Scientific Journal of Computer Science

futuristech
PT. Teknologi Futuristik Indonesia

Received: February 6, 2026; Revised: May 10, 2026; Accepted: May 12, 2026; Published: May 15, 2026; Current version: May 15, 2026. e-ISSN: 3110-3170
Digital Object Identifier 10.64539/sjcs.v2i2.2026.438

Review

A Structured Survey of Attention Mechanisms in Audio-Visual Fusion: Architectures, Challenges, and Evaluation Frameworks

Rexcharles Enyinna Donatus[1,2,*], Oludele Awodele[3], Osondu Everestus Oguike[4], Amina Sambo-Magaji[5]

[1] Africa Centre of Excellence on Technology Enhanced Learning, National Open University of Nigeria, Abuja 900108, Nigeria; e-mail: charlly4eyims@yahoo.com (R. E. Donatus).
[2] Aerospace Engineering Department, Air Force Institute of Technology, Kaduna and 800283, Nigeria.
[3] Department of Computer Science, Babcock University, Ilishan-Remo and 121103, Ogun, Nigeria.
[4] Department of Computer Science, University of Nigeria, Nsukka and 410101, Enugu, Nigeria.
[5] Digital Literacy & Capacity Development Department, National Information Technology Development Agency, Abuja and 900104, Nigeria.
* Correspondence

The authors received no financial support for the research, authorship, and/or publication of this article.

Abstract: Audio-visual fusion plays an important role in multimodal artificial intelligence, particularly in applications such as speech processing, emotion recognition, and video understanding, where information from sound and vision improves performance and contextual understanding. Recent developments are driven by attention mechanisms and transformer-based models, which enable more flexible and context-aware interaction within and across modalities compared to conventional fusion approaches. Despite these advances, challenges remain, including sensitivity to noisy or missing modalities, modality imbalance, limited interpretability, and high computational cost. This paper presents a structured survey of attention mechanisms in audio-visual fusion, with emphasis on architectural design and evaluation practices across multiple application domains. A structured survey methodology inspired by PRISMA principles is used to identify and select relevant studies, followed by comparative analysis of model architectures, training strategies, and evaluation methods. The findings show that transformer-based and attention-centered architectures have become increasingly prominent and achieve strong performance across tasks. However, these approaches involve trade-offs between robustness, interpretability, and computational efficiency, and remain sensitive to noise and modality imbalance. Evaluation practices are also inconsistent, with limited use of standardized and robustness-focused metrics. The survey provides an attention-centered taxonomy of audio-visual fusion methods and synthesizes current approaches and evaluation strategies. It identifies key challenges and outlines directions for improving robustness, interpretability, and efficiency in practical deployment.

Keywords: Multimodal fusion; Audio-visual; Deep learning; Attention mechanisms; Temporal modeling; Cross-modal attention.

1. Introduction

Multimodal learning seeks to jointly exploit complementary signals such as audio and vision to enable richer and more reliable machine perception than unimodal approaches [1]-[4]. In audio-visual settings, combining spectro-temporal audio patterns with spatio-temporal visual cues supports more robust interpretation of speech, affect, and events, particularly under challenging conditions [2], [5], [6]. This integration has become central in applications such as speech processing, where visual lip movements complement degraded acoustic signals, and emotion recognition, where facial expressions and vocal prosody jointly convey affect [7]-[10]. Beyond these domains, audio-visual fusion supports event and action recognition as well as human–computer interaction, enabling systems to better interpret complex real-world scenarios [3], [6], [8].

Early approaches to multimodal fusion relied on feature-level concatenation or decision-level aggregation, which provided limited capacity to model cross-modal dependencies [4], [6], [11]. With the advancement of deep learning, convolutional and recurrent architectures became standard for encoding audio spectrograms and visual frames, improving unimodal representation learning but often still relying on relatively simple fusion strategies [2], [4], [12], [13]. More recently, attention mechanisms and transformer-based models have enabled more structured intra- and inter-modal interactions. Self-attention and cross-modal attention allow models to focus on salient regions within each modality and to learn fine-grained correspondences between modalities, while transformer-based designs support flexible and scalable fusion strategies [5], [7], [8], [14].

