Computer Engineering and Applications Vol. 15, No. 2, June 2026

A Comprehensive Survey of Audio-Visual Fusion with Attention Mechanisms: Trends, Challenges, and Future Directions

Rexcharles Enyinna Donatus[1*]

Aerospace Engineering Department , Air Force Institute of Technology, Kaduna, Nigeria
** charlly4eyims@yahoo.com*

ABSTRACT

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, Temporal Modeling, Cross-Modal Attention

1. 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[1], [2]. For instance, audio-visual emotion recognition systems are now pivotal in mental health diagnostics, autonomous driving, and affective computing, where understanding nuanced human emotions from both speech and facial expressions is essential for robust, context-aware responses [3], [4]. In real-world scenarios, such as video conferencing or surveillance, the fusion of audio and visual cues allows for more accurate detection of intent, sentiment, and action, especially under challenging conditions like low lighting or background noise [5]. In emotion and behavior recognition, for example, fusing prosody, facial expressions, and motion dynamics improves robustness to environmental distortions commonly encountered in real-world settings [6]. Deep learning has become the backbone of these advances, offering powerful architectures

such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers that can learn complex, hierarchical representations from multimodal data, facilitating the extraction and fusion of complementary features across modalities [5], [7].

Despite these advances, current multimodal fusion methods face persistent challenges. Traditional fusion strategies early, late, and hybrid often struggle to capture intricate cross-modal dependencies, leading to suboptimal performance in dynamic, noisy, or incomplete data environments [3]. These limitations highlight the need for refined attention mechanisms capable of selective and context-aware integration [8]. Existing attention mechanisms, while effective in highlighting salient features, frequently fail to address modality imbalance, where one modality dominates or suppresses the other, resulting in reduced robustness and generalizability [4]. Dataset inconsistencies, such as varying annotation standards and limited diversity, further hinder the development of universally applicable models. Moreover, evaluation metrics often focus narrowly on accuracy or F1-score, neglecting deeper insights into modality contributions and attention interpretability. These limitations manifest in practical deployments as decreased reliability, poor adaptability to real-world noise, and limited transparency critical barriers for applications in healthcare, autonomous systems, and interactive AI.

Prior research has attempted to address these challenges through a variety of architectural and algorithmic innovations. Cross-modal attention mechanisms, such as those leveraging transformers or dual-attention modules, have improved the modeling of inter-modal relationships, enabling more nuanced feature integration [3]. Hierarchical attention frameworks and dual-pathway designs have been proposed to capture both fine-grained and global dependencies, enhancing robustness to missing or noisy modalities [9]. However, these approaches often fall short in fully resolving modality imbalance, as they may still prioritize dominant modalities or fail to adapt dynamically to context shifts [4]. Interpretability remains a significant concern; while attention maps offer some transparency, the underlying decision processes are frequently opaque, limiting trust and diagnosability. Furthermore, cross-modal transformers and co-attention networks, though promising, are not yet systematically evaluated for their ability to handle attention failure cases or provide actionable interpretability in complex, real-world settings [5], [7].

A critical gap persists in the comprehensive mapping and evaluation of advanced attention mechanisms within multimodal audio-visual deep learning. There is a lack of systematic reviews that dissect the nuances of modality-specific attention, cross-modal transformers, hierarchical attention, and the circumstances under which attention mechanisms fail or succeed. The field lacks standardized metrics for quantifying attention interpretability and modality contribution, impeding the objective comparison of competing architectures. Addressing this gap is crucial for advancing both theoretical understanding and practical deployment, as it would enable the design of more robust, interpretable, and adaptable multimodal systems.

This paper aims to: (i) critically review advanced attention mechanisms in multimodal audio-visual deep learning, moving beyond simple fusion strategies; (ii) categorize and analyze cross-modal attention architectures, including hierarchical and dual-attention frameworks; and (iii) propose novel metrics for evaluating attention

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

interpretability and modality contribution, facilitating more transparent and reliable model assessment.

The key contributions of this survey are as follows:

i. We provide a unified taxonomy of advanced attention mechanisms used in multimodal audio–visual deep learning.
ii. We systematically compare self-attention, cross-modal, co-attention, hierarchical, and correction-based attention modules.
iii. We identify critical gaps in multimodal fusion, including modality imbalance, interpretability limitations, and robustness challenges.
iv. We introduce emerging evaluation metrics such as modality contribution and cross-modal saliency drift.
v. We propose future research directions, including attention diagnostics frameworks and lightweight cross-modal transformers.

The remainder of this paper is structured as follows. Section 2 reviews the landscape of multimodal datasets, deep learning architectures, and fusion strategies. Section 3 delves into advanced attention mechanisms, categorizing and analyzing their design and performance. Section 4 introduces evaluation metrics, highlighting current limitations and proposing new approaches for interpretability and modality assessment. Section 5 discusses underexplored areas and synthesizes key findings, while Section 6 outlines future research directions and concludes the study. Figures throughout the paper illustrate key concepts, such as cross-modal attention mechanisms and fusion strategies, to support the analytical narrative**.**

2. METHODOLOGY

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 [10]. The systematic approach ensures identification of patterns, gaps, and methodological limitations ,models transparency, reproducibility, and objectivity, which are essential for mapping the landscape of attention mechanisms ranging from self-attention to hierarchical and cross-modal variants across diverse application contexts. By adhering to established review protocols, the study minimizes bias and provides a comprehensive, evidence-based foundation for future research and practical deployment [11].

2.1 DATA SOURCES

The literature search targeted peer-reviewed journal articles, conference proceedings, and preprints to capture both foundational and cutting-edge developments. Databases included IEEE Xplore, ACM Digital Library, Scopus, and arXiv, ensuring broad coverage of computer vision, machine learning, and multimodal AI research [10]. Selection criteria prioritized works published in English, with a focus on Q1 and Q2 journals and top-tier conferences. The search strategy combined keywords such as “multimodal,” “audio-visual,” “deep learning,” “attention

mechanism," "cross-modal," "co-attention," and "hierarchical attention." This approach facilitated the identification of studies that specifically address the integration and evaluation of attention mechanisms within deep learning-based audio-visual fusion systems.

2.2 INCLUSION AND EXCLUSION CRITERIA

To ensure relevance and 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 on at least one multimodal dataset. Studies focusing solely on unimodal architectures were excluded unless they served as comparative baselines for multimodal approaches. Editorials, and non-peer-reviewed sources were omitted to maintain methodological quality. The inclusion of both application-driven and methodological papers allowed for a holistic synthesis of advances in attention design, interpretability, and robustness.

2.3 DATA EXTRACTION AND SYNTHESIS

Data extraction was guided by a structured protocol, capturing key attributes of each study: attention mechanism type (self, cross, co-attention, hierarchical), deep learning architecture, evaluation metrics, datasets used, and reported strengths and limitations. Methods were categorized according to the dominant attention paradigm, following established taxonomies [12]. For example, self-attention mechanisms were distinguished from cross-modal and co-attention approaches based on their operational focus and integration within the network [13]. Hierarchical attention models, which operate at multiple abstraction levels, were separately identified due to their unique capacity for multi-scale feature integration.

