Modality-Aware Multimodal Fusion for Sleep Event Classification

Kevin Hayes
Department of Computer Science
University of North Carolina at Greensboro
Greensboro, North Carolina, USA
Email: khhayes2@uncg.edu

Yingcheng Sun*
Department of Computer Science
University of North Carolina at Greensboro
Greensboro, North Carolina, USA
Email: y_sun4@uncg.edu

***Abstract*—Sleep event classification plays a crucial role in diagnosing and managing neurological and sleep-related disorders. However, the underlying data are inherently multimodal—ranging from EEG waveforms and structured metadata to free-text clinical notes—posing significant challenges for scalable, generalizable machine learning systems. Most existing approaches either ignore this heterogeneity or require fixed architectures tailored to specific modality combinations.**

We present a flexible, modality-aware multimodal fusion framework for sleep event classification that integrates heterogeneous data using pretrained submodels and transformer-based fusion strategies. Our system supports diverse modalities—including EEG (EDF), clinical notes, tabular metadata, and synthetic image encodings—through a plug-and-play architecture that allows rapid testing and adaptation. We introduce and compare four variants of our architecture based on two key design axes: fusion timing (early vs. late) and representation strategy (lightweight vs. modality-specific encoders).

Our evaluation spans both real-world data from the CAP Sleep Database and synthetic data designed to expose model-level decision boundaries. While real data experiments are constrained by limited training examples and modality-specific bottlenecks, our synthetic experiments show that late fusion with modality-specific encoders significantly improves accuracy (up to 93%) over early fusion or naive concatenation strategies. This work contributes a modular and extensible foundation for multimodal learning in clinical domains, highlighting key trade-offs in fusion design and laying the groundwork for future extensions in more complex, real-world biomedical tasks.

***Index Terms*—Multimodal systems, Multi-modal recognition, Sleep Event Classification,**

I. Introduction

Understanding and classifying sleep events is essential for diagnosing neurological conditions such as epilepsy, narcolepsy, insomnia, and sleep apnea. [1] These events are often identified through a combination of physiological signals (e.g., EEG), structured metadata (e.g., sleep stage, electrode location), and unstructured inputs such as clinical notes. As a result, effective computational support for sleep event analysis requires the ability to process heterogeneous, multimodal data.

Traditional machine learning approaches tend to rely on a single modality (e.g., EEG time-series) [2] or require tightly coupled model architectures, limiting their adaptability and reuse. Meanwhile, recent advances in *multimodal transformer architectures* offer flexible mechanisms to fuse diverse inputs—but many still depend on fixed modality combinations or large end-to-end pretraining, which is difficult in the medical domain due to data scarcity and privacy constraints.

In this work, we introduce a modular and modality-aware multimodal fusion framework for sleep event classification. Our architecture is designed to support diverse data modalities through interchangeable submodules and to explore the effects of different fusion strategies. We instantiate four model variants across two design axes: (1) fusion timing—*early* (concatenation before fusion) vs. *late* (independent modality encoders followed by fusion), and (2) representation design—*lightweight* vs. *modality-specific* transformer encoders.

This framework is inspired by and extends the ideas from Meta-Transformer [3] and LLaMA-Adapter [4], with a focus on modular flexibility and clinical applicability. Our problem setting is grounded in the CAP Sleep Database [5], which contains EEG and structured sleep annotations, and is supplemented by synthetic modalities as discussed in [6].

To summarize, our main contributions are:

- **A modular multimodal architecture** for biomedical event classification that supports plug-and-play modality integration.
- **Four architectural variants** combining early vs. late fusion with lightweight vs. modality-specific encoders, enabling systematic exploration of design trade-offs.
- **Evaluation on both real-world and synthetic datasets**[1], demonstrating the superiority of late fusion strategies with pretrained encoders under constrained conditions.

II. Related Work

Multimodal learning has been widely explored, particularly in areas such as visual question answering, speech-text alignment, and video-language tasks [7]. In the medical domain, recent studies have applied multimodal fusion to tasks like disease prediction, report generation, and survival analysis [8]. However, most prior work targets either well-aligned

*Corresponding author: Yingcheng Sun (y_sun4@uncg.edu).

[1]Code: https://github.com/Kevin-Hayes-UNCG/Combination_transformer

modalities or relies on high-quality paired datasets, which are not always available in real-world clinical applications.

