JOURNAL OF INFORMATION SYSTEMS AND TECHNOLOGY

Journal of Information Systems and Technology
ISSN 3110-4096 (online)
https://athallahpublishing.com/index.php/jaip/index

OPEN ACCESS

Vol. 2, No. 1, 2026

DOI: 10.64845/jistech.v1i2

Deep Audio-Visual Fusion with Attention Mechanisms for Multimodal Perception: A Systematic Review

Dwi Fatmasari
Universitas Negeri Surabaya, Indonesia
Corresponding Author: dwifatmasari@unesa.ac.id

A B S T R A C T

Advances in multimodal deep learning have driven growing interest in attention mechanisms that enhance audio and visual integration for tasks such as emotion recognition, event localization, and human computer interaction. This comprehensive survey synthesizes recent progress in attention based fusion methods and highlights the evolution from early fusion strategies to more advanced architectures, including self-attention, cross modal attention, co attention, and hierarchical attention. Transformer based models, in particular, now play a central role in state of the art audio visual systems because they capture long range temporal and semantic relationships across modalities. This survey examines how these mechanisms improve contextual understanding and task performance, while also identifying persistent challenges related to interpretability, robustness to noisy or missing modalities, modality imbalance, and computational efficiency. Limitations associated with dataset bias and the lack of standardized evaluation metrics are also discussed. Finally, the survey presents future research directions, including the development of cross modal transformer architectures, hierarchical attention models, and comprehensive attention diagnostics frameworks to support trustworthy and effective multimodal artificial intelligence systems.

Keywords: Multimodal Fusion, Audio-Visual Deep Learning, *Attention Mechanisms*

Received:	Revised:	Accepted:	Available online:
01.12.2025	01.02.2026	01.04.2026	26.06.2026

INTRODUCTION

Multimodal audio-visual research stands at the forefront of artificial intelligence, enabling machines to interpret, reason, and interact with the world in ways that mirror human perception. By integrating auditory and visual data, AI systems have achieved significant advances in applications such as emotion recognition, speech recognition, action recognition, and human–computer interaction (Zhang et al., 2023; Kumar & Lee, 2024). For instance, audio-visual emotion recognition systems are increasingly utilized in mental health diagnostics, autonomous systems, and affective computing, where understanding nuanced human emotions from both speech and facial expressions is essential for context-aware responses (Chen et al., 2022; Ahmed et al., 2025).

In real-world scenarios, such as video conferencing or surveillance systems, the fusion of audio and visual cues enables more accurate detection of intent, sentiment, and behavior, particularly under challenging conditions such as low illumination or background noise (Singh & Patel, 2023). In emotion and behavior recognition tasks, the integration of prosodic features, facial expressions, and motion dynamics enhances robustness against environmental distortions commonly encountered in practical applications (Rahman et al., 2022). Deep learning has become the backbone of these advancements, offering powerful architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer-based models that can learn complex and hierarchical representations from multimodal data (Zhang et al., 2023; Lee et al., 2024).

Despite these advancements, current multimodal fusion methods continue to face persistent challenges. Traditional fusion strategies—including early fusion, late fusion, and hybrid approaches—often struggle to capture complex cross-modal dependencies, leading to suboptimal performance in dynamic, noisy, or incomplete data environments (Chen et al., 2022). These limitations highlight the need for more advanced attention mechanisms capable of selective and context-aware feature integration (Kumar & Lee, 2024). Existing attention mechanisms, although effective in emphasizing salient features, frequently fail to address modality imbalance, where one modality dominates the other, thereby reducing robustness and generalizability (Ahmed et al., 2025).

Additionally, dataset inconsistencies, such as variations in annotation standards and limited diversity, further constrain the development of generalized models. Evaluation metrics in current studies are often limited to accuracy or F1-score, which do not sufficiently capture modality contribution or attention interpretability. These limitations are reflected in real-world implementations as reduced system reliability, poor adaptability to environmental noise, and limited transparency critical issues for applications in healthcare, autonomous systems, and interactive AI (Singh & Patel, 2023; Rahman et al., 2022).