3. BACKGROUND AND FUNDAMENTAL CONCEPTS

3.1 OVERVIEW OF MULTIMODAL FUSION

Multimodal fusion is foundational to audio-visual deep learning, enabling systems to integrate complementary information from disparate sensory streams[14]. The three principal strategies early fusion, late fusion, and hybrid (or model-level) fusion each offer distinct trade-offs. Early fusion concatenates raw or low-level features from audio and visual modalities before feeding them into a unified model. While this approach allows direct interaction between modalities, it often fails to capture complex inter-modal relationships and can suffer from data sparsity, resulting in only marginal performance gains [3]. Late fusion, in contrast, processes each modality independently and combines their predictions at the decision level [2]. This method is straightforward and robust to missing modalities but neglects mutual dependencies, limiting its ability to exploit cross-modal correlations [3], [9]. Hybrid or model-level fusion leverages deep learning architectures to extract and integrate features at intermediate layers, capturing both intra- and inter-modal relationships. This approach, especially when combined with attention mechanisms, has demonstrated superior performance in emotion recognition, action recognition, and other tasks [4], [9]. Understanding these multimodal fusion strategies is crucial for optimizing performance in audio-visual emotion recognition tasks.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

3.2 ATTENTION MECHANISMS IN MULTIMODAL LEARNING

Attention mechanisms have become central to multimodal deep learning, enabling models to dynamically weigh the importance of features within and across modalities [15], [16]. Self-attention (intra-modal) mechanisms capture dependencies within a single modality, enhancing contextual understanding [4], [7], [14]. Cross-attention (inter-modal) mechanisms allow the model to focus on relevant information in one modality based on cues from another, facilitating richer feature integration [3], [4], [17], [18]. More sophisticated frameworks such as co-attention and dual-attention introduce bidirectional information flow, improving the modeling of complex, context-dependent relationships between audio and visual streams [4], [9], [17]. Hierarchical attention structures extend these capabilities by aggregating multi-level information, thereby capturing both local and global dependencies that are particularly valuable in audio–visual sentiment and emotion analysis [9]. The following sections define and classify various attention mechanisms, offering operational descriptions and empirical evidence of their effectiveness. For a detailed explanation of the underlying mathematical intuition, readers are referred to other sources [19].

3.2.1 SELF-ATTENTION

Self-attention allows a model to weigh the importance of different parts of a single modality's input data by computing attention scores for each element in the sequence with respect to all other elements. The attention score is calculated using a scaled dot-product of query, key, and value vectors derived from the input data. Empirical evidence shows that self-attention effectively captures long-range dependencies within a modality, as demonstrated in tasks such as language modeling and image classification [20].

3.2.2 CROSS-ATTENTION

Cross-attention focuses on aligning elements from different modalities, for example aligning text with corresponding image regions [21], [22]. It computes attention scores between query vectors from one modality and key-value pairs from another modality. Cross-attention has shown strong success in tasks requiring semantic alignment, such as image-text matching, by effectively associating relevant sub-elements across modalities [23]. Although widely applied in text–image tasks,, co-attention architectures have also been effective in audio–visual emotion recognition and event localization. Figure 1 illustrates an A–V fusion mechanism designed to encode inter-modal correlations and retain essential intra-modal features.

3.2.3 CO-ATTENTION

Co-attention simultaneously attends to multiple modalities, enabling joint representation learning through a bidirectional attention mechanism in which each

modality attends to the other [21], [24]. This facilitates mutual information exchange. Empirical studies show that co-attention mechanisms enhance performance in visual question answering by better capturing interactions between text and image data [20], [25].

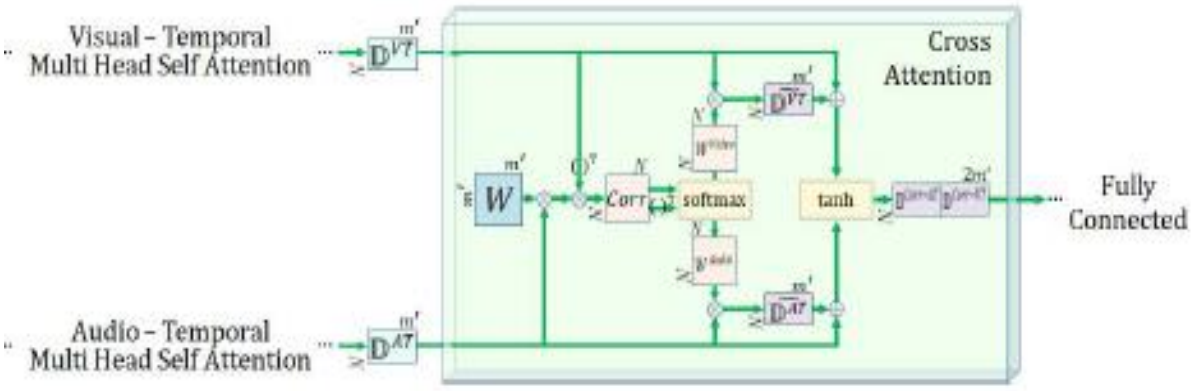

FIGURE 1. Schematic of the Cross-modal Attention Fusion Module [3].

3.2.4 HIERARCHICAL ATTENTION

Hierarchical attention structures the attention process in multiple layers to capture both local and global dependencies. It aggregates multi-level features through progressive fusion, improving the model's ability to learn complex patterns. Evidence shows that hierarchical attention is effective in multimodal classification tasks, outperforming traditional methods by leveraging higher-order correlations [26].

3.2.5 RESIDUAL/CORRECTION ATTENTION

Residual or correction attention refines initial attention scores by incorporating residual connections that correct potential errors in the attention process. This is particularly beneficial in audio–visual fusion where one modality may dominate under noisy condition. This approach adds a residual term to the attention scores, which stabilizes the learning procedure. Empirical findings indicate that residual attention improves model robustness and performance across a variety of multimodal tasks [25].

While attention mechanisms significantly enhance model performance, their interpretability remains a challenge. Attention maps, although widely used to visualize model focus, do not always provide clear insights into the decision-making process. The decision-making process is often opaque, especially in deep transformer-based models [12], [13], [27]. Failure cases frequently arise when modalities are imbalanced or when noise and missing data disrupt the attention distribution, underscoring the need for robust and adaptive mechanisms [4], [7], [9]. Notably, recent advances such as dual cross-modality attention and hybrid attention modules have improved robustness and performance, even under adverse conditions [9], [17].

The limitations of attention mechanisms are particularly evident in complex multimodal settings, where interactions between modalities are intricate and not fully captured by existing explainability methods [28]. Consequently, ongoing research is

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

needed to develop more interpretable and transparent attention-based models. Figure 2 presents the overall framework. Audio–visual pairs are first encoded into high-dimensional embeddings, which are refined in the fusion module. The SMA block captures intramodal dependencies through self-attention, while the PCMA block models cross-modal interactions via parallel attention. Finally, both multimodal and unimodal representations feed into the emotion recognition module to produce the outputs.