Classical fusion strategies are broadly categorized into early, late, and hybrid fusion [9]. Early fusion (also called feature-level fusion) merges raw or embedded features before model training, allowing joint representation learning but often suffering from modality imbalance and alignment issues. Late fusion (decision-level) combines modality-specific predictions and has the advantage of modularity and robustness to missing data, though it may lose cross-modal synergy. Several works propose hybrid schemes involving cross-attention [10], gating mechanisms [11], or graph-based reasoning [12] to balance these trade-offs.

In biomedical signal analysis, most fusion work focuses on combining signals like ECG and EMG [2] or multimodal imaging [13], but fewer have tackled the challenges of integrating heterogeneous data types such as EEG, tabular clinical data, and unstructured text. Our work contributes to this space by proposing a modular and modality-aware architecture capable of operating on diverse and partially missing modalities, which is especially relevant in sleep medicine and critical care.

Compared to prior multimodal benchmarks like MIMIC-CXR [14] or PhysioNet 2021 [15], our study incorporates both controlled synthetic evaluations and real-world clinical EEG data, emphasizing the fusion mechanism itself rather than task-specific fine-tuning. This positions our work as a systematic foundation for future studies on multimodal architecture design in healthcare.

III. Methodology

Our architecture adopts a modular fusion framework tailored to low-resource, heterogeneous clinical data settings. The system consists of three stages: (1) *modality-specific encoding*, (2) *fusion*, and (3) *prediction*. Let $\mathcal{X} = \{x^{(1)}, x^{(2)}, \ldots, x^{(M)}\}$ denote the set of M modalities for each input sample. The full pipeline transforms each modality $x^{(m)}$ into a latent representation $z^{(m)}$ for downstream integration and classification.

A. Modality-Specific Encoding

Each input modality $x^{(m)}$ is processed through a dedicated encoder $f^{(m)}$, producing an intermediate representation:

$$h^{(m)} = f^{(m)}(x^{(m)}) \in \mathbb{R}^{L_m \times d} \tag{1}$$

where L_m is the token length and d is the shared hidden size across modalities.

Text: Unstructured or semi-structured text (e.g., clinician annotations, sleep stage descriptions, or event labels) is first tokenized and then transformed into dense semantic vectors, producing contextual embeddings $h^{(txt)}$.

Time-series signals: Physiological recordings like EEG signals stored in EDF format, are first preprocessed into log-mel spectrograms, which preserve both temporal dynamics and spectral energy patterns relevant to sleep events. These time–frequency representations are then encoded using the audio transformer, producing $h^{(eeg)}$.

Image: Image-based inputs such as hypnogram plots or visual snapshots of EEG event detections are processed using a pretrained Vision Transformer convert images into patch-level embeddings and outputs, producing $h^{(img)}$.

Numerical: The numerical data are metadata associated with each sleep event, such as sleep stage, event type or event duration. These discrete metadata fields capture clinically meaningful attributes about each event. We encode them as embeddings $h^{(num)}$.

B. Fusion Strategy

We explore two fusion strategies:

a) Early Fusion: All $h^{(m)}$ are concatenated along the token dimension to form:

$$H = \text{Concat}\left(h^{(1)}, h^{(2)}, \ldots, h^{(M)}\right) \in \mathbb{R}^{L \times d} \tag{2}$$

where $L = \sum_{m=1}^{M} L_m$. This unified sequence H is passed through a shared transformer encoder T_{shared}:

$$Z = T_{shared}(H) \tag{3}$$

Final representation z is derived via mean pooling or a special [CLS] token.

b) Late Fusion: Each modality is first passed through a modality-specific transformer $T^{(m)}$:

$$z^{(m)} = T^{(m)}(h^{(m)}) \in \mathbb{R}^{L_m \times d} \tag{4}$$

A cross-modal attention block $\mathcal{A}$ then computes the fused representation:

$$z = \mathcal{A}(z^{(1)}, z^{(2)}, \ldots, z^{(M)}) \tag{5}$$

C. Prediction and Optimization

The fused vector z is passed through a classification head:

$$\hat{y} = \text{softmax}(Wz + b) \tag{6}$$

where $W \in \mathbb{R}^{C \times d}$ and C is the number of output classes. The model is trained using cross-entropy loss:

$$\mathcal{L} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \tag{7}$$

D. Model Variants

We design and evaluate four architectural variants derived from two principal design axes: the timing of fusion and the complexity of modality-specific encoders. These combinations allow us to investigate how early versus late integration of modality streams and shallow versus deep encoding affect downstream classification performance in both synthetic and clinical contexts.