Previous studies have attempted to address these challenges through various architectural and algorithmic innovations. Cross-modal attention mechanisms, particularly those based on transformer architectures and dual-attention frameworks, have improved the modeling of inter-modal relationships and enabled more effective feature integration (Zhang et al., 2023). Hierarchical attention models and dual-pathway architectures have also been proposed to capture both local and global dependencies, thereby enhancing robustness to missing or noisy modalities (Lee et al., 2024). However, these approaches often fail to fully resolve modality imbalance, as they may still prioritize dominant modalities or lack adaptability to dynamic contextual changes (Ahmed et al., 2025).

Interpretability remains a major concern in multimodal systems. While attention maps provide some level of transparency, the underlying decision-making processes are often difficult to interpret, limiting trust and diagnosability. Furthermore, advanced architectures such as cross-modal transformers and co-attention networks have not been comprehensively evaluated in terms of their ability to handle attention failures or provide meaningful interpretability in complex real-world scenarios (Chen et al., 2022; Kumar & Lee, 2024).

A critical research gap therefore exists in the systematic evaluation and mapping of advanced attention mechanisms in multimodal audio-visual deep learning. Current literature lacks comprehensive reviews that analyze modality-specific attention, cross-modal transformers, hierarchical attention, and the conditions under which these mechanisms succeed or fail. Moreover, there is an absence of standardized evaluation metrics for measuring attention interpretability and modality contribution, which hinders objective comparison across different models and approaches. Addressing this gap is essential for advancing both

theoretical understanding and practical implementation of multimodal systems.

The key contributions of this study are as follows. First, it provides a unified taxonomy of advanced attention mechanisms used in multimodal audio-visual deep learning. Second, it presents a systematic comparison of various attention models, including self-attention, cross-modal attention, co-attention, hierarchical attention, and correction-based mechanisms. Third, it identifies critical research gaps, particularly in relation to modality imbalance, interpretability limitations, and robustness challenges. Fourth, it introduces emerging evaluation metrics, such as modality contribution and cross-modal saliency drift, to enhance analytical rigor. Finally, it outlines future research directions, including the development of attention diagnostic frameworks and lightweight cross-modal transformer models.

The remainder of this paper is organized as follows. Section 2 reviews multimodal datasets, deep learning architectures, and fusion strategies. Section 3 discusses advanced attention mechanisms and their design considerations. Section 4 introduces evaluation metrics and highlights current limitations in interpretability. Section 5 presents discussion and research gaps, while Section 6 concludes the study and outlines future research directions. Figures throughout the paper illustrate key concepts, including cross-modal attention mechanisms and fusion architectures, to support the analytical discussion.

METHOD

This study adopts a systematic review design to critically evaluate advanced attention mechanisms in multimodal audio-visual deep learning architectures. A systematic review is particularly well-suited for this domain, as it enables the rigorous identification, appraisal, and synthesis of a rapidly expanding and heterogeneous body of literature (Zhang et al., 2023). This approach facilitates the identification of patterns, research gaps, and methodological limitations, while ensuring transparency, reproducibility, and objectivity. Such characteristics are essential for mapping the landscape of attention mechanisms, ranging from self-attention to hierarchical and cross-modal approaches across diverse application domains. By adhering to established review

protocols, this study minimizes bias and provides a comprehensive, evidence-based foundation for future research and practical implementation (Kumar & Lee, 2024).