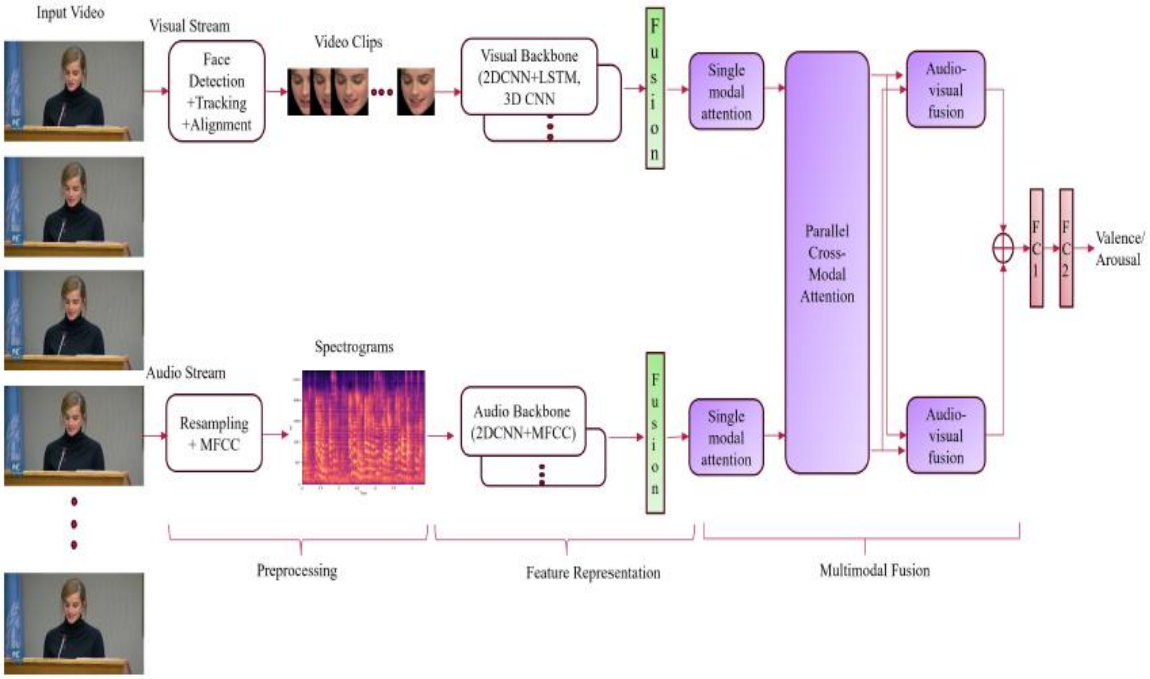

FIGURE 2. Cross-Modal Attention Workflow [4].

3.3 STRENGTHS AND WEAKNESSES ACROSS ATTENTION MECHANISMS

Self-attention mechanisms excel at capturing intra-modal dependencies and are highly effective for extracting salient features within each modality [3], [13], [29]. Cross-attention and co-attention mechanisms, by contrast, explicitly model the interactions between modalities, allowing the network to focus on complementary or correlated features [5], [27], [30]. Hierarchical attention structures further enhance this by operating at multiple abstraction levels, improving both interpretability and robustness [4], [7]. Despite these strengths, attention-based models can be sensitive to modality imbalance and may propagate noise if one modality is corrupted or missing [5], [7]. Some studies have addressed this by introducing correction or residual attention modules that adaptively reweight modalities based on their reliability [4], [5], [30].

Given the diversity of attention mechanisms and their variants, it is essential to organize them within a structured taxonomy that clarifies their functional roles in multimodal audio-visual models. In this survey, we present a classification framework that groups attention mechanisms into major categories based on their functional role within audio-visual models. Figure 3 illustrates the taxonomy, showing how the

different mechanisms and their variants can be organized into coherent categories and subcategories. The primary branches distinguish mechanisms designed to operate on specific feature representations, mechanisms that target particular query structures, and mechanisms that serve more general purposes without direct dependence on feature or query types.

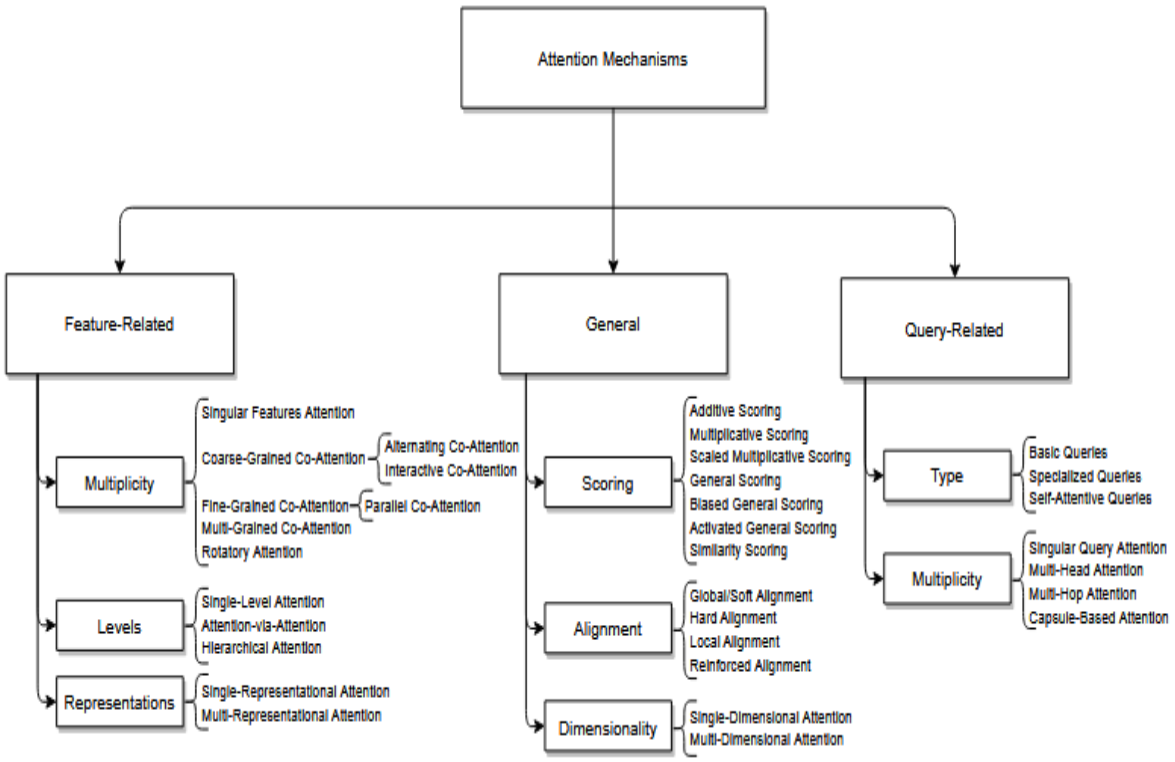

FIGURE 3. Taxonomy of Attention Mechanisms [16][12].

Building on the taxonomy introduced above, the following section examines how these attention mechanisms operate within multimodal deep learning systems, emphasizing their strengths, limitations, and typical failure modes. Attention mechanisms are central to modern multimodal fusion, enabling models to focus on the most relevant information across and within modalities. Table 1 presents a concise taxonomy and comparative overview of the core attention types.

TABLE 1.
Comparison of Attention Types, Mechanisms, and Trade-Offs in Multimodal Fusion.

Type	Mechanism Description	Pros	Cons	Refs
Self-attention	Attends to relationships within a single modality	Captures long-range dependencies; flexible	Ignores cross-modal cues; may miss complementarity	[31], [32], [33]
Cross-attention	Attends to one modality using features from another	Models inter-modal	Computationally intensive; sensitive to misalignment	[12], [27], [33]

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

Type	Mechanism Description	Pros	Cons	Refs
		interactions; flexible fusion		
Co-attention	Jointly learns attention over two (or more) modalities simultaneously	Captures bidirectional dependencies	Complex to train; risk of overfitting	[12], [27]
Hierarchical	Stacks multiple attention layers (e.g., word→sentence→document or frame→clip→video)	Multi-level context; interpretable	High complexity; more parameters	[27], [34]
Residual/Correction	Adds correction or residual connections to attention outputs	Stabilizes training; improves gradient flow	May mask attention errors; harder to interpret	[32]

3.3.1 WHEN ATTENTION FAILS: COMMON FAILURE MODES

To better understand the practical behavior of attention mechanisms, it is important to examine both their strengths and the conditions under which they tend to fail.

- **Modality imbalance:** Over-reliance on a dominant modality can suppress weaker but informative signals, reducing fusion effectiveness [27], [33].
- **Noise Sensitivity:** Attention may amplify noisy or irrelevant features, especially if modalities are misaligned or corrupted[35], [36] .
- **Bias:** Learned attention weights can reflect dataset or modality biases, leading to unfair or suboptimal decisions [37], [38].
- **Adversarial Attacks:** Attention mechanisms can be manipulated by adversarial inputs, causing models to focus on misleading cues and degrade robustness [37], [38], [39].