The first design axis—*fusion timing*—determines when the multimodal information streams are merged during processing. In the early fusion strategy, modality-specific features are concatenated or embedded into a joint representation before being passed through a shared transformer encoder. This approach encourages early interactions and potential synergies

TABLE I
COMPARISON OF MULTIMODAL FUSION MODEL VARIANTS ALONG THE TWO DESIGN AXES: FUSION TIMING (EARLY VS. LATE) AND ENCODER COMPLEXITY (LIGHTWEIGHT VS. DEEP MODALITY-AWARE).

Model	Fusion Timing	Modality Encoders	Cross-Modal Attention	Encoder Depth
Basic (Lightweight) Early Fusion	Early	Tokenizers / minimal embeddings	No	Low
Basic (Lightweight) Late Fusion	Late	Frozen pretrained encoders	No	Low
Enhanced (Modality-Aware) Early Fusion	Early	Trainable modality-specific encoders	Yes	High
Enhanced (Modality-Aware) Late Fusion	Late	Pretrained + trainable encoders	Yes	High

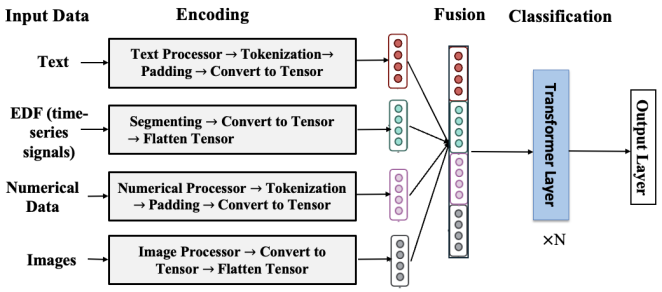


Fig. 1. Computation graph of the Basic Early Fusion model. Inputs are tokenized and concatenated before passing through a shared transformer stack.

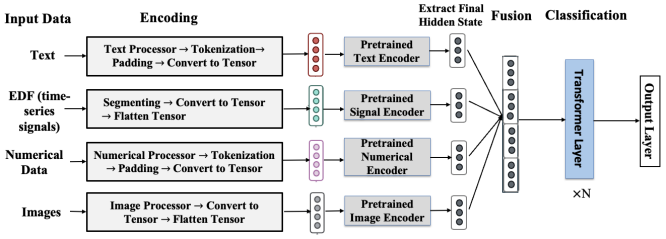


Fig. 2. Computation graph of the Basic Late Fusion model. Each modality is encoded by a frozen submodel, concatenated, and passed through a shallow transformer.

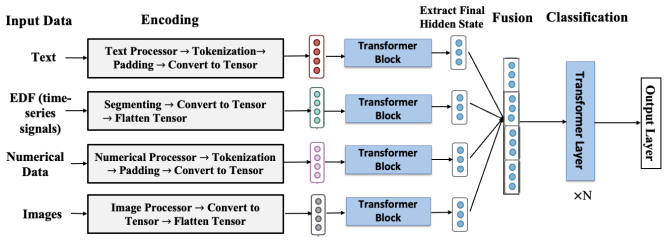


Fig. 3. Computation graph of the Enhanced (Modality-Aware) Early Fusion model. Modality-specific transformer blocks are used before fusion.

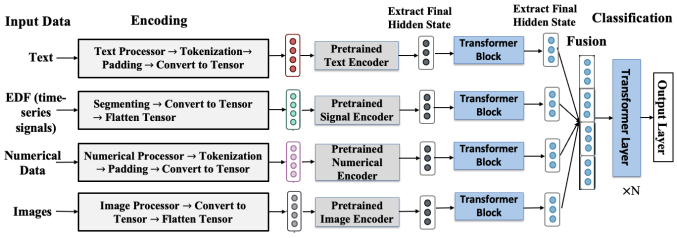


Fig. 4. Computation graph of the Enhanced (Modality-Aware) Late Fusion model. Each modality is processed by a frozen encoder and a transformer block before attention-based fusion.

between modalities but may dilute domain-specific information, especially when modalities are semantically heterogeneous. In contrast, late fusion defers the integration process until each modality has been independently encoded, typically via separate transformer or pretrained backbone networks. The resulting modality-specific embeddings are then aligned and aggregated via attention pooling, concatenation, or learned fusion layers. This strategy preserves semantic integrity of each modality and facilitates more interpretable and modular representation learning.