These developments have led to strong performance gains across tasks such as audio-visual speech recognition, emotion recognition, and video classification, where attention-based models can explicitly capture both intra-modal structure and cross-modal alignment [6], [7], [11], [15]. However, their effectiveness is often conditioned on data quality and modality balance. Modality imbalance, where one modality dominates the learning process, can reduce the contribution of complementary signals, while noisy or weakly aligned data can degrade cross-modal learning [3], [9], [12], [16], [17]. Although attention mechanisms improve selective feature integration, they remain sensitive to these conditions and do not fully resolve robustness challenges [18]-[20]. Despite these advances, a key challenge remains: understanding how different attention mechanisms behave under varying data conditions and how their design choices influence robustness, interpretability, and fusion effectiveness.

Additional limitations arise in interpretability and computational efficiency. While attention weights provide some insight into model behavior, understanding global attention patterns and modality contributions remains challenging without specialized analysis tools [21]. Furthermore, the computational cost of dense attention, which scales with sequence length and modality size, presents practical constraints for real-world deployment. This has motivated the development of more efficient designs, including bottleneck-based fusion, sparse attention, and lightweight transformer variants [1], [5], [6], [22].

A number of surveys have examined multimodal learning, deep fusion strategies, and multimodal architectures across different application domains [1]-[3], [8], [21], [23]. These studies provide broad taxonomies and cover a wide range of modalities and model families. However, they often treat attention as one component among many and do not provide a focused analysis of how different attention mechanisms influence audio-visual fusion. In addition, architectural design, evaluation practices, and robustness considerations are frequently discussed separately, making it difficult to understand their combined impact. Existing reviews are also often organized around specific applications, such as emotion recognition or speech processing, which limits cross-domain comparison of attention-based approaches. As a result, there remains a need for a structured, attention-focused synthesis that connects architectural design choices with evaluation strategies and robustness properties across audio-visual tasks.

To address these limitations in a unified manner, this paper provides a structured and attention-focused survey of audio–visual fusion. The main contributions are as follows:

- A systematic taxonomy of attention mechanisms in audio–visual systems, covering intra-modal, inter-modal, hierarchical, and hybrid attention designs.
- A unified analysis linking attention mechanisms to multimodal fusion strategies and deep learning architectures.
- A structured review of evaluation practices, including performance, robustness, and computational efficiency considerations.
- An identification of key challenges, including modality imbalance, noise sensitivity, interpretability, and scalability constraints.
- A synthesis of open research directions to support the development of more robust and interpretable multimodal systems.

The remainder of this paper is organized as follows. Section 2 describes the methodology, including the structured and PRISMA-inspired process used for literature selection and analysis. Section 3 presents the taxonomy of attention-based audio–visual fusion architectures. Section 4 critically synthesizes existing studies and identifies major research gaps and challenges. Section 5 reviews evaluation metrics, benchmarks, and assessment frameworks. Section 6 concludes the study, while Section 7 outlines future research directions.

2. Methodology

This study adopts a structured survey methodology guided by systematic literature review (SLR) principles to ensure transparency, reproducibility, and analytical rigor. While the study aims to synthesize trends rather than perform statistical meta-analysis, PRISMA reporting concepts are incorporated to organize the literature identification and screening procedures. Therefore, this work is positioned as a semi-systematic survey, combining systematic search procedures with qualitative synthesis of research developments in attention-based audio-visual fusion.

The methodological workflow consists of three main stages:

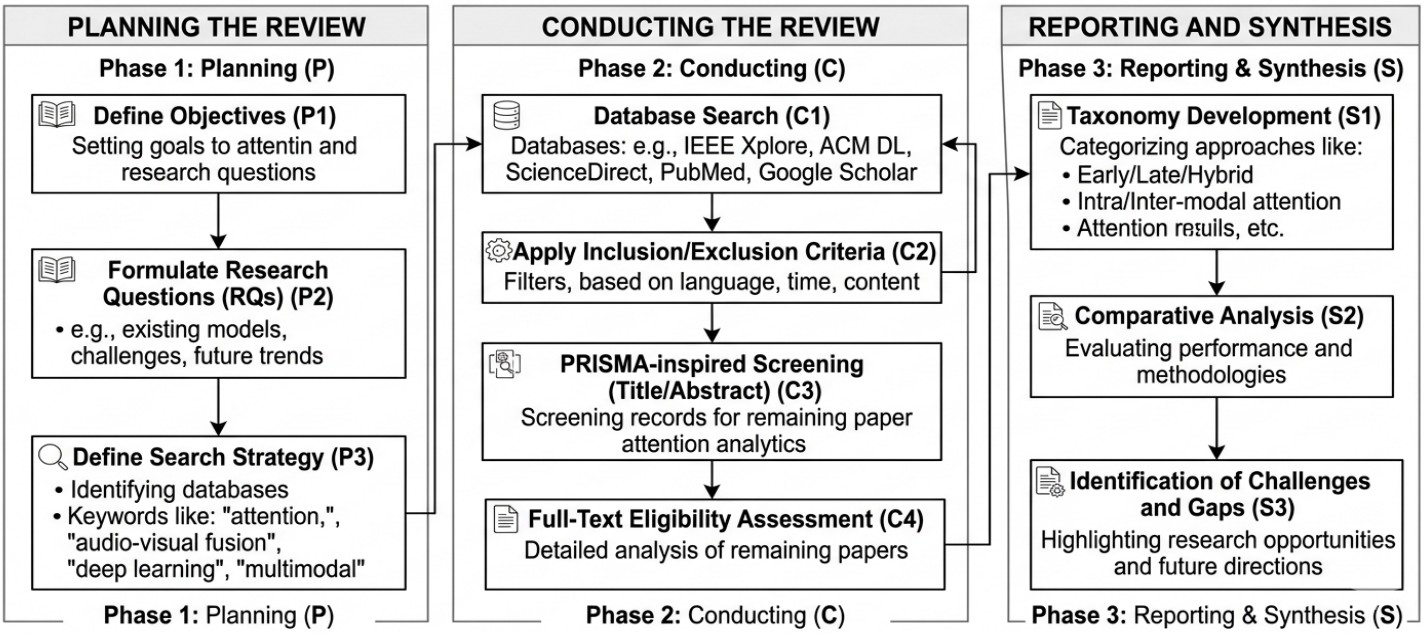


Figure 1. Workflow for Semi-Systematic Literature Review on Attention-Based Audio – Visual Fusion.

Table 1. Research Questions Guiding the Survey.

RQ ID	Research Question	Purpose
RQ1	What types of attention mechanisms are used in audio-visual fusion architectures?	To categorize and analyze different attention designs such as self-attention, cross-modal attention, co-attention, and hierarchical attention.
RQ2	How do different fusion strategies interact with attention mechanisms in audio-visual systems?	To examine how early, late, and hybrid fusion approaches integrate with attention-based models.
RQ3	What evaluation metrics and protocols are used to assess attention-based audio-visual fusion models?	To review standard performance metrics as well as robustness- and efficiency-oriented evaluation practices.
RQ4	What are the key challenges affecting attention-based audio-visual fusion?	To identify issues such as modality imbalance, noise sensitivity, missing modalities, interpretability, and computational complexity.
RQ5	What research gaps and future directions exist in attention-based audio-visual fusion?	To highlight limitations in current approaches and guide future research on robust, interpretable, and efficient models.

1) Planning the review
2) Conducting the review
3) Reporting and synthesizing the review

An overview of the research workflow is illustrated in Figure 1.

2.1. Planning the Review

2.1.1. Research Objectives

The objective of this study is to systematically examine how attention mechanisms are designed, evaluated, and applied within audio-visual fusion systems. The review focuses on identifying architectural patterns, evaluation practices, strengths, limitations, and emerging research challenges.

2.1.2. Research Questions

To guide the analysis and maintain methodological consistency, five research questions (RQs) were formulated (see Table 1). These questions define the scope of literature selection and structure the subsequent discussion sections.