3.3.2 SPECIALIZED ATTENTION FOR ROBUSTNESS AND ALIGNMENT

Current research emphasizes the development of specialized attention modules to tackle specific challenges inherent to multimodal data. For instance, Hierarchical Attention is employed to capture dependencies across multiple granularities, simultaneously modeling local temporal features and global contextual correlations [26]. To directly confront issues of noise and modality imbalance where one data stream may dominate the prediction models are incorporating correction-based attention mechanisms, such as the Adaptive Interaction and Correction Attention Network (AICANet). These mechanisms dynamically modulate features to ensure they are aligned and appropriately weighted, thereby actively mitigating the influence of low-quality or misaligned inputs [30]. This move represents a critical shift from passive fusion to active, corrective feature alignment.

3.4 FUSION STRATEGIES IN AUDIO–VISUAL DEEP LEARNING

Fusion strategies define how information from audio and visual modalities is integrated within a multimodal architecture [19]. The evolution of fusion has progressed from simple concatenation-based methods to advanced attention-driven

mechanisms capable of modeling fine-grained cross-modal relationships. This section provides a structured taxonomy of fusion paradigms and highlights their strengths, limitations, and applications in modern audio–visual system.

3.4.1 EARLY FUSION

Early fusion combines raw data from multiple modalities at the input level before processing. It merges features from different modalities into a single representation, and Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are commonly used [14], [40]. It captures inter-modal relationships effectively and reduces dimensionality, but it is sensitive to noise from any single modality, which can degrade overall performance [41].

3.4.2 LATE FUSION

Late fusion integrates the outputs of separate models trained on individual modalities. Each modality is processed independently, and their predictions are combined at the decision level. Ensemble methods and voting classifiers are prevalent. It is robust to noise in individual modalities and allows for easy integration of new modalities, but it may overlook inter-modal dependencies, leading to suboptimal performance [42], [43].

3.4.3 HYBRID/MODEL-LEVEL FUSION

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 [42], [44], [45] . However, its complexity can lead to overfitting, especially with limited data [46]. The comparative performance of multimodal fusion strategies is presented in Table 2.

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]

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

Fusion Type	When to Use	Strengths	Weaknesses	Typical Tasks	Citations
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.

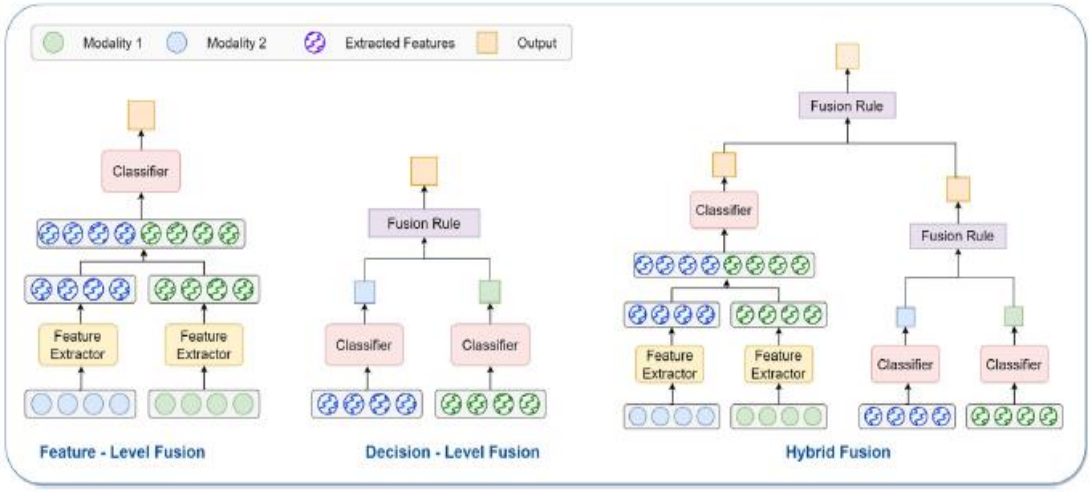

FIGURE 4. Representation of Fundamental Fusion Methods [19].

3.5 DEEP LEARNING ARCHITECTURES

The evolution of deep learning architectures has profoundly shaped multimodal audio-visual research. CNN-based models excel at extracting spatial features from visual data and, when adapted, can process spectrogram representations of audio [3], [4], [50]. Recent studies have demonstrated that CNNs can achieve high accuracy in emotion recognition by effectively capturing spatial features from video frames [51], [52]. RNNs and LSTMs are adept at modeling temporal dependencies, crucial for sequential data such as speech and video [4], [7], [22]. However, RNNs and LSTMs struggle with long-range dependencies, which can cause performance degradation in tasks requiring integration of information over extended sequences [51]. Although they are generally less computationally intensive than CNNs, their performance may be hindered by vanishing gradient issues in long sequences. Bidirectional LSTMs (BiLSTMs) extend standard LSTMs by processing data in both forward and backward directions. They are particularly effective in emotion recognition tasks where context from both past and future frames enhances accuracy. BiLSTMs handle long-range

dependencies better than standard LSTMs, making them suitable for complex temporal tasks [51], [53], [54]. More recently, transformer-based architectures have revolutionized the field by enabling parallelized processing and capturing long-range dependencies through self-attention mechanisms [4], [15], [42], [55], [56]. Despite their computational intensity, their parallel processing capabilities significantly reduce training time and make them suitable for large datasets and complex tasks. For example, transformer encoders applied to both audio and visual embeddings have improved temporal modeling and facilitated more effective late joint fusion, reducing overfitting and enhancing generalization [15], [42]. Hybrid models that combine CNNs for feature extraction and transformers for temporal fusion are increasingly prevalent, offering state-of-the-art results in continuous emotion recognition and video classification [4], [42], [55].

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 [8], [57], [58], [59].

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

Architecture	Input Type(s)	Attention Use	Example References
CNN	Images, spectrograms	Spatial/channel attention	[59], [60]
RNN	Audio, text	Temporal attention	[61]
BiLSTM	Audio, video	Temporal (bidirectional)	[61], [62]
Transformer	Audio, video, text	Self/cross-modal attention	[8], [57], [58], [59], [63]
CNN→BiLSTM→Fusion	Audio, video	Spatial + temporal	[62], [64], [65]
CNN→Transformer	Audio, video	Spatial + self-attention	[42], [58], [59], [63]
Bottleneck Transformer Fusion	Audio, video	Bottleneck/cross-modal attention	[8]

3.5.1 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

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

approach lies in the Transformer's intrinsic Self-Attention mechanism, which excels at capturing complex, long-range dependencies within individual modalities [15], [42]. 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 [4]. 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 generalizability [3].