The second design axis—*encoder complexity*—governs the depth and specialization of processing applied to each modality prior to fusion. In lightweight models, raw inputs are either tokenized (e.g., clinical text or structured metadata) or embedded via frozen pretrained models (e.g., ViT for images), with minimal learnable parameters specific to each modality. This leads to efficient training and reduced memory usage, though it limits the model's capacity to extract rich, domain-aware features. By contrast, deep modality-aware models allocate modality-specific transformer blocks or adapters that transform raw inputs into deeper semantic embeddings. These components are trained end-to-end, enabling the model to learn modality-specific inductive biases and improve discriminability for tasks like EEG-based event classification.

These two axes yield four variants: (1) **Basic Early Fusion**, in which all tokenized or minimally embedded modalities are concatenated and passed through a shared transformer, as shown in Fig. 1; (2) **Basic Late Fusion**, where each modality is independently encoded by a frozen pretrained model and their outputs are directly concatenated for classification, as shown in Fig. 2; (3) **Enhanced Early Fusion**, which augments the basic early model with lightweight modality-specific encoders prior to shared transformer modeling, as Fig. 3 shown; and (4) **Enhanced Late Fusion**, which applies trainable deep modality-specific encoders followed by attention-based fusion, as Fig. 4 shown. This most expressive variant supports both rich per-modality representation learning and structured cross-modal interaction. The distinctions among these designs are formally summarized in Table I and guide our empirical analysis in both synthetic and clinical settings.

IV. Experiments and Results

A. Dataset

Our experiments used a hybrid dataset consisting of real sleep data from the CAP Sleep Database [5] and additional synthetic modalities created to support a complete multimodal fusion study.

The real portion of the dataset included four modalities obtained or derived from the CAP Sleep Database: EDF signals, numerical metadata, synthetic clinical text, and control images. Because the CAP Sleep Database does not include narrative clinician notes, synthetic doctor-style descriptions were generated for each event using GPT-4o. These notes approximate the interpretive text commonly found in clinical sleep studies and provide a simulated text modality.

To evaluate whether the fusion models could learn structured decision boundaries, we created a synthetic dataset aligned with the EDF preprocessing pipeline. We generated 20,000 feature pairs (x_1, x_2), each sampled uniformly from $[0, 500]$. For compatibility with the Distil-AST encoder, which requires fixed-length input vectors, each two-dimensional pair was padded with 500 zeros. This synthetic dataset provided a controlled setting to probe the ability of the multimodal models to learn nontrivial feature interactions.

B. Experimental Setup

For the lightweight fusion variants, each modality was encoded using pretrained models without additional modality-specific transformer blocks. The EDF modality was processed using the audio transformer [16], which converts log-mel spectrograms into compact embeddings. Numerical metadata were converted into short textual descriptions and tokenized using the same tokenizer employed for the text modality. Both the numerical and text modalities used the tokenizer and text encoder [17]. The image modality was represented using the pretrained Vision Transformer [18], which produces embeddings from 224×224 image patches.

All experiments were conducted on the UNCG Computer Science high-performance computing server *knuth*. The system is equipped with 1.5 TiB of RAM and two CPU cores operating at approximately 3.7 GHz. GPU acceleration was provided by four NVIDIA H100 PCIe GPUs, each with sufficient memory to support multimodal transformer fine-tuning. This environment enabled efficient large-batch training and supported the computational demands of the EDF encoder, text encoder, image encoder, and multimodal fusion transformers. We run each experiment three times and report mean performance on the metrics:

- **Accuracy:** Overall correctness of predictions.
- **Cross-Entropy Loss:** Evaluated per epoch.
- **F1 Score:** Harmonic mean of precision and recall for multiclass classification.

C. Results and Analysis

We first evaluate our models on the CAP Sleep Database. Table II reports performance averaged across three runs. Enhanced Late Fusion achieves the best accuracy (65.17%), macro F1 score (0.4936), and lowest cross-entropy loss (1.6923), showing the advantage of modality-specific transformers and late fusion design. As shown in Table II, Enhanced Late Fusion again performs best with an average accuracy of 65.17%, macro F1 of 0.4936, and cross-entropy loss of 1.6923.

TABLE II
CAP Sleep Dataset performance (avg. of 3 runs).