2.1.3. Data Sources

Literature was collected from indexing databases widely used in computer science and multimodal learning research:

- Google Scholar
- Semantic Scholar
- PubMed
- Scopus

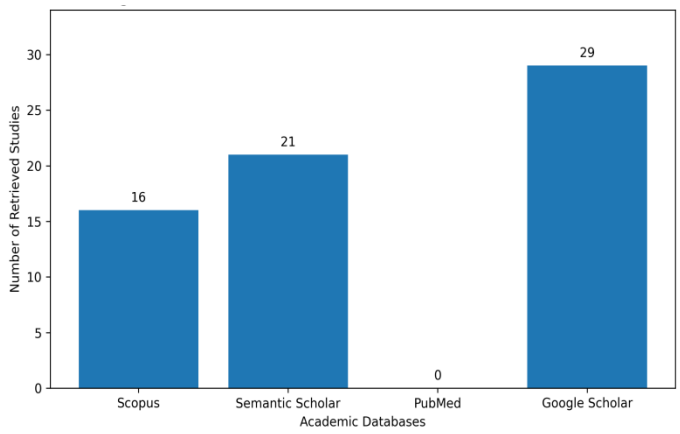


Figure 2. Distribution of Retrieved Studies Across Academic Databases.

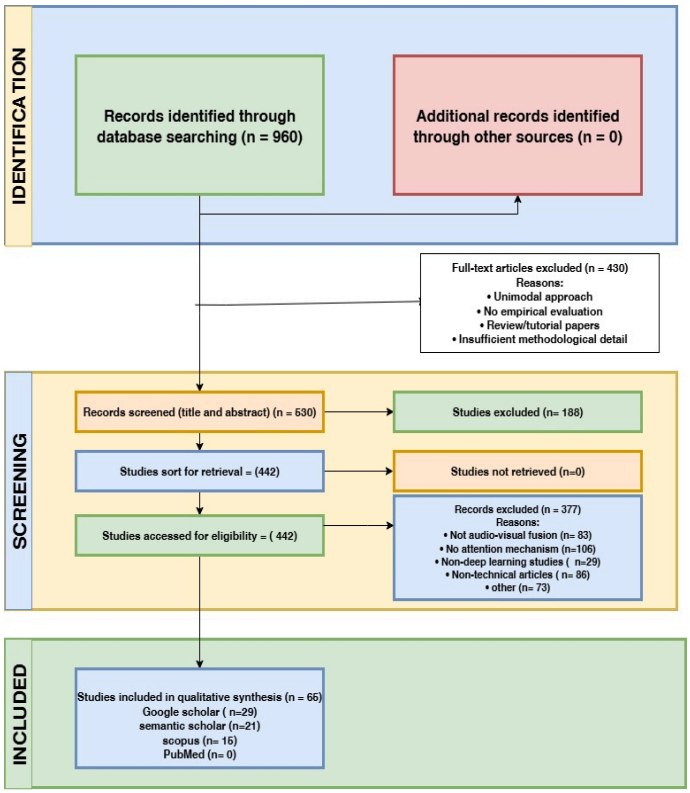


Figure 3. PRISMA Flow Diagram.

These databases were selected due to their complementary coverage of peer-reviewed journals, conference proceedings, and emerging preprint research relevant to deep learning and multimodal artificial intelligence.

The distribution of retrieved studies across databases is presented in Figure 2.

2.1.4. Search Strategy and Query Formulation

Search queries were designed by combining keywords related to modality, fusion strategy, and attention mechanisms.

Search Query: ("audio-visual" OR multimodal) AND ("attention mechanism" OR "cross-modal attention" OR co-attention OR transformer OR self-attention) AND ("deep learning" OR neural network) AND (fusion OR multimodal fusion) NOT (survey OR review)

Equivalent keyword combinations were adapted for each database according to its indexing syntax.

2.2. Conducting the Review

2.2.1. Eligibility Criteria

Eligibility criteria were established prior to screening to ensure consistency and reproducibility (see Table 2).

2.2.2. Literature Identification

The database search initially identified **960 records** across all sources. Duplicate entries were removed using automated and manual verification procedures. The literature selection workflow follows PRISMA reporting principles [14] (See Figure 3).

2.2.3. Screening Process

The screening process was conducted in three stages:

1) **Stage 1 — Title and Abstract Screening**
   Irrelevant studies unrelated to attention-based audio-visual fusion were excluded.
2) **Stage 2 — Full-Text Assessment**
   Remaining articles were evaluated based on methodological relevance, experimental validation, and alignment with research questions.
3) **Stage 3 — Quality and Relevance Assessment**
   Studies were assessed according to:
   - clarity of attention mechanism design
   - completeness of evaluation methodology
   - contribution to multimodal fusion research

After screening, **65 studies** were selected for qualitative synthesis.