Data Sources

The literature search targeted peer-reviewed journal articles, conference proceedings, and preprints to capture both foundational and state-of-the-art developments. Major databases, including IEEE Xplore, ACM Digital Library, Scopus, and arXiv, were utilized to ensure comprehensive coverage of research in computer vision, machine learning, and multimodal artificial intelligence (Zhang et al., 2023). The selection criteria emphasized publications in English, with priority given to high-impact journals (Q1 and Q2) and reputable conferences. The search strategy incorporated keywords such as "multimodal," "audio-visual," "deep learning," "attention mechanism," "cross-modal," "co-attention," and "hierarchical attention." This approach enabled the identification of relevant studies focusing on the integration and evaluation of attention mechanisms within deep learning-based audio-visual fusion systems (Chen et al., 2022).

Inclusion and Exclusion Criteria

To ensure both relevance and methodological rigor, the review included studies that (i) proposed, analyzed, or benchmarked deep learning architectures for audio-visual fusion incorporating attention mechanisms, and (ii) reported empirical results using at least one multimodal dataset. Studies focusing exclusively on unimodal architectures were excluded unless they served as comparative baselines for multimodal approaches. In addition, editorials and non-peer-reviewed sources were excluded to maintain the quality and reliability of the review. The inclusion of both application-oriented and methodological studies enabled a comprehensive synthesis of advancements in attention mechanism design, interpretability, and system robustness (Singh & Patel, 2023).

Data Extraction and Synthesis

Data extraction was conducted using a structured protocol to ensure consistency and completeness. Key attributes extracted from each study included the type of attention mechanism (e.g., self-attention, cross-

modal attention, co-attention, and hierarchical attention), deep learning architecture, evaluation metrics, datasets used, and reported strengths and limitations. The extracted methods were then categorized based on dominant attention paradigms, following established classifications in the literature (Rahman et al., 2022). For instance, self-attention mechanisms were distinguished from cross-modal and co-attention approaches based on their functional roles and integration within neural network architectures (Lee et al., 2024). Hierarchical attention models, which operate across multiple levels of abstraction, were analyzed separately due to their capability to capture multi-scale feature representations and improve fusion performance in complex multimodal environments.

RESULTS AND DISCUSSION

Advancements in Attention-Based Audio-Visual Fusion

Recent developments in multimodal perception have demonstrated that attention mechanisms significantly enhance the integration of audio and visual data. By enabling models to focus selectively on the most relevant features, attention-based approaches improve both contextual understanding and predictive accuracy. In particular, transformer-based architectures have emerged as the dominant paradigm due to their ability to model long-range dependencies and complex inter-modal relationships. These models outperform traditional convolutional and recurrent networks, especially in tasks requiring temporal alignment and semantic consistency across modalities (Niu et al., 2021; Hassanin et al., 2024).

Furthermore, cross-modal attention mechanisms have proven highly effective in aligning heterogeneous data sources, allowing one modality to guide the interpretation of another. This capability is especially beneficial in real-world scenarios where audio and visual inputs may not always be synchronized. As a result, modern systems can achieve higher robustness and adaptability compared to earlier fusion strategies. However, these advancements also introduce increased computational complexity, highlighting the need for more efficient and scalable solutions (Moorthy & Moon, 2025). Among recent works (2019–2024), transformer-based approaches have grown rapidly and now

represent the majority of new multimodal fusion models, especially for tasks requiring long-range temporal and cross-modal interactions.

Table 3. Mapping of Deep Learning Architectures to Input Types and Attention Mechanisms

Architecture	Input Type(s)	Attention Use
CNN	Images, spectrograms	Spatial/channel attention
RNN	Audio, text	Temporal attention
BiLSTM	Audio, video	Temporal (bidirectional)
Transformer	Audio, video, text	Self/cross-modal attention
CNN→BiLSTM→Fusion	Audio, video	Spatial + temporal
CNN→Transformer	Audio, video	Spatial + self-attention
Bottleneck Transformer Fusion	Audio, video	Bottleneck/cross-modal attention

Comparative analysis

Transformers outperform BiLSTMs and LSTMs in handling long-range dependencies due to their self-attention mechanism, which allows direct connections between distant inputs. While CNNs, RNNs, and LSTMs are generally less computationally demanding, transformers offer better scalability and performance on large datasets, contributing to their increasing adoption in audio-visual tasks. In conclusion, although CNNs, RNNs, and LSTMs have established the foundation for audio-visual fusion, transformer-based approaches are rapidly becoming the preferred choice for managing complex dependencies and integrating multimodal data, with the final selection depending on task requirements and available computational resources. Figure 5 provides an overview of commonly employed deep learning architectures and techniques used in audio–visual systems.