Model	Acc.	F1	Loss
Basic Early Fusion	57.04	0.4427	1.9485
Basic Late Fusion	62.40	0.4727	1.7667
Enhanced Early Fusion	61.50	0.4634	1.8154
Enhanced Late Fusion	**65.17**	**0.4936**	**1.6923**

On synthetic data, modality-aware late fusion models outperform all other architectures, particularly when the decision rule involves nonlinear interactions. We first validate model behavior using a controlled synthetic dataset. Each input consists of four modalities with scalar values sampled from normal distributions. Labels are generated using a logic-based rule (e.g., AND, XOR) applied to the modalities. This setting isolates the effect of fusion strategies and model capacity without real-world noise. Table III shows that Enhanced Late Fusion outperforms all other designs, with the highest accuracy (93.37%) and lowest loss (0.1634).

TABLE III
Synthetic dataset performance for binary classification (AND logic).

Model	Accuracy (%)	Loss
Basic Early Fusion	89.79	0.2409
Basic Late Fusion	92.44	0.1785
Enhanced Early Fusion	91.67	0.1937
Enhanced Late Fusion	**93.37**	**0.1634**

When the classification task is extended to 3 classes (e.g., using trinary logic or composite rules), the same performance trend holds. Enhanced Late Fusion again outperforms the rest, achieving 93.20% accuracy with the lowest loss of 0.2645 (Table IV).

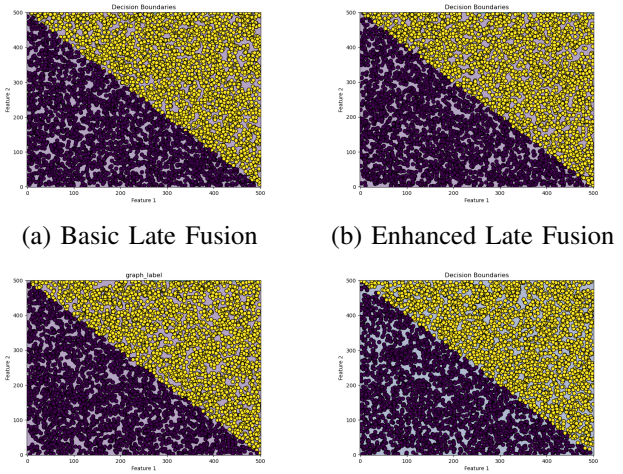
TABLE IV
Synthetic dataset performance for 3-class classification.

Model	Accuracy (%)	Loss
Basic Early Fusion	88.92	0.3655
Basic Late Fusion	91.20	0.3031
Enhanced Early Fusion	90.57	0.3218
Enhanced Late Fusion	**93.20**	**0.2645**

Figures 5 shows decision boundaries for each architecture. Enhanced Late Fusion produces smoother and more accurate separation between classes, especially in high-entropy regions.

D. Ablation Study

To better understand the contribution of each modality and architectural component, we perform a series of ablation experiments using the CAP Sleep Database. We remove one



(a) Basic Late Fusion (b) Enhanced Late Fusion

(c) Basic Early Fusion (d) Enhanced Early Fusion

Fig. 5. Decision boundaries of different fusion models.

modality at a time from the Enhanced Late Fusion model and retrain under identical settings.

Results are summarized in Table V. The removal of EEG features leads to the most significant drop in accuracy (from 65.17% to 54.82%), confirming that EEG is the primary modality for sleep classification. Interestingly, the removal of structured metadata reduces F1 score more than accuracy, suggesting its role in refining class boundaries. Text modality contributes less individually but improves robustness when combined with others.

These results support the importance of multimodal integration. Each modality carries complementary information, and their joint encoding through late fusion allows for more nuanced decisions than any single modality alone.

TABLE V
ABLATION STUDY RESULTS FOR ENHANCED LATE FUSION (CAP DATASET).

Ablated Modality	Acc. (%)	F1	Loss
None (Full Model)	**65.17**	**0.4936**	**1.6923**
w/o EEG	54.82	0.3935	2.0472
w/o Metadata	60.03	0.4217	1.8511
w/o Text	62.10	0.4591	1.7734
w/o Image	63.47	0.4780	1.7328

Our experimental results yield several important observations regarding architectural design. First, fusion timing plays a critical role in model performance. Across both synthetic and real-world datasets, late fusion consistently outperforms early fusion. This is especially evident in scenarios where the target label depends on nonlinear relationships between modalities, such as XOR-based logic. Delaying the fusion until each modality has been independently encoded allows the model to preserve and refine modality-specific features before integration, leading to better separability in decision space.

