Leveraging Foundational Models and Simple Fusion for Multi-modal Physiological Signal Analysis

Youssef Ghallab[1,2] **Omar Iraqy**[1] **Mohamed Kandil**[1] **Mohamed Ashraf**[1]

Saadeldine Eletter[1,2] **Morougue Ghazal**[1] **Ayman Khalafallah**[1] **Nagwa El-Makky**[1]

[1]Computer and Communication Engineering Department, Alexandria University
[2]Mohamed bin Zayed University of Artificial Intelligence

{es-youssifghalab2025, es-omarmohammed2025, es-mohamed.abdelmoneim2025,
es-mohameda.hamdy2025, es-saadeddin2025, es-morojmahmoud2025, ayman.khalafallah,
nagwamakky}@alexu.edu.eg

{youssef.ghallab, saadeldine.eletter}@mbzuai.ac.ae

Abstract

Physiological signals such as electrocardiograms (ECG) and electroencephalograms (EEG) provide complementary insights into human health and cognition, yet multi-modal integration is challenging due to limited multi-modal labeled data, and modality-specific differences . In this work, we adapt the CBraMod encoder [1] for large-scale self-supervised ECG pretraining, introducing a dual-masking strategy to capture intra- and inter-lead dependencies. To overcome the above challenges, we utilize a pre-trained CBraMod encoder [1] for EEG and pre-train a symmetric ECG encoder, equipping each modality with a rich foundational representation. These representations are then fused via simple embedding concatenation, allowing the classification head to learn cross-modal interactions, together enabling effective downstream learning despite limited multi-modal supervision. Evaluated on emotion recognition, our approach achieves near state-of-the-art performance, demonstrating that carefully designed physiological encoders, even with straightforward fusion, substantially improve downstream performance. These results highlight the potential of foundation-model approaches to harness the holistic nature of physiological signals, enabling scalable, label-efficient, and generalizable solutions for healthcare and affective computing.

1 Introduction

Physiological signals represent a rich source of information about human health, cognitive states, and emotional responses. Among these, electrocardiograms (ECG) and electroencephalograms (EEG) offer complementary perspectives on human physiology—ECG capturing cardiac dynamics that reflect autonomic nervous system activity, while EEG provides direct measurements of neural oscillations underlying cognitive and emotional processes.

The integration of these modalities holds significant promise for advancing healthcare diagnostics, brain-computer interfaces, and affective computing applications. Despite this potential, multi-modal physiological signal analysis faces several fundamental challenges. First, the scarcity of multi-modal labeled datasets limits the development of supervised learning approaches, as simultaneous recording

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundation Models for the Brain and Body.

of multiple physiological signals with expert annotations remains expensive and time-intensive. Second, the inherent differences in signal characteristics—ECG's relatively regular morphology versus EEG's complex spatiotemporal patterns—create modality-specific representational gaps that complicate joint analysis. Third, traditional approaches often rely on hand-crafted features and task-specific architectures, limiting their generalizability across diverse applications and patient populations.

Recent advances in foundation models have demonstrated remarkable success in learning universal representations across domains such as natural language processing and computer vision [3]. These models leverage self-supervised pretraining on large-scale unlabeled data to capture generalizable patterns that can be adapted to downstream tasks with minimal supervision. In the physiological signal domain, this paradigm offers particular promise given the abundance of unlabeled biosignal data and the potential for transferable representations that capture fundamental physiological processes.

However, adapting foundation model approaches to physiological signals presents unique challenges. Unlike text or images, physiological signals exhibit complex temporal dependencies, multi-scale patterns, and inherent variability across subjects and recording conditions. Furthermore, the effective integration of multiple physiological modalities requires careful consideration of how to align and fuse representations that capture both modality-specific characteristics and cross-modal relationships.

In this work, we address these challenges through a foundation model approach that combines self-supervised pretraining with strategic multi-modal fusion. We adapt the CBraMod architecture [1], originally designed for EEG analysis, to create a symmetric ECG encoder capable of learning robust cardiac representations. Our ECG pretraining framework introduces a dual-masking strategy that encourages the model to capture both intra-lead temporal patterns and inter-lead spatial dependencies, enabling comprehensive understanding of cardiac dynamics.