2.2.4. PRISMA Selection Summary

The PRISMA process can be summarized as follows:

- Records identified: 960
- After duplicate removal: 530
- After title/abstract screening: 442
- Full-text assessed: 442
- Record further screened for eligibility
- Final included studies: 65

2.3. Reporting and Data Extraction

For each selected study, structured data extraction was performed using a standardized review template capturing:

- attention mechanism type
- fusion strategy
- model architecture
- datasets used
- evaluation metrics
- strengths and limitations

Attention mechanisms were categorized into:

Table 2. Inclusion and Exclusion Criteria.

Criteria	Inclusion	Exclusion
Study Focus	Audio-visual fusion with attention mechanisms	Unimodal studies
Methodology	Deep learning architectures	Non-DL methods
Evaluation	Empirical validation reported	No experiments
Publication Type	Journal or conference papers	Editorials, opinions
Language	English	Non-English
Scope	Multimodal learning applications	Irrelevant domains

- self-attention
- cross-modal attention
- co-attention
- hierarchical attention
- temporal attention

This structured extraction enabled systematic comparison across heterogeneous audio-visual learning tasks.

2.4. Data Synthesis and Analysis Strategy

The synthesis process combined quantitative overview analysis with qualitative thematic comparison. The analysis was organized according to the predefined research questions:

- architectural trends (RQ1–RQ2)
- evaluation practices (RQ3)
- challenges and limitations (RQ4)
- research gaps and future opportunities (RQ5)

Rather than claiming methodological superiority, the study emphasizes trade-offs, limitations, and contextual effectiveness of attention mechanisms, aligning with the balanced evaluation perspective recommended in recent survey methodologies.

To guide the structured analysis presented in this survey, a set of research questions is defined. These questions provide a framework for organizing the review of attention mechanisms, fusion strategies, evaluation practices, and associated challenges in audio-visual fusion.

3. Background and Fundamental Concepts

3.1. Multimodal Fusion in Audio–Visual Learning

Multimodal fusion integrates heterogeneous data sources to exploit complementary information and improve robustness compared with unimodal systems [6], [15], [24]. In audio–visual (AV) learning, acoustic and visual streams are typically processed through modality-specific encoders and subsequently combined via a fusion mechanism for tasks such as emotion recognition, event classification, and segmentation [11], [25], [26].

Fusion strategies are commonly categorized into early (feature-level), late (decision-level), and hybrid/ intermediate fusion, although these distinctions are often less rigid in deep learning systems where feature extraction and fusion are jointly optimized [24].

- **Early fusion** combines features at the input or shallow layers, enabling direct interaction between modalities. This facilitates joint representation learning but is sensitive to temporal misalignment and noise.
- **Late fusion** aggregates predictions from independently processed modalities. It is robust to missing or degraded modalities but does not explicitly model inter-modal relationships.
- **Hybrid (intermediate) fusion** integrates modalities at multiple levels within the network, allowing both independent processing and interaction. This approach supports more flexible modeling of intra- and inter-modal dependencies.

Recent developments emphasize attention-driven fusion modules as a unifying framework that enables adaptive interaction between modalities rather than static fusion operations [6].

3.2. Attention Mechanisms: Types and Definitions

Attention mechanisms dynamically reweight features by modeling relationships between queries, keys, and values, allowing models to focus on salient information [27], [28]. In AV fusion, attention mechanisms are central to modeling both intra- and inter-modal dependencies, directly addressing **RQ1 (identification and classification of attention mechanisms)**.

3.2.1. Self-Attention (Intra-modal Attention)

Self-attention operates within a single modality, where each element attends to all other elements in the same sequence. Its primary role is to capture long-range temporal or spatial dependencies within audio or visual streams. It is widely used in transformer-based architectures to model contextual relationships in spectrograms, video frames, or tokenized representations [6], [26].