Second, encoder complexity significantly impacts generalization and interpretability. Models that include modality-specific transformer encoders—our Enhanced variants—demonstrate more accurate and smoother decision boundaries. This indicates that separate attention mechanisms per modality help disentangle the contribution of each data source, enabling more robust multimodal reasoning.

Lastly, we observe that real-world constraints such as missing modality values, noise, and class imbalance can dampen the benefits of more complex architectures. Although Enhanced Late Fusion remains the best-performing model on the CAP Sleep dataset, the performance gap between models narrows compared to synthetic experiments. This suggests that practical deployment of such systems will require additional strategies for handling data incompleteness and modality dropout.

V. DISCUSSION

This work introduces a modular, modality-aware framework for multimodal sleep event classification, tailored for medical settings characterized by signal heterogeneity, missing data, and limited annotation. Our architecture is designed around two key axes: fusion timing (early vs. late) and encoder complexity (lightweight vs. modality-aware). Through extensive experimentation on both synthetic and real-world datasets, we demonstrate that late fusion with modality-aware encoders consistently yields the strongest performance, particularly in scenarios with nonlinear inter-modality dependencies.

The performance advantage of late fusion stems from its ability to preserve modality-specific representations before integration. Pretrained encoders can extract high-level features aligned with domain semantics—such as spectral EEG patterns or linguistic phrasing in clinical notes—while late fusion retains these distinct signals until higher levels of abstraction, allowing for more informed cross-modal reasoning. This behavior is most evident in synthetic datasets, where Enhanced Late Fusion achieves up to 93% accuracy under XOR-based logic. Visualizations of decision boundaries further support this finding, revealing that modality-aware late fusion leads to smoother, more confident class separations, particularly in regions with overlapping or ambiguous input clusters.

Real-world experiments on the CAP Sleep Database, however, illustrate the trade-offs involved in deploying these models in practical clinical scenarios. Despite Enhanced Late Fusion still outperforming its counterparts, performance gains are moderated by challenges such as missing modalities, class imbalance, and variability in signal quality. These findings underscore that architectural sophistication alone is insufficient—robust multimodal learning also depends heavily on data completeness and quality.

Moreover, resource constraints impose an upper bound on the model's scalability. Processing high-resolution EEG waveforms and visual embeddings requires significant memory overhead, making it essential to consider parameter-efficient strategies such as adapter tuning, low-rank updates (e.g., LoRA), or dynamic fusion policies that skip missing or noisy inputs during inference.

A. Fusion Strategy Trade-Offs

Fusion timing plays a pivotal role in how effectively multimodal information is integrated. Early fusion benefits from joint modeling of cross-modal token-level interactions from the outset but suffers when modality-specific nuances are diluted due to lack of disentangled attention. In contrast, late fusion allows each modality to preserve high-level semantic structures before fusion. It is particularly useful when modalities exhibit divergent structures—e.g., temporal EEG signals vs. clinical text—because $T^{(m)}$ can specialize its attention to modality-specific priors.

Empirically, late fusion performs better when modalities are semantically or structurally diverse, as shown in our synthetic XOR logic experiments and real-world EEG + metadata setting. However, early fusion can outperform when modalities are weakly coupled and low-dimensional (e.g., metadata + tabular signals) since it avoids redundancy and requires fewer parameters.

The performance gap between early and late fusion narrows as modality correlations become linear or when input representations are already aligned in a shared feature space. Therefore, fusion timing should be selected based on:

- Modality similarity(homogeneous vs. heterogeneous)
- Task complexity (linear vs. non-linear decision boundaries)
- Resource constraints (early fusion is typically cheaper)

B. Broader Impacts and Future Work

This work contributes a flexible and reproducible architecture for multimodal clinical signal fusion, addressing core challenges in data heterogeneity, modality integration, and real-world deployment. In doing so, it offers a foundation for developing generalizable clinical decision support systems that can adapt across diverse tasks and data settings. Although our experiments focus on sleep event classification, the underlying design is modality-agnostic and can be extended to other domains such as ICU monitoring, neurodegenerative disease detection, or multimodal triage.