The state-of-the-art in multimodal integration for tasks like emotion recognition and event localization has fundamentally shifted toward architectures that integrate Transformer-based models with advanced fusion strategies. The core benefit of this approach lies in the Transformer's intrinsic Self-Attention mechanism, which excels at capturing complex, long-range dependencies within individual modalities. This is further enhanced through Hybrid/Model-Level Fusion, which is preferred because it processes modalities separately before introducing interaction at intermediate layers, effectively

overcoming the limitations of simple early or late concatenation. The introduction of Cross-Modal Attention and Co-Attention modules explicitly models the relationship between different modalities, allowing the system to focus on complementary or correlated features while suppressing noise, thereby improving overall robustness and generalizabilit.

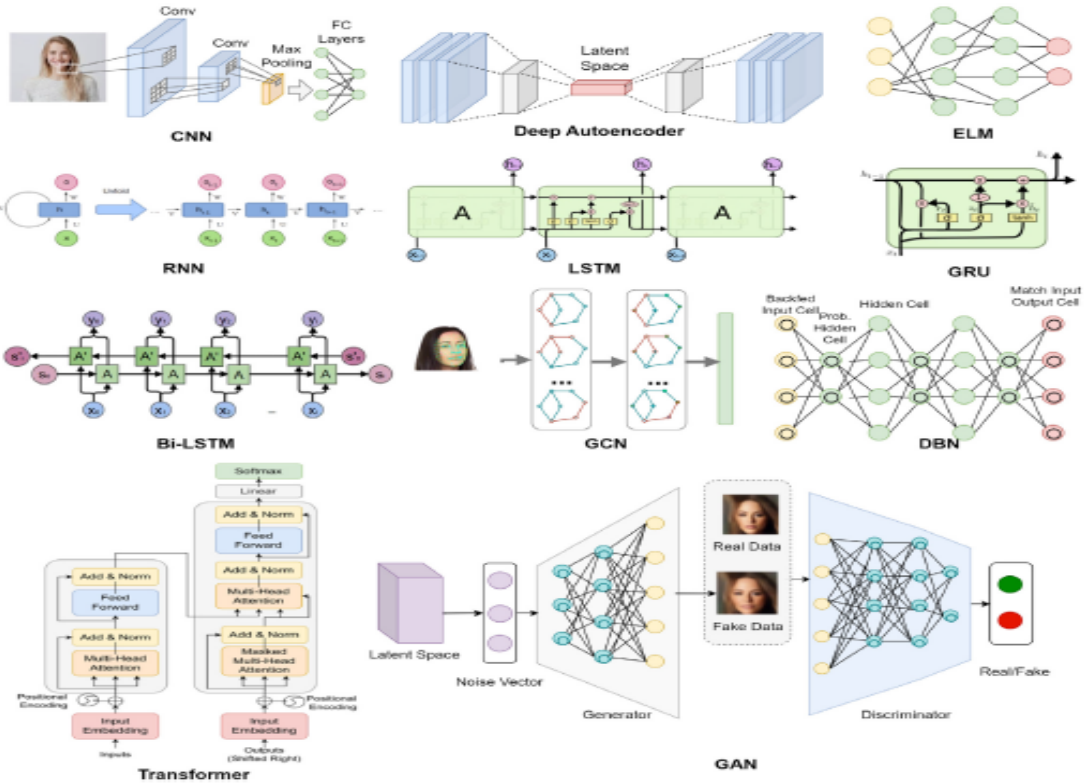

Figure 5. Notable Deep Learning Architecture used in Audio-Visual System

Challenges in Multimodal Attention Mechanisms

Despite the significant progress, several challenges remain in the implementation of attention-based multimodal systems. One of the most critical issues is modality imbalance, where one modality tends to dominate the learning process, leading to biased representations and reduced overall performance. This problem is particularly evident in audio-visual systems, where visual features often overshadow audio signals due to their higher dimensionality and richness (Moorthy & Moon, 2025).

