s\to t,i}^{\ell,c} = (x_i^{t,\ell})^\top d_{s,c}^{\ell}, \qquad c \in \{\text{full}, \text{res}\}, \qquad d_{s,\text{full}}^{\ell} = d_s^{\ell}. \tag{22}$$

The target representations are kept unchanged. No additional cross-modal mapping or target-side classifier is fitted, and the direction is not flipped according to target labels.

For a fixed source–target pair, network layer, and direction component, we abbreviate the projection score as q_i. The projection AUROC is defined as

$$\text{AUROC}_{\text{proj}} = \frac{1}{n_t^+ n_t^-} \sum_{i:y_i^t=1} \sum_{j:y_j^t=0} \left[\mathbf{1}(q_i > q_j) + \frac{1}{2}\mathbf{1}(q_i = q_j) \right], \tag{23}$$

where n_t^+ and n_t^- denote the numbers of participants in the two target classes, and $\mathbf{1}(\cdot)$ is the indicator function. This metric measures the proportion of positive–negative participant pairs for which the positive participant receives a higher projection score, with ties counted as one half. It therefore quantifies the ranking association between the continuous projection score and the target labels, rather than model classification accuracy or cosine similarity between activation vectors.

A value of 0.5 corresponds to no ranking advantage, whereas values above or below 0.5 indicate concordant or reversed ordering tendencies, respectively. For example, the 6 positive and 12 negative participants in CMDC form 72 positive–negative pairs. If 58 pairs are concordantly ordered with no ties, the projection AUROC is

$$58/72 = 0.806.$$

These pairs share participants and therefore should not be treated as statistically independent observations.

Results. Figure 3 reports projection AUROC for the full and residual directions at the common layers L15, L16, L22, and L32. At L15, the full directions of MODMA→CMDC and CMDC→MODMA reach 0.722 and 0.800, respectively, indicating bidirectional concordant ordering in the current sample.

At L16, the full-direction AUROC for MODMA→CMDC is 0.542, whereas the residual-direction AUROC increases to 0.806, suggesting that the class-ordering information for this source–target

pair is not carried exclusively by the dominant source-side variance component. At L32, the full-direction AUROC for EDRA→CMDC reaches 0.958, whereas the residual-direction AUROC is 0.208, further showing that cross-modal associations across cohorts and layers may be carried by different representational components. The current validation cohorts exhibit localized cross-modal class-ordering associations that depend on both the source–target cohort pair and the network layer, with their behavior further varying across direction components.