Rather than pursuing complex multi-modal alignment schemes, we demonstrate that carefully designed modality-specific encoders combined with straightforward embedding concatenation can achieve highly effective cross-modal learning. This approach leverages the rich foundational representations learned by each encoder while allowing the downstream classification head to discover cross-modal interactions relevant to the target task. We evaluate our framework on emotion recognition using the DREAMER dataset [2], where our approach achieves near state-of-the-art performance across valence, arousal, and dominance dimensions. Our results demonstrate that foundation model approaches can effectively harness the complementary nature of physiological signals, providing a scalable and label-efficient pathway for multi-modal biosignal analysis.

2 Method

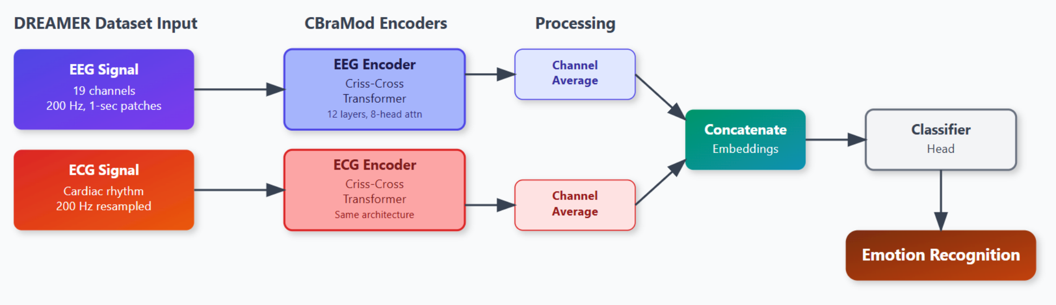

Figure 1: Architecture of the proposed multimodal emotion recognition framework. Electroencephalogram (EEG) and electrocardiogram (ECG) signals are processed by separate encoders to extract modality-specific features. The resulting embeddings are averaged, concatenated, and passed to a classifier for emotion prediction.

Figure 1 provides an overview of our multi-modal framework for emotion recognition using ECG and EEG signals from the DREAMER dataset [2]. EEG and ECG signals are first processed by separate respective encoders with similar architectures, capturing modality-specific features. After encoding, a channel-wise averaging step reduces each representation to a compact embedding. The embeddings from both modalities are then concatenated and passed through a lightweight classifier head, which learns cross-modal interactions to perform the downstream emotion recognition task.

2.1 ECG Pre-training

To address the scarcity of large-scale multimodal datasets, we pre-train an encoder tailored for ECG signals, ensuring it can provide rich and transferable representations.

Architecture We adopt the CBraMod [1] architecture, specifically a criss-cross transformer with 12 layers and 8 attention heads, where input ECG patches are first embedded and enriched with positional encodings. The criss-cross attention mechanism enables the model to jointly capture local temporal dynamics and global contextual dependencies. This design is kept identical to the EEG encoder, ensuring architectural symmetry across modalities and facilitating comparable representation spaces for downstream fusion.

Pre-processing For ECG pre-training, we employ a multi-stage preprocessing pipeline to ensure signal quality and robustness across acquisition settings. First, a Butterworth bandpass filter (0.5–40 Hz) is applied to isolate clinically relevant ECG components while suppressing baseline wander and high-frequency noise, followed by a notch filter (50 Hz or 60 Hz, depending on region) to attenuate powerline interference. To address variability in acquisition hardware, we adopt a multi-rate resampling strategy, standardizing signals to 100, 200, 500, and 1000 Hz, which exposes the model to diverse sampling resolutions and promotes invariant feature learning. Each ECG is then segmented into fixed 200-sample windows and reshaped into tensors of size (12, 10, 200), corresponding to 12 clinical leads, 10 temporal patches, and 200 samples per patch, with leads reordered to follow the conventional clinical configuration.

ECG Patching & Dual-masking We partition each ECG sample $S \in \mathbb{R}^{C \times T}$, where C denotes the number of leads and T the number of timepoints, into $n = \lfloor T/t \rfloor$ temporal patches per channel, resulting in the patch set:

$$X = \{x_{i,j} \mid i \in [1, \ldots, C],\ j \in [1, \ldots, n]\}. \tag{1}$$

To enhance robustness across both spatial (lead-wise) and temporal contexts, we apply a *dual-masking* scheme:

- **Patch masking**: a binary mask $M \in \{0,1\}^{C \times n}$ is generated, with each $m_{i,j}$ sampled from a Bernoulli distribution with masking ratio r; if $m_{i,j} = 1$, the corresponding patch $x_{i,j}$ is masked.
- **Channel masking**: independently, we select a subset of channels $\mathcal{C}_m \subseteq \{1, \ldots, C\}$ using masking proportion r_c; all patches in these channels are masked to encourage learning across lead configurations.