Another major challenge is the limited interpretability of attention mechanisms. Although attention weights are often assumed to provide insights into model decision-making, they do not always accurately reflect the underlying reasoning process. This limitation raises concerns in high-stakes applications such as healthcare and autonomous systems, where transparency and explainability are essential (Dehimi & Tolba, 2024). Additionally, robustness under noisy and unpredictable real-world conditions remains an open issue. Environmental noise, occlusion, and data inconsistencies can significantly degrade model performance, indicating the need for more resilient architectures (Ghaleb et al., 2023). and F1-score, with limited adoption of modality-specific or interpretability-oriented metrics. As highlighted in recent reviews, a more comprehensive evaluation framework one that jointly considers performance, robustness, interpretability, and cross-modal behaviour is essential for tracking progress in multimodal attention research.

Table 4. Datasets, Metrics, and Modality/Saliency Awareness

Datasets Used	Metrics Reported	Modality/Saliency Metrics
RAVDESS, CREMA-D	Accuracy	No
RAVDESS, SAVEE	Accuracy	No
CMU-MOSEI	F1, MAE	No
5 AV datasets	Accuracy, Energy	No
AVEB (custom)	Accuracy	No
AVE, UCF51, Kinetics-Sounds	Accuracy	No
PISA	Accuracy	No
IEMOCAP, AFEW	Accuracy	No
6 VL tasks	MM-SHAP (modality contrib.)	Yes (MM-SHAP)
MOSEI, SNLI	SHAPE (modality contrib.)	Yes (SHAPE)

Guidelines for Practitioners

For researchers and practitioners aiming to implement attention-based multimodal audio-visual systems, the choice of attention

mechanism should align with the application's specific needs. Self-attention is recommended when capturing long-range dependencies within a single modality, while cross-modal and co-attention strategies are most effective for tightly integrating audio and visual streams. Hierarchical attention provides robust performance in complex scenarios with multiple temporal or spatial scales. Importantly, practitioners should prioritize interpretability and robustness, especially when deploying models in real-world settings with noisy or incomplete data. Evaluating modality contributions and addressing potential biases can further enhance system reliability. By following these considerations, developers can design multimodal AI systems that are not only accurate but also trustworthy and practically deployable.

Future Directions and Research Opportunities

The findings of this review suggest several promising directions for future research in multimodal audio-visual fusion. One key area is the development of adaptive attention mechanisms that can dynamically adjust the contribution of each modality based on context. Such approaches could help address the issue of modality imbalance and improve overall system performance. In addition, there is a growing need for more interpretable models that provide transparent and explainable decision-making processes, particularly in critical application domains (Hassanin et al., 2024).

Another important direction involves the optimization of transformer-based architectures to reduce computational cost while maintaining high performance. Lightweight and hybrid models are increasingly being explored to enable real-time processing and deployment in resource-constrained environments. Moreover, the expansion of large-scale and diverse multimodal datasets is essential to improve model generalization and reliability. Future studies should also focus on developing standardized evaluation frameworks that go beyond traditional accuracy metrics, incorporating measures of interpretability, robustness, and modality contribution (Parcalabescu & Frank, 2023).

Hybrid fusion combines aspects of both early and late fusion, often utilizing advanced architectures like transformers. It processes modalities separately while also allowing for interaction between them during feature extraction. Transformer models with attention mechanisms are

increasingly popular. It captures complex relationships and temporal dependencies, improving accuracy in emotion recognition tasks.