From a societal perspective, multimodal AI systems have the potential to improve diagnostic accuracy, reduce clinician burden, and extend access to care in under-resourced settings. However, deploying such models in practice also raises issues around bias amplification, data privacy, and the interpretability of model decisions. Our modular architecture provides a natural basis for auditing these behaviors by isolating modality contributions and allowing fine-grained error tracing across components.

In future work, we plan to extend this framework in several directions. First, we will incorporate additional physiological signals such as ECG and EMG to capture broader diagnostic patterns. Second, we aim to implement robust missing modality handling via gated fusion or modality dropout during training. Third, we will explore trust calibration and uncertainty estimation using techniques such as Monte Carlo dropout or deep ensembles. Overall, our results highlight the need to balance model complexity with real-world feasibility. While modality-aware late fusion architectures offer state-of-the-art performance in clean, controlled settings, their effectiveness in noisy, heterogeneous clinical environments will depend on innovations not only in architecture but also in training data curation and human-in-the-loop interpretability.

ACKNOWLEDGMENTS

This work was supported by the 2025 UNCG Chancellor's Initiative for Transformative Research (CITR) Award.

REFERENCES

[1] M. M. Ohayon and C. F. Reynolds III, "Epidemiological and clinical relevance of insomnia diagnosis algorithms according to the dsm-iv and the international classification of sleep disorders (icsd)," *Sleep medicine*, vol. 10, no. 9, pp. 952–960, 2009.

[2] J. Zhao, W. Liu, X. He, and H. Li, "Multimodal sensor fusion for human activity recognition with deep learning," *Sensors*, vol. 20, no. 17, p. 4821, 2020.

[3] Y. Zhang, K. Gong, K. Zhang, H. Li, Y. Qiao, W. Ouyang, and X. Yue, "Meta-transformer: A unified framework for multimodal learning," 2023. [Online]. Available: https://arxiv.org/abs/2307.10802

[4] R. Zhang, J. Han, C. Liu, P. Gao, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, and Y. Qiao, "Llama-adapter: Efficient fine-tuning of language models with zero-init attention," 2024. [Online]. Available: https://arxiv.org/abs/2303.16199

[5] M. G. Terzano, L. Parrino, A. Sherieri, R. Chervin, S. Chokroverty, C. Guilleminault, M. Hirshkowitz, M. Mahowald, H. Moldofsky, A. Rosa, R. Thomas, and A. Walters, "Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep," *Sleep Med*, vol. 2, no. 6, pp. 537–553, Nov. 2001.

[6] K2View, "What is synthetic data generation? a practical guide to synthetic data generation tools," Dec 2024. [Online]. Available: https://www.k2view.com/what-is-synthetic-data-generation/

[7] Y. Yuan, Z. Li, and B. Zhao, "A survey of multimodal learning: Methods, applications, and future," *ACM Computing Surveys*, vol. 57, no. 7, pp. 1–34, 2025.

[8] K. Yan, T. Li, J. A. L. Marques, J. Gao, and S. J. Fong, "A review on multimodal machine learning in medical diagnostics," *Math. Biosci. Eng*, vol. 20, no. 5, pp. 8708–8726, 2023.

[9] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[10] Y.-H. H. Tsai, S. Bai, M. Yamada, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of ACL*, 2019.

[11] E. Perez, K. Cho, and Z. Dai, "Merging modalities: A lightweight gate for multimodal fusion," in *Proceedings of EMNLP*, 2021.

[12] J. Liu, F. Xu, B. Zhao, and B. Wang, "Mm-gnn: Multimodal graph neural networks for clinical outcome prediction," in *Proceedings of NeurIPS*, 2022.

[13] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," in *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, 2016, pp. 1285–1298.

[14] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "Mimic-cxr: A large publicly available database of labeled chest radiographs," *arXiv preprint arXiv:1901.07042*, 2019.

[15] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physionet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[16] A. Joyoadikusumo, D. S. Setiawan, and W. Wongso, "Bookbot/distil-ast-audioset · hugging face," Mar 2023. [Online]. Available: https://huggingface.co/bookbot/distil-ast-audioset

[17] S. Patil, P. Platen, M. Davaadorj, J. Chaumond, Lysandre, and L. Saulnier, "Openai/clip-vit-large-patch14 · hugging face." [Online]. Available: https://huggingface.co/openai/clip-vit-large-patch14

[18] R. Wightman, "Pytorch image models," https://github.com/huggingface/pytorch-image-models, 2019.