The resulting masked patch $\tilde{x}_{i,j}$ is defined as:

$$\tilde{x}_{i,j} = \begin{cases} x_M, & \text{if } m_{i,j} = 1 \text{ or } i \in \mathcal{C}_m, \\ x_{i,j}, & \text{otherwise,} \end{cases} \tag{2}$$

and the full masked input becomes:

$$\tilde{X} = \{\tilde{x}_{i,j} \mid i \in [1, \ldots, C],\ j \in [1, \ldots, n]\}. \tag{3}$$

This dual-masking encourages the encoder to learn both local temporal patterns and global inter-lead dependencies, improving robustness to channel-specific artifacts and missing data.

Learning Objective The learning objective in our ECG pre-training is designed to capture both fine-grained temporal dynamics and inter-lead dependencies by employing a composite mean squared error (MSE) loss. Similar to CBraMod [1], a reconstruction head is used for reconstructing the masked patches and channels.

We use the same patch reconstruction loss as in CBraMod and add a new channel reconstruction loss, which ensures the model can recover signals from entirely masked leads. we compute the reconstruction loss for channels that were masked entirely and those that were not:

$$\mathcal{L}_{\text{channel}} = \frac{1}{2}\left(\|\hat{X}^c_M - X^c_M\|_2^2 + \|\hat{X}^c_{\bar{M}} - X^c_{\bar{M}}\|_2^2\right) \tag{4}$$

where X^c_M and $\hat{X}^c_M$ represent the masked channel data and its reconstruction, and $X^c_{\bar{M}}$, $\hat{X}^c_{\bar{M}}$ denote the unmasked channels.

To promote balanced learning, the loss is computed over both masked and inverse-masked regions, assigning equal weight to each.

The total loss is an equally weighted summation of both components:

$$\mathcal{L}_{\text{total}} = \frac{1}{2}\mathcal{L}_{\text{patch}} + \frac{1}{2}\mathcal{L}_{\text{channel}} \tag{5}$$

This loss design ensures the model learns to reconstruct both localized temporal content and complete signals across leads, leading to generalized and transferable ECG representations.

2.2 Multi-modal Fusion Framework

To leverage complementary information from ECG and EEG signals, we adopt a dual-backbone approach, employing our pre-trained ECG foundational model as the ECG encoder and the CBraMod model [1] as the EEG encoder. The embeddings produced by each backbone are then fused through straightforward concatenation, which is subsequently passed to a feed-forward classification head.

Let:

$$\mathbf{E}_{\text{ECG}} = f_{\text{ECG}}(\mathbf{X}_{\text{ECG}}) \tag{6}$$

be the embedding produced by the pre-trained ECG backbone for an input ECG signal $\mathbf{X}_{\text{ECG}}$, and

$$\mathbf{E}_{\text{EEG}} = f_{\text{EEG}}(\mathbf{X}_{\text{EEG}}) \tag{7}$$

be the embedding produced by the pre-trained CBraMod EEG backbone for an input EEG signal $\mathbf{X}_{\text{EEG}}$.

Fusion via concatenation:

$$\mathbf{E}_{\text{fusion}} = \text{Concat}(\mathbf{E}_{\text{ECG}}, \mathbf{E}_{\text{EEG}}) \tag{8}$$

Classification head:

$$\hat{\mathbf{y}} = g(\mathbf{E}_{\text{fusion}}) = g(\text{Concat}(f_{\text{ECG}}(\mathbf{X}_{\text{ECG}}), f_{\text{EEG}}(\mathbf{X}_{\text{EEG}}))) \tag{9}$$

where $g(\cdot)$ denotes a feed-forward network that maps the fused embedding to the task-specific prediction $\hat{\mathbf{y}}$.

This framework enables the classification network to learn a unified representation that captures cross-modal interactions, making it well-suited for downstream tasks that require multi-modal information while preserving meaningful modality-specific features through the respective encoders.

3 Experiments & Results

3.1 ECG Pretraining Setup

We utilize the PhysioNet/Computing in Cardiology Challenge 2021 dataset [4], which is a collection of seven distinct datasets, for ECG pretraining, comprising 88,253 12-lead ECG recordings from diverse geographic and clinical sources across three continents.. This multi-dataset composition introduces substantial variability in patient demographics, healthcare settings, and device configurations, with recordings varying in length (up to 60 seconds) and sampling rates (500–1000 Hz). During self-supervised pretraining, diagnostic labels are omitted to learn task-agnostic representations purely from signal structure and temporal-spatial patterns. All pretraining experiments are conducted on dual NVIDIA A100 GPUs.