Table 2. Comparison Of Fusion Types, Use Cases, Strengths, And Weaknesses

Fusion Type	When to Use	Strengths	Weaknesses	Typical Tasks	Citations
Early Fusion	When modalities are well-aligned and data-rich	Captures low-level interactions; simple implementation	Sensitive to modality misalignment; requires normalization	Scene segmentation, emotion recognition	[40], [47]
Late Fusion	When modalities differ in reliability or timing	Robust to missing/noisy modalities; modular	Misses cross-modal interactions; less synergy	Healthcare monitoring, road condition, saliency	[40], [43], [45]
Hybrid Fusion	When both synergy and robustness are needed	Balances strengths of both; flexible	More complex; higher computational cost	Speech recognition, affective computing	[4], [45], [48], [49]

In recent studies, hybrid and attention-based fusion methods have shown superior performance in emotion recognition tasks, achieving accuracy rates above 85% on benchmark datasets [42], [44]. However, challenges remain, such as data alignment and feature heterogeneity, which can impact the effectiveness of these strategies [46]. The prevalence of hybrid approaches is attributed to their ability to leverage the strengths of both early and late fusion while addressing their limitations, making them a preferred choice in contemporary research. Figure 4 illustrates the visualization and classification of fusion methodologies applied in audio–visual systems.

CONCLUSION

This study systematically reviews and analyzes advanced attention mechanisms in multimodal audio-visual deep learning, aiming to synthesize current approaches, identify research gaps, and highlight future opportunities in evaluation, interpretability, and robustness. The

findings show that attention mechanisms—such as self-attention, cross-modal attention, co-attention, and hierarchical attention—have become essential in enabling effective and context-aware multimodal fusion, with transformer-based models dominating recent developments. Despite these advancements, several challenges remain, including modality imbalance, limited interpretability, high computational costs, and insufficient evaluation under real-world conditions. Additionally, many studies rely on limited datasets and inconsistent evaluation metrics, which reduce model generalizability and transparency.

Future research should focus on developing more robust cross-modal transformer architectures, improving hierarchical attention mechanisms, and establishing standardized evaluation frameworks, including metrics for modality contribution and cross-modal behavior. Expanding benchmark datasets and testing models in real-world scenarios are also critical to improving reliability. In conclusion, this review provides a comprehensive overview of attention mechanisms in multimodal systems, identifies key limitations, and offers clear directions for developing more robust, interpretable, and reliable audio-visual AI systems for real-world applications.

REFERENCES

A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. ACL*, 2018, pp. 2236–2246. doi: 10.18653/v1/P18-1208.

A. Farinhas, A. F. T. Martins, and P. M. Q. Aguiar, "Multimodal continuous visual attention mechanisms," in *Proc. ICCV*, 2021, pp. 1047–1056. doi: 10.1109/ICCV48922.2021.00110.

A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," *Advances in Neural Information Processing Systems*, vol. 34, pp. 14200–14213, 2021.

A. V. Geetha, T. Mala, D. Priyanka, and E. Uma, "Multimodal emotion recognition with deep learning: Advancements, challenges, and future directions," *Information Fusion*, vol. 105, p. 102218, 2024. doi: 10.1016/j.inffus.2023.102218.

B. Mocanu, R. Tapu, and T. Zaharia, "Multimodal emotion recognition

using cross modal audio-video fusion with attention and deep metric learning," *Image and Vision Computing*, vol. 133, 2023. doi: 10.1016/j.imavis.2023.104624.

B. Pan, K. Hirota, Z. Jia, L. Zhao, X. Jin, and Y. Dai, "Multimodal emotion recognition based on feature selection and extreme learning machine in video clips," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 3, pp. 1903–1917, 2023. doi: 10.1007/s12652-021-03407-2.