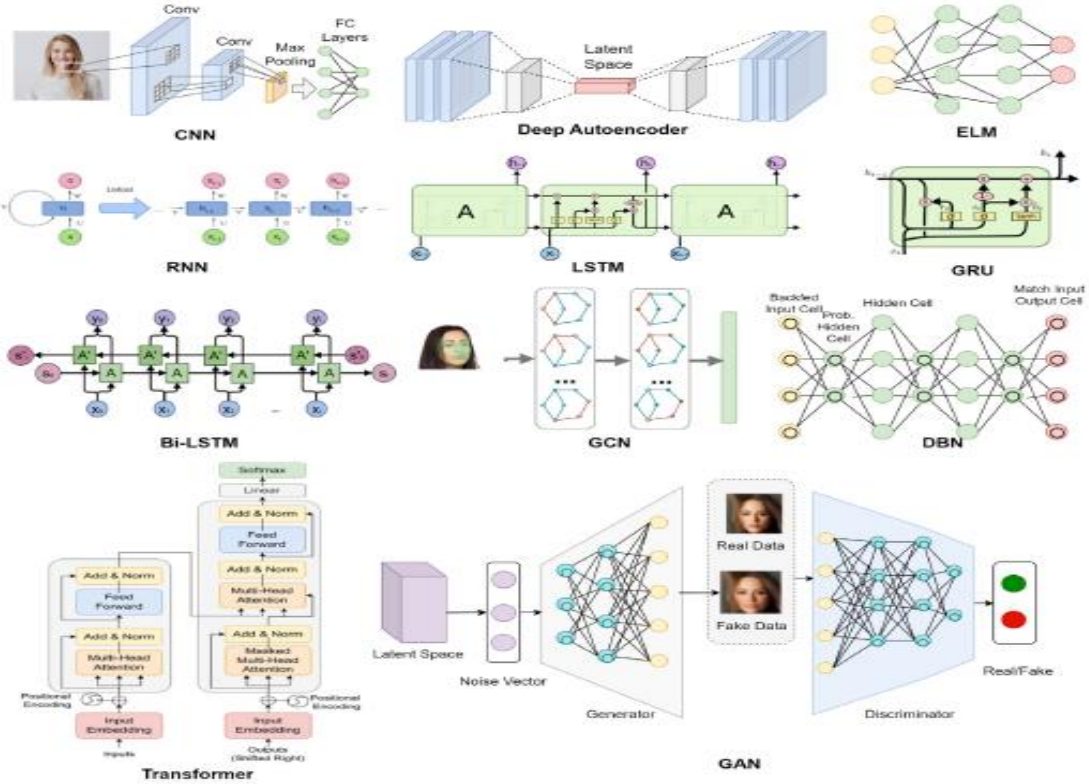

FIGURE 5. Notable Deep Learning Architecture used in Audio-Visual System [19].

3.6 ANALYSIS APPROACH

To contextualize the taxonomy and its implications, this section examines how different attention mechanisms have been applied across multimodal audio-visual architectures. A qualitative synthesis was conducted to compare and contrast the identified attention mechanisms, with particular emphasis on their effectiveness in addressing modality imbalance, interpretability, and cross-modal integration. Our comparative analysis highlights trends in architectural design, such as the increasing adoption of transformer-based models and the integration of dual- and co-attention modules for enhanced feature fusion [12], [13]. The review also identified research gaps, including the limited use of standardized interpretability metrics and the scarcity of studies stress-testing attention mechanisms under real-world conditions [40].

By systematically mapping the landscape of attention mechanisms in multimodal audio-visual deep learning, this review provides a critical foundation for advancing both theoretical understanding and practical deployment. The structured methodology ensures that the synthesis is comprehensive, transparent, and aligned with the evolving needs of the research community.

3.7 SYNTHESIS OF KEY FINDINGS

Recent progress in multimodal audio-visual deep learning has been driven by increasingly sophisticated attention architectures that enable models to align, weigh, and integrate features across modalities more effectively. The field has moved from simple concatenation and early/late fusion strategies to models that explicitly learn to align, weigh, and integrate information across modalities using attention mechanisms.

Cross-modal, co-attention, and hierarchical attention frameworks now dominate state-of-the-art systems, enabling more nuanced and context-aware fusion of audio and visual cues [3], [4], [5], [7], [30]. For example, adaptive attention modules such as those in AICANet and S-CMRL dynamically correct and align modality-specific features, improving robustness to noise and enhancing the consistency of learned representations [5], [30]. Hybrid models that combine intra-modal (self-attention) and inter-modal (cross/co-attention) mechanisms have demonstrated superior performance in emotion recognition, event localization, and deepfake detection, often outperforming unimodal or naïve fusion baselines by significant margins [3], [66], [67], [68].

A key trend is the integration of transformer-based architectures, which leverage multi-head self- and cross-attention to capture long-range dependencies and complex inter-modal relationships [3], [4], [29], [69]. These models are particularly effective at modeling temporal dynamics and have shown strong generalization across diverse datasets. However, the computational cost and data requirements of transformers remain a challenge, especially for real-time or resource-constrained applications [12].

4. CRITICAL SYNTHESIS & RESEARCH GAPS

The literature on multimodal audio-visual deep learning reveals a clear trend: hybrid fusion strategies and transformer-based architectures, augmented with advanced attention mechanisms, consistently outperform traditional approaches by capturing nuanced temporal and semantic relationships [3], [9], [42]. Cross-modal and hierarchical attention mechanisms have been shown to be particularly effective in modeling complex dependencies and enhancing robustness to noise and missing data[4], [9], [17] . This progress has established attention-based multimodal fusion as superior to early or late fusion methods, especially in tasks requiring fine-grained temporal and semantic alignment [3], [4], [7], [12].

Despite these advances, several gaps and limitations persist. Interpretability remains a central challenge**,** as attention weights often provide only partial insight into model reasoning, and outputs may not correspond to human-understandable decision-making. Efforts such as causal or correction-based attention modules and attention visualization have begun addressing these issues [4], [30], [70], yet the field lacks

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

systematic evaluation of attention interpretability under adversarial, noisy, or real-world conditions [12], [13], [70].

Robustness to modality imbalance and data limitations is another key concern. While attention-based models generally improve resilience to missing or noisy modalities, performance can degrade if a model over-relies on a single modality or is trained on biased or limited datasets [5], [7]. Adaptive attention, residual correction, and hybrid attention modules show promise in mitigating these effects, but consistent real-world performance is not yet guaranteed [4], [5], [30].

Evaluation practices remain inconsistent**,** with few studies employing standardized metrics such as modality contribution analysis or saliency drift assessment [3], [9]. This lack of uniform evaluation makes it difficult to compare architectures fairly, especially regarding generalizability and robustness across diverse datasets.

Contradictions in reported robustness further highlight the need for systematic stress-testing. While many studies claim improved resilience of attention models, others reveal vulnerabilities, particularly when models encounter noisy, missing, or culturally varied input [5], [7]. Moreover, the computational demands of transformer-based fusion architectures may limit deployment in resource-constrained environments, raising concerns about scalability and practical applicability [4], [66], [71], [72], [73].

Implications for real-world applications are significant. Advanced attention mechanisms enhance performance in emotion recognition, event localization, and deepfake detection, offering tangible benefits for human-computer interaction, surveillance, and content moderation. However, achieving fairness, transparency, and reliability remains a pressing challenge, particularly as these systems are deployed in sensitive domains [4], [66], [71], [72], [73].

In summary, the main research gaps include:

1. Limited interpretability of attention mechanisms in deep multimodal architectures.
2. Inconsistent robustness under modality imbalance, noisy data, or biased datasets.
3. Lack of standardized evaluation protocols and metrics for cross-study comparisons.
4. Insufficient stress-testing and validation in real-world, diverse, or adverse scenarios.
5. High computational demands limiting scalability and practical deployment.

Addressing these gaps will be essential for the development of more transparent, robust, and generalizable multimodal audio-visual AI systems, bridging the gap between laboratory research and real-world applications.

5. EVALUATION METRICS & BENCHMARKS

Rigorous evaluation of multimodal attention mechanisms requires metrics that capture not only predictive accuracy but also how models distribute, balance, and stabilize attention across modalities.Evaluating multimodal attention mechanisms requires metrics that extend beyond traditional accuracy-based measures. While accuracy and F1-score remain the most commonly reported metrics particularly across standard benchmark datasets such as RAVDESS**,** CREMA-D**,** and AVE [11], [19], they provide only a partial view of model behavior in multimodal settings.