3.2 Evaluation on Emotion Recognition

We evaluate our approach on the DREAMER dataset [2], which contains EEG and ECG recordings from 23 subjects watching emotional video stimuli. Each subject viewed 18 video clips while continuous EEG (14 channels at 128 Hz) and ECG (2 channels at 256 Hz) were recorded. Emotional responses were self-assessed on continuous scales for valence, arousal, and dominance, which we discretize into binary classification tasks (Low: < 3, High: ≥ 3).

Our model processes both modalities through independent CBraMod encoders, where the EEG encoder uses pretrained weights from the Temple University Hospital EEG Corpus (TUEG) and the ECG encoder uses our self-supervised pretraining weights. Embeddings are averaged across channels, concatenated, and passed to three parallel binary classification heads. We employ subject-independent 3:1:1 train/validation/test splitting and train using Adam optimizer with learning rate 10^{-3} for 10 epochs.

3.3 Results

Table 1 compares our approach against state-of-the-art methods on the DREAMER dataset. Our pretrained model achieves competitive performance across all emotional dimensions, with particularly strong results in AUC metrics for arousal (84.79) and dominance (86.69), ranking best and second-best respectively. For valence prediction, we achieve second-best accuracy (69.44%) and best F1-score (81.14), demonstrating strong discriminative power while approaching the performance of Brant-X [14] with a more efficient architecture.

Table 1: Performance comparison on DREAMER dataset. Values show mean across subjects. **Bold** = best, Underline = second-best.

Method	**Valence**			**Arousal**			**Dominance**		
	Acc.	F1	AUC	Acc.	F1	AUC	Acc.	F1	AUC
TF-C [5]	66.20	78.09	69.71	76.45	85.86	80.40	78.17	87.01	85.20
SimMTM [6]	63.84	75.52	69.73	76.16	86.21	76.42	78.54	87.81	82.84
OneFitsAll [7]	63.51	76.12	68.26	73.88	84.15	78.64	77.41	86.92	85.59
Time-LLM [8]	68.03	72.22	**80.83**	76.39	85.63	80.00	78.47	85.92	79.68
MiniRocket [9]	60.54	65.68	64.36	75.73	85.75	77.90	75.11	85.28	86.69
MLF-CapsNet [10]	65.67	77.06	71.05	74.56	84.98	79.80	77.13	86.94	82.61
EEGConformer [11]	59.82	69.53	71.94	73.07	83.21	75.11	81.82	89.50	83.19
Lin et al. [12]	66.47	79.50	67.10	75.54	85.87	79.06	78.46	87.83	79.40
Wang et al. [13]	66.95	79.84	66.20	76.47	86.44	80.29	81.87	89.96	83.97
Brant-X [14]	**70.61**	80.51	72.48	**78.64**	**87.59**	82.14	**83.54**	**90.97**	**90.19**
Ours	69.44	**81.14**	68.69	75.00	84.21	**84.79**	79.63	88.30	86.69

4 Conclusion

In this work, we presented a multi-modal fusion framework that leverages foundation models for both ECG and EEG, addressing the challenges of limited labeled data and modality-specific differences. By pretraining a symmetric ECG encoder and employing the CBraMod EEG encoder, we obtained rich modality-specific representations, which were fused via simple concatenation to enable effective cross-modal learning. Our approach demonstrated near state-of-the-art performance on emotion recognition, underscoring the effectiveness of combining carefully designed physiological encoders with straightforward fusion strategies. These findings highlight the potential of foundation-model approaches to advance scalable, label-efficient, and generalizable solutions for healthcare and affective computing.

References

[1] Wang, J., Zhao, S., Luo, Z., Zhou, Y., Jiang, H., Li, S., Li, T., & Pan, G. (2025). CBraMod: A Criss-Cross Brain Foundation Model for EEG Decoding. *The Thirteenth International Conference on Learning Representations*.

[2] Katsigiannis, S., & Ramzan, N. (2018). DREAMER: A Database for Emotion Recognition Through EEG and ECG Signals From Wireless Low-cost Off-the-Shelf Devices. *IEEE Journal of Biomedical and Health Informatics*, 22(1), 98–107.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[4] Moody, G. B., Li, Q., Johnson, A. E. W., Mark, R. G., & Clifford, G. D. (2021). PhysioNet/Computing in Cardiology Challenge 2021. `https://physionet.org/content/challenge-2021/1.0.3/`.