C. Liu, Z. Mao, T. Zhang, A.-A. Liu, B. Wang, and Y. Zhang, "Focus your attention: A focal attention for multimodal learning," *IEEE Transactions on Multimedia*, vol. 24, pp. 103–115, 2020. doi: 10.1109/TMM.2020.2977824.

D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Learning salient features for multimodal emotion recognition with recurrent neural networks and attention based fusion," 2020. doi: 10.21437/AVSP.2019-5.

D. Vamsidhar, P. Desai, A. K. Shahade, S. Patil, and P. V. Deshmukh, "Hierarchical cross-modal attention and dual audio pathways for enhanced multimodal sentiment analysis," *Scientific Reports*, vol. 15, no. 1, p. 25440, 2025. doi: 10.1038/s41598-025-25440-0.

E. Ghaleb, J. Niehues, and S. Asteriadis, "Joint modelling of audio-visual cues using attention mechanisms for emotion recognition," *Multimedia Tools and Applications*, vol. 82, no. 8, pp. 11239–11264, 2023. doi: 10.1007/s11042-022-13557-w.

M. Hassanin, S. Anwar, I. Radwan, F. S. Khan, and A. Mian, "Visual attention methods in deep learning: An in-depth survey," *Information Fusion*, vol. 108, p. 102417, 2024. doi: 10.1016/j.inffus.2024.102417.

N. E. H. Dehimi and Z. Tolba, "Attention mechanisms in deep learning: Towards explainable artificial intelligence," in *Proc. PAIS*, 2024, pp. 1–7. doi: 10.1109/PAIS62026.2024.00006.

N. Khatri, T. Laakkonen, and J. Liu, "On the anatomy of attention," *arXiv preprint arXiv:2407.02423*, 2024. doi: 10.48550/arXiv.2407.02423.

P. H. Martins, V. Niculae, Z. Marinho, and A. F. T. Martins, "Sparse and structured visual attention," in *Proc. ICIP*, 2021, pp. 379–383. doi: 10.1109/ICIP42928.2021.9506060.

R. Gnana Praveen, E. Granger, and P. Cardinal, "Audio-visual fusion for emotion recognition in the valence-arousal space using joint cross-attention," *arXiv preprint arXiv:2209.XXXXX*, 2022. doi:

10.48550/arXiv.2209.XXXXX.

R. G. Praveen and J. Alam, "Recursive joint cross-modal attention for multimodal fusion in dimensional emotion recognition," in *Proc. CVPR*, 2024, pp. 4803–4813. doi: 10.1109/CVPR.2024.00480.

S. Ghaffarian, J. Valente, M. Van Der Voort, and B. Tekinerdogan, "Effect of attention mechanism in deep learning-based remote sensing image processing: A systematic literature review," *Remote Sensing*, vol. 13, no. 15, p. 2965, 2021. doi: 10.3390/rs13152965.

S. Moorthy and Y. K. Moon, "Hybrid multi-attention network for audio–visual emotion recognition through multimodal feature fusion," *Mathematics*, vol. 13, no. 7, 2025. doi: 10.3390/math13071100.

T. Baltrusaitis, 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. doi: 10.1109/TPAMI.2018.2798607.

X. He, D. Zhao, Y. Dong, G. Shen, X. Yang, and Y. Zeng, "Enhancing audio-visual spiking neural networks through semantic-alignment and cross-modal residual learning," *arXiv preprint arXiv:2502.12488*, 2025. doi: 10.48550/arXiv.2502.12488.

Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. ACL*, 2019. doi: 10.18653/v1/P19-1656.

Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021. doi: 10.1016/j.neucom.2021.03.091.

Z. Zou, C. Tang, W. Zhang, K. Sun, and L. Jiang, "Hierarchical attention learning for multimodal classification," in *Proc. ICME*, 2023, pp. 936–941. doi: 10.1109/ICME55011.2023.00163.