Recent studies have introduced modality-aware metrics designed to quantify how attention mechanisms distribute importance across audio and visual streams. Modality contribution scores measure the relative influence of each modality within fusion layers and have shown value in diagnosing modality imbalance in attention-based systems [3] Complementing this, saliency drift metrics capture how attention distributions shift across time or under perturbations, offering insight into the stability of cross-modal representations [9].

Another emerging class of evaluation tools focuses on cross-modal consistency**,** assessing whether a model preserves coherent predictions when one modality is noisy, occluded, or entirely missing. Such tests are particularly useful for understanding robustness in real-world applications where multimodal inputs are often imperfect [7].

Despite these advances, most benchmark evaluations still rely heavily on accuracy 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 [11], [12], [19].

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

Paper (Year) & Ref	Datasets Used	Metrics Reported	Modality/Saliency Metrics
[3]	RAVDESS, CREMA-D	Accuracy	No
[74]	RAVDESS, SAVEE	Accuracy	No
[75]	CMU-MOSEI	F1, MAE	No
[76]	5 AV datasets	Accuracy, Energy	No
[77]	AVEB (custom)	Accuracy	No
[78], [79]	AVE, UCF51, Kinetics-Sounds	Accuracy	No
[78], [80]	PISA	Accuracy	No
[80], [81]	IEMOCAP, AFEW	Accuracy	No
[82]	6 VL tasks	MM-SHAP (modality contrib.)	Yes (MM-SHAP)
[83]	MOSEI, SNLI	SHAPE (modality contrib.)	Yes (SHAPE)

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

5.1 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. [4], [40], [43], [45], [47].

6. RECOMMENDATIONS FOR FUTURE RESEARCH

Future research should prioritize the development of a comprehensive attention diagnostics framework that systematically evaluates modality contribution, cross-modal saliency drift, and interpretability. Metrics such as modality contribution score and cross-modal saliency drift should be incorporated to more accurately characterize the behavior of attention mechanisms under diverse conditions [3], [5], [16], [30]. Stress-testing multimodal architectures using adversarial noise, missing-modality scenarios, and realistic data variability remains essential for assessing robustness and generalizability [5], [7].

Further investigation into lightweight attention architectures, as well as causal and correction-based attention modules, may help bridge the persistent tension between performance and interpretability [4], [30], [70]. Additionally, expanding the scale, diversity, and ecological validity of benchmark datasets will be crucial to ensuring that advances in attention mechanisms translate effectively to real-world deployments.

In summary, while attention mechanisms continue to strengthen multimodal audio-visual deep learning, meaningful progress requires deeper focus on interpretability, robustness, and standardized evaluation. Addressing these gaps will be central to ensuring responsible and reliable use of audio-visual AI systems.

7. CONCLUSION

This study set out to systematically review and critically analyze advanced attention mechanisms in multimodal audio-visual deep learning architectures. The main objectives were to synthesize the current landscape of attention-based fusion methods, identify persistent gaps, and outline emerging opportunities for improving evaluation, interpretability, and robustness. By consolidating recent literature, the review demonstrates how attention mechanisms have reshaped multimodal integration while revealing areas that remain underexplored.

The findings show that self-attention, cross-modal attention, co-attention, and hierarchical attention have become central to modern multimodal systems, enabling more effective and context-aware fusion of audio and visual streams [12], [13], [16]. Transformer-based models and hybrid attention architectures now dominate state-of-the-art approaches, delivering improved performance in multimodal tasks such as emotion recognition, sentiment analysis, and audio-visual matching [3], [84], [85]. Their ability to model complex inter-modal relationships and temporal dependencies has led to notable gains in accuracy and robustness [3], [7], [9].

Despite these advances, key challenges persist. Current systems continue to grapple with modality imbalance, interpretability limitations, and the computational cost of sophisticated attention modules [12], [86], [87]. Moreover, attention behavior under noisy or missing modality conditions and its alignment with human interpretability remain insufficiently understood [4], [9], [88]. The review also highlights limitations arising from narrow benchmark datasets and inconsistent evaluation practices, which often overlook interpretability and cross-modal behavior [7], [16].

Future research should prioritize the development of cross-modal Transformer architectures and hierarchical attention models capable of capturing multi-scale dependencies and improving robustness in real-world settings [7], [9], [12] . Establishing an attention diagnostics framework including metrics such as modality contribution score and cross-modal saliency drift will be essential for systematic comparison and transparent evaluation of attention modules [7], [16]. Expanding benchmark datasets and stress-testing models under real-world conditions will further improve the generalizability and reliability of multimodal AI systems [7], [9].

In conclusion, this review advances the understanding of attention mechanisms in multimodal audio-visual deep learning by synthesizing existing approaches, identifying fundamental gaps, and proposing actionable future directions. By emphasizing interpretability, robustness, and comprehensive evaluation, the study provides a clear roadmap for developing next-generation multimodal AI systems that are both effective and trustworthy, demonstrating the critical role of attention mechanisms in enabling robust, transparent, and practical real-world applications across domains such as healthcare, autonomous systems, and human–computer interaction.

ACKNOWLEDGEMENTS

This research was conducted independently without external funding or support. The author gratefully acknowledges the contributions of the referenced literature that informed this study.

REFERENCES

[1] T. Baltrusaitis, C. Ahuja, and L. P. Morency, “Multimodal Machine Learning: A Survey and Taxonomy,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2019, doi: 10.1109/TPAMI.2018.2798607.

[2] B. Pan, K. Hirota, Z. Jia, L. Zhao, X. Jin, and Y. Dai, “Multimodal emotion

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

recognition based on feature selection and extreme learning machine in video clips," *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 3, pp. 1903–1917, 2023, doi: 10.1007/s12652-021-03407-2.

[3] B. Mocanu, R. Tapu, and T. Zaharia, "Multimodal Emotion Recognition using Cross Modal Audio-Video Fusion with Attention and Deep Metric Learning," *Image Vis. Comput.*, vol. 133, pp. 1–18, 2023.

[4] S. Moorthy and Y. K. Moon, "Hybrid Multi-Attention Network for Audio–Visual Emotion Recognition Through Multimodal Feature Fusion," *Mathematics*, vol. 13, no. 7, pp. 1–30, 2025, doi: 10.3390/math13071100.

[5] 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 Prepr. arXiv2502.12488*, 2025.

[6] 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 *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.

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

[8] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 14200–14213, 2021.

[9] 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," *Sci. Rep.*, vol. 15, no. 1, p. 25440, 2025.

[10] A. de Santana Correia and E. L. Colombini, "Attention, please! A survey of neural attention models in deep learning," *Artif. Intell. Rev.*, vol. 55, no. 8, pp. 6037–6124, 2022.

[11] 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 Sens.*, vol. 13, no. 15, p. 2965, 2021.

[12] M. Hassanin, S. Anwar, I. Radwan, F. S. Khan, and A. Mian, "Visual attention methods in deep learning: An in-depth survey," *Inf. Fusion*, vol. 108, p. 102417, 2024.

[13] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[14] 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," pp. 21–26, 2020, doi: 10.21437/avsp.2019-5.

[15] 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 *Proceedings of the conference. Association for computational linguistics. Meeting*, 2019, p. 6558.

[16] G. Brauwers and F. Frasincar, "A general survey on attention mechanisms in deep learning," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 3279–3298, 2021.

[17] Y.-H. Lee, D.-W. Jang, J.-B. Kim, R.-H. Park, and H.-M. Park, "Audio–visual speech recognition based on dual cross-modality attentions with the transformer model," *Appl. Sci.*, vol. 10, no. 20, p. 7263, 2020.