[5] Zhang, X., Zhao, Z., Tsiligkaridis, T., & Zitnik, M. (2022). Self-Supervised Contrastive Pre-Training For Time Series via Time-Frequency Consistency. *arXiv preprint arXiv:2206.08496*.

[6] Dong, J., Wu, H., Zhang, H., Zhang, L., Wang, J., & Long, M. (2023). SimMTM: A Simple Pre-Training Framework for Masked Time-Series Modeling. *Advances in Neural Information Processing Systems*, 36, 29996–30025.

[7] Zhou, T., Niu, P., Wang, X., Sun, L., & Jin, R. (2023). One Fits All: Power General Time Series Analysis by Pretrained LM. *arXiv preprint arXiv:2302.11939*.

[8] Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., Shi, X., Chen, P.-Y., Liang, Y., Li, Y.-F., Pan, S., & Wen, Q. (2024). Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. *arXiv preprint arXiv:2310.01728*.

[9] Dempster, A., Schmidt, D. F., & Webb, G. I. (2021). MiniRocket: A Very Fast (Almost) Deterministic Transform for Time Series Classification. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 248–257.

[10] Liu, Y., Ding, Y., Li, C., Cheng, J., Song, R., Wan, F., & Chen, X. (2020). Multi-channel EEG-based emotion recognition via a multi-level features guided capsule network. *Computers in Biology and Medicine*, 123, 103927.

[11] Song, Y., Zheng, Q., Liu, B., & Gao, X. (2023). EEG Conformer: Convolutional Transformer for EEG Decoding and Visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31, 710–719.

[12] Lin, X., Chen, J., Ma, W., Tang, W., & Wang, Y. (2023). EEG emotion recognition using improved graph neural network with channel selection. *Computer Methods and Programs in Biomedicine*, 231, 107380.

[13] Wang, X., Zhang, J., He, C., Wu, H., & Cheng, L. (2024). A novel emotion recognition method based on the feature fusion of single-lead EEG and ECG signals. *IEEE Internet of Things Journal*, 11(5), 8746–8756.

[14] Zhang, D., Yuan, Z., Chen, J., Chen, K., & Yang, Y. (2024). Brant-X: A Unified Physiological Signal Alignment Framework. *arXiv preprint arXiv:2409.00122*.

A Pre-training Additional Details

To evaluate the effectiveness of our self-supervised pretraining approach, we monitor the reconstruction loss—specifically, the mean squared error (MSE) between the masked input segments and the model's reconstructed outputs. This metric directly reflects the model's ability to learn meaningful latent representations of the ECG signal, capturing both temporal and inter-lead dependencies.