[18] R. Gnana Praveen, E. Granger, and P. Cardinal, "Audio-Visual Fusion for Emotion Recognition in the Valence-Arousal Space Using Joint Cross-Attention," *arXiv e-prints*, p. arXiv-2209, 2022.

[19] A. V. Geetha, T. Mala, D. Priyanka, and E. Uma, "Multimodal Emotion Recognition with Deep Learning: Advancements, challenges, and future directions," *Inf. Fusion*, vol. 105, no. December 2023, p. 102218, 2024, doi: 10.1016/j.inffus.2023.102218.

[20] A. Farinhas, A. F. T. Martins, and P. M. Q. Aguiar, "Multimodal continuous visual attention mechanisms," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1047–1056.

[21] N. Khatri, T. Laakkonen, and J. Liu, "On the anatomy of attention," *arXiv Prepr. arXiv2407.02423*, 2024.

[22] R. G. Praveen and J. Alam, "Recursive Joint Cross-Modal Attention for Multimodal Fusion in Dimensional Emotion Recognition," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 4803–4813, 2024.

[23] C. Liu, Z. Mao, T. Zhang, A.-A. Liu, B. Wang, and Y. Zhang, "Focus your attention: A focal attention for multimodal learning," *IEEE Trans. Multimed.*, vol. 24, pp. 103–115, 2020.

[24] N. E. H. Dehimi and Z. Tolba, "Attention mechanisms in deep learning: Towards explainable artificial intelligence," in *2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, IEEE, 2024, pp. 1–7.

[25] P. H. Martins, V. Niculae, Z. Marinho, and A. F. T. Martins, "Sparse and structured visual attention," in *2021 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2021, pp. 379–383.

[26] X. Zou, C. Tang, W. Zhang, K. Sun, and L. Jiang, "Hierarchical attention learning for multimodal classification," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2023, pp. 936–941.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

[27] S. Lu, M. Liu, L. Yin, Z. Yin, and X. Liu, "The multi-modal fusion in visual question answering : a review of attention mechanisms," pp. 1–29, 2023, doi: 10.7717/peerj-cs.1400.

[28] M. R. Kibria, S. Lafond, and J. Arslan, "Decoding the Multimodal Maze: A Systematic Review on the Adoption of Explainability in Multimodal Attention-based Models," *arXiv Prepr. arXiv2508.04427*, 2025.

[29] R. Flores, M. L. Tlachac, A. Shrestha, and E. A. Rundensteiner, "WavFace: A Multimodal Transformer-based Model for Depression Screening," *IEEE J. Biomed. Heal. Informatics*, 2025.

[30] J. Wang, A. Zheng, L. Liu, C. Li, R. He, and J. Tang, "Adaptive Interaction and Correction Attention Network for Audio-Visual Matching," *IEEE Trans. Inf. Forensics Secur.*, 2025.

[31] X. Jiang, X. Bai, and L. Yin, "The Latest Research Progress of Attention Mechanism in Deep Learning," pp. 82–89, 2025.

[32] T. Ruan and S. Zhang, "Towards understanding how attention mechanism works in deep learning," *arXiv Prepr. arXiv2412.18288*, 2024.

[33] F. Zhao, C. Zhang, and B. Geng, "Deep multimodal data fusion," *ACM Comput. Surv.*, vol. 56, no. 9, pp. 1–36, 2024.

[34] Y. Yue, "Multimodal Learning Data Fusion and Analysis Based on Self-Attention Mechanism," in *2025 IEEE 5th International Conference on Electronic Technology, Communication and Information (ICETCI)*, IEEE, 2025, pp. 1040–1047.

[35] S. Qian and C. Wang, "COM: Contrastive Masked-attention model for incomplete multimodal learning," *Neural Networks*, vol. 162, pp. 443–455, 2023.

[36] Z. Yuan, Y. Liu, H. Xu, and K. Gao, "Noise imitation based adversarial training for robust multimodal sentiment analysis," *IEEE Trans. Multimed.*, vol. 26, pp. 529–539, 2023.

[37] E. Mozhegova, A. M. Khattak, A. Khan, R. Garaev, and B. Rasheed, "Assessing the adversarial robustness of multimodal medical AI systems: insights into vulnerabilities and modality interactions," *Front. Med.*, vol. 12, p. 1606238, 2025.

[38] J. Wang, A. Liu, X. Bai, and X. Liu, "Universal adversarial patch attack for automatic checkout using perceptual and attentional bias," *IEEE Trans. Image Process.*, vol. 31, pp. 598–611, 2021.

[39] S. Zhang, W. Chen, X. Li, Q. Liu, and G. Wang, "APBAM: Adversarial perturbation-driven backdoor attack in multimodal learning," *Inf. Sci. (Ny).*, vol. 700, p. 121847, 2025.

[40] D. Michelsanti *et al.*, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1368–1396, 2021.

[41] R. Karani and S. Desai, "Review on multimodal fusion techniques for human emotion recognition," *Int. J. Adv. Comput. Sci. Appl*, vol. 13, pp. 287–296, 2022.

[42] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Multimodal Transformer Fusion For Continuous Emotion Recognition National Laboratory of Pattern Recognition , Institute of Automation , Chinese Academy of CAS Center for Excellence in Brain Science and Intelligence Technology , Beijing , China School of A," *ICASSP 2020 - 2020 IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 3507–3511, 2020.

[43] I. Galanakis, R. F. Soldatos, N. Karanikolas, A. Voulodimos, I. Voyiatzis, and M. Samarakou, "Early and Late Fusion for Multimodal Aggression Prediction in Dementia Patients: A Comparative Analysis," *Appl. Sci.*, vol. 15, no. 11, p. 5823, 2025.

[44] D. Ortiz-Perez, M. Benavent-Lledo, D. Mulero-Pérez, D. Tomás, and J. Garcia-Rodriguez, "Multimodal Fusion Strategies for Emotion Recognition," in *2024 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2024, pp. 1–8.

[45] N. Saeed, M. Alam, and R. G. Nyberg, "A multimodal deep learning approach for gravel road condition evaluation through image and audio integration," *Transp. Eng.*, vol. 16, p. 100228, 2024.

[46] S. Li and H. Tang, "Multimodal alignment and fusion: A survey," *arXiv Prepr. arXiv2411.17040*, 2024.

[47] Y. Zhang, D. Sidibé, O. Morel, and F. Mériaudeau, "Deep multimodal fusion for semantic image segmentation: A survey," *Image Vis. Comput.*, vol. 105, p. 104042, 2021.

[48] H. Liu, W. Li, and B. Yang, "Robust audio-visual speech recognition based on hybrid fusion," in *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, 2021, pp. 7580–7586.

[49] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," *IEEE Trans. circuits Syst. video Technol.*, vol. 28, no. 10, pp. 3030–3043, 2017.

[50] H. Kumar and M. Aruldoss, "Advanced optimal cross-modal fusion mechanism for audio-video based artificial emotion recognition," *Informatica*, vol. 49, no. 12, 2025.

[51] V. John and Y. Kawanishi, "Audio and video-based emotion recognition using multimodal transformers," in *2022 26th International Conference on Pattern Recognition (ICPR)*, IEEE, 2022, pp. 2582–2588.

[52] J. Huang *et al.*, "Continuous multimodal emotion prediction based on long

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

short term memory recurrent neural network," *AVEC 2017 - Proc. 7th Annu. Work. Audio/Visual Emot. Challenge, co-located with MM 2017*, pp. 11–18, 2017, doi: 10.1145/3133944.3133946.

[53] A. Alasiry, M. Al-Hussain, M. Turki-Hadj Alouane, and N. Ben Hadj-Alouane, "Efficient audio-visual emotion recognition approach," *Multimed. Tools Appl.*, 2025, doi: 10.1007/s11042-024-20572-6.