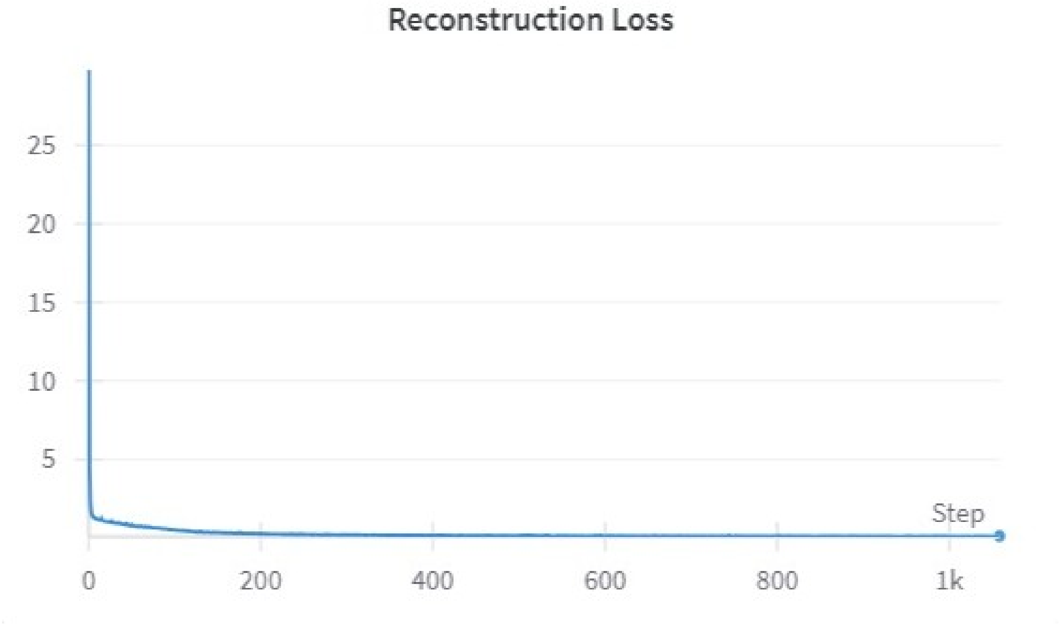

Figure 2: Reconstruction loss in pre-training

As shown in Figure 2, the reconstruction loss dropped significantly from an initial value of approximately 30 to a final value of 0.1136. This steady decline indicates that the model has progressively improved its understanding of the underlying ECG structures, even in the absence of supervised labels.

B Training details for the multi-modal experiment

This section provides details about the experimental setup and plots for the multi-modal fusion experiment.

Table 2: Training Hyperparameters

Parameter	**Value**
Epochs	10
Batch size	8
Initial learning rate	1×10^{-3}
Classifier dropout	0.2
Optimizer	Adam
LR Schedular	StepLR
Gamma	0.1
Step Size	4 epochs
Loss Function	BCE Loss

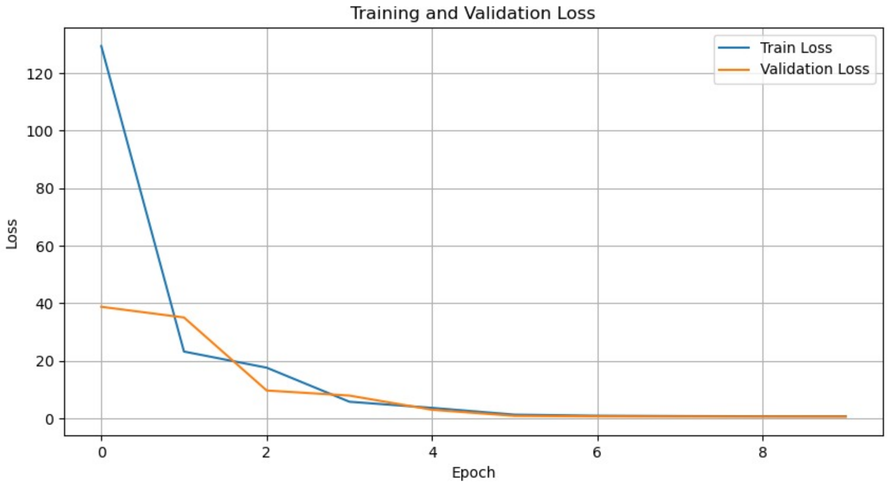

Figure 3: Training and validation loss curves over epochs for the multi-modal fusion experiment on emotion recognition task using Dreamer dataset.

C Limitations and Future Work

Despite the promising results, our work has several limitations. First, the fusion strategy relies on simple embedding concatenation, which, while effective, may not fully exploit fine-grained temporal or cross-modal dependencies. Second, our evaluation is constrained to emotion recognition task, leaving open questions regarding generalization to broader physiological applications and more diverse populations. Future work will explore more advanced fusion mechanisms, extend evaluation to additional downstream tasks, and investigate new strategies for mitigating data scarcity, such as weak supervision or synthetic data augmentation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims are made at the end of the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in Appendix C

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This paper focuses on a domain application. No proofs or theory assumptions provided in this paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper discloses all necessary information to reproduce the main experi mental results in sections 3.1 and 3.2. Additionaly, researchers can follow the provided code to re-produce the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: Upon acceptance, we will release a public GitHub repository containing the full implementation and instructions necessary to reproduce the main experimental results. All datasets employed in this work are standard publicly available benchmarks, with references and access links provided in the paper, ensuring that the experiments can be faithfully reproduced.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper mentions the experimental details in sections 3.1 and 3.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper doesn‘t include error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Pre-training compute resources are mentioned in section 3.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our work adheres to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: In our opinion, there is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks. We think the data and models have probably no risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Cbramod model used in this work is properly credited. The license (MIT license) is mentioned, respected, and present in the Github Repository. All data/models used are open-source and referenced properly.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: datasets,code and model details are provided in the sections 2,3 and in the supplementary materials.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: As noted above, the paper does not involve any research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Our paper does not involve LLMs.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.