[54] F. Harby, M. Alohali, A. Thaljaoui, and A. S. Talaat, "Exploring Sequential Feature Selection in Deep Bi-LSTM Models for Speech Emotion Recognition," *Comput. Mater. Contin.*, vol. 78, no. 2, pp. 2689–2719, 2024, doi: 10.32604/cmc.2024.046623.

[55] R. S. Kiziltepe, J. Q. Gan, and J. J. Escobar, "Integration of feature and decision fusion with deep learning architectures for video classification," *IEEE Access*, vol. 12, pp. 19432–19446, 2024.

[56] J. Kim, "applied sciences Audio-Visual Action Recognition Using Transformer Fusion Network," 2024.

[57] J. Selva, A. S. Johansen, S. Escalera, K. Nasrollahi, T. B. Moeslund, and A. Clapés, "Video transformers: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 12922–12943, 2023.

[58] J.-H. Kim and C. S. Won, "Audio-visual action recognition using transformer fusion network," *Appl. Sci.*, vol. 14, no. 3, p. 1190, 2024.

[59] M. B. Shaikh, D. Chai, S. M. S. Islam, and N. Akhtar, "From CNNs to transformers in multimodal human action recognition: A survey," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 20, no. 8, pp. 1–24, 2024.

[60] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 2, no. 2, pp. 117–128, 2018.

[61] M. Muzammel, H. Salam, and A. Othmani, "End-to-end multimodal clinical depression recognition using deep neural networks: A comparative analysis," *Comput. Methods Programs Biomed.*, vol. 211, p. 106433, 2021.

[62] X. Zhao *et al.*, "Integrating audio and visual modalities for multimodal personality trait recognition via hybrid deep learning," *Front. Neurosci.*, vol. 16, p. 1107284, 2023.

[63] S. Islam *et al.*, "A comprehensive survey on applications of transformers for deep learning tasks," *Expert Syst. Appl.*, vol. 241, p. 122666, 2024.

[64] K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa, "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets," *Vis. Comput.*, vol. 38, no. 8, pp. 2939–2970, 2022, doi: 10.1007/s00371-021-02166-7.

[65] A. Zubair and F. Al Rashed, "Deep Learning Algorithms for Multimodal Interaction Using Speech and Motion Data in Virtual Reality Systems," *PIQM*, vol. 1, no. 4, 2024.

[66] P. Zhang, J. Wang, M. Wan, S. Chang, L. Ding, and P. Shi, "Multi-Relation Learning Network for audio-visual event localization," *Knowledge-Based Syst.*, vol. 310, p. 112925, 2025.

[67] S. Liu, W. Quan, C. Wang, Y. Liu, B. Liu, and D.-M. Yan, "Dense modality interaction network for audio-visual event localization," *IEEE Trans. Multimed.*, vol. 25, pp. 2734–2748, 2022.

[68] T. Hongping, "Multi-Modal Recognition Via Multi-Head Attention Fusion of Text, Visual and Audio Content on Video Datasets," in *2025 International Conference on Electronics and Renewable Systems (ICEARS)*, IEEE, 2025, pp. 1405–1410.

[69] L. A. Hendricks, J. Mellor, R. Schneider, J.-B. Alayrac, and A. Nematzadeh, "Decoupling the role of data, attention, and losses in multimodal transformers," *Trans. Assoc. Comput. Linguist.*, vol. 9, pp. 570–585, 2021.

[70] W. Song, S. Ren, and B. Hu, "Interpretable Learning Method Based on Causal Interactive Attention," *IEEE Access*, 2025.

[71] I. Kukanov and J. W. Ng, "KLASSify to Verify: Audio-Visual Deepfake Detection Using SSL-based Audio and Handcrafted Visual Features," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 13707–13713.

[72] Z. Liu, Y. Cao, J. Chen, and J. Li, "A hierarchical reinforcement learning algorithm based on attention mechanism for UAV autonomous navigation," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 11, pp. 13309–13320, 2022.

[73] W. Wang and J. Ma, "A review: Applications of machine learning and deep learning in aerospace engineering and aero-engine engineering," *Adv. Eng. Innov.*, vol. 6, no. 1, pp. 54–72, 2024, doi: 10.54254/2977-3903/6/2024060.

[74] A. I. Middya, B. Nag, and S. Roy, "Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities," *Knowledge-Based Syst.*, vol. 244, p. 108580, 2022, doi: 10.1016/j.knosys.2022.108580.

[75] S. Peerbasha, M. I. Habelalmateen, and T. Saravanan, "Multimodal Transformer Fusion for Sentiment Analysis using Audio, Text, and Visual Cues," in *2025 International Conference on Intelligent Systems and Computational Networks (ICISCN)*, IEEE, 2025, pp. 1–6.

[76] X. Liu, N. Xia, J. Zhou, Z. Li, and D. Guo, "Towards energy-efficient audio-visual classification via multimodal interactive spiking neural network," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 21, no. 5, pp. 1–24, 2025.

[77] G. Sun *et al.*, "Fine-grained audio-visual joint representations for multimodal large language models," *arXiv Prepr. arXiv2310.05863*, 2023.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

[78] M. Brousmiche, J. Rouat, and S. Dupont, "Multimodal attentive fusion network for audio-visual event recognition," *Inf. Fusion*, vol. 85, pp. 52–59, 2022.

[79] J. Li and Y. Tian, "From waveforms to pixels: A survey on audio-visual segmentation," *arXiv Prepr. arXiv2508.03724*, 2025.

[80] X. Zhao, Y. Wang, and X. Cai, "A ResNet-based audio-visual fusion model for piano skill evaluation," *Appl. Sci.*, vol. 13, no. 13, p. 7431, 2023.

[81] H. Zhou, J. Du, Y. Zhang, Q. Wang, Q.-F. Liu, and C.-H. Lee, "Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition," *IEEE/ACM Trans. audio, speech, Lang. Process.*, vol. 29, pp. 2617–2629, 2021.

[82] L. Parcalabescu and A. Frank, "Mm-shap: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 4032–4059.

[83] S. Khaled, M. E. Ragab, A. K. Helmy, W. Medhat, and E. H. Mohamed, "Ar-MuSA: A Multimodal Benchmark Dataset and Evaluation Framework for Arabic Sentiment Analysis.," *Int. J. Intell. Eng. Syst.*, vol. 18, no. 4, 2025.

[84] J. Dhanith, S. Venkatraman, V. Sharma, S. Malarvannan, and M. Narendra, "Multimodal Emotion Recognition using Audio-Video Transformer Fusion with Cross Attention," 2024.

[85] A. Lamichhane and G. Karn, "CNN-BiLSTM based Facial Emotion Recognition," *Int. J. Eng. Technol.*, vol. 2, no. 1, pp. 227–236, 2024, doi: 10.3126/injet.v2i1.72579.

[86] G. Udahemuka, K. Djouani, and A. M. Kurien, "Multimodal Emotion Recognition Using Visual, Vocal and Physiological Signals: A Review," *Appl. Sci.*, vol. 14, no. 17, 2024, doi: 10.3390/app14178071.

[87] Y. Wu, Q. Mi, and T. Gao, "A Comprehensive Review of Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions," *Biomimetics*, vol. 10, no. 7, 2025, doi: 10.3390/biomimetics10070418.

[88] E. Ghaleb, J. Niehues, and and S. Asteriadis, "MULTIMODAL ATTENTION-MECHANISM FOR TEMPORAL EMOTION RECOGNITION," *IEEE Int. Conf. Image Process.*, pp. 251–255, 2020.