3.2.2. Cross-Attention (Inter-modal Attention)

Cross-attention models interactions between modalities by allowing one modality (query) to attend to another (key–value pairs). Its functional role is to enable direct alignment and information transfer across modalities, such as aligning speech cues with facial expressions in emotion recognition [11], [15], [29]. Figure 4 illustrates a

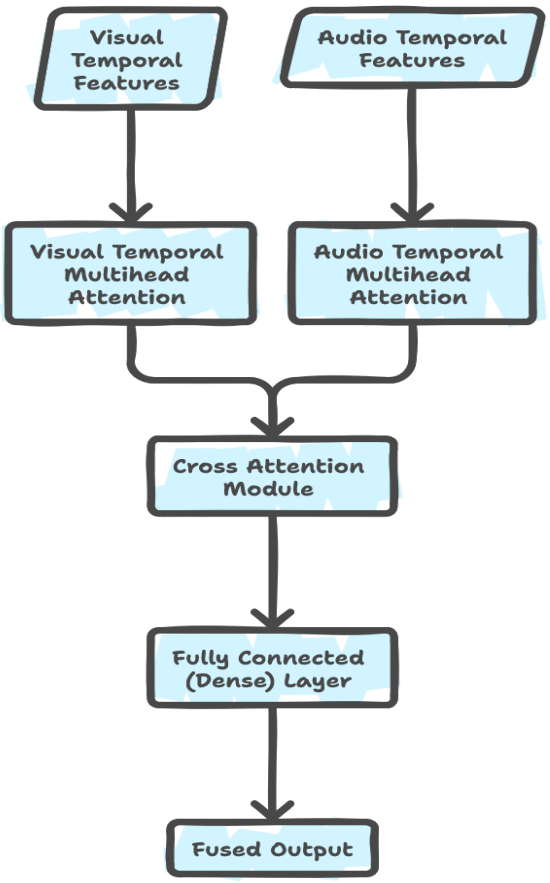


Figure 4. Schematic of the Cross-modal Attention Fusion Module.

typical cross-modal attention module, where modality-specific representations interact through directed attention.

3.2.3. Co-Attention (Collaborative Attention)

Co-attention extends cross-attention by enabling bi-directional or joint attention between modalities. Both modalities simultaneously attend to each other, producing interdependent representations. This mechanism supports mutual feature refinement, preserving both intra- and inter-modal relationships [26], [30].

3.2.4. Hierarchical Attention

Hierarchical attention applies attention at multiple levels (e.g., frame-level, segment-level, sequence-level), capturing both local and global dependencies. Its role is to model structured relationships across temporal and spatial scales before or during fusion [15].

3.2.5. Residual/Correction Attention

Residual or correction attention introduces additive or corrective pathways to refine attention outputs. Its primary function is to stabilize learning and mitigate errors in attention weighting, particularly when one modality is noisy or dominant. This mechanism supports adaptive reweighting and improved robustness in multimodal settings [31].

To consolidate the classification of attention mechanisms discussed above (RQ1) and to summarize their functional roles and limitations (RQ2), Table 3 presents a structured comparison of the principal attention types used in audio–visual fusion systems.

Figure 5 illustrates a representative architecture of attention-based multimodal fusion systems. In this framework, audio and visual inputs are first processed through modality-specific encoders to obtain high-dimensional feature representations. These representations are then passed to a fusion module, where different attention mechanisms are applied.

The self-attention module (SMA) captures intra-modal dependencies within each modality, while the parallel cross-modal attention module (PCMA) facilitates interaction between modalities by modeling inter-modal relationships. The resulting unimodal and multimodal representations are subsequently integrated and passed to a downstream prediction module for tasks such as emotion recognition.

This architecture highlights how intra-modal and inter-modal attention mechanisms are jointly incorporated within a unified processing pipeline.

3.3. Strengths and Limitations of Attention Mechanisms

This subsection addresses RQ2 (strengths, weaknesses, and failure conditions of attention mechanisms).

3.3.1. Strengths

Attention mechanisms contribute to multimodal learning in several ways:

a) They enable selective focus on salient features, improving representation efficiency [11], [15].
b) They support explicit modeling of cross-modal relationships, which is essential for tasks requiring alignment between modalities [25], [26].
c) They provide flexible integration across architectures, including CNNs, RNNs, and transformers.
d) They allow adaptive weighting of modalities, which can improve robustness when modality quality varies.

Attention improves robustness particularly when:

a) modalities are complementary and temporally aligned,
b) informative features are sparsely distributed,
c) and the model benefits from dynamic feature selection across time or space.

3.3.2. Limitations and Failure Conditions

Attention mechanisms do not inherently guarantee robustness and may fail under several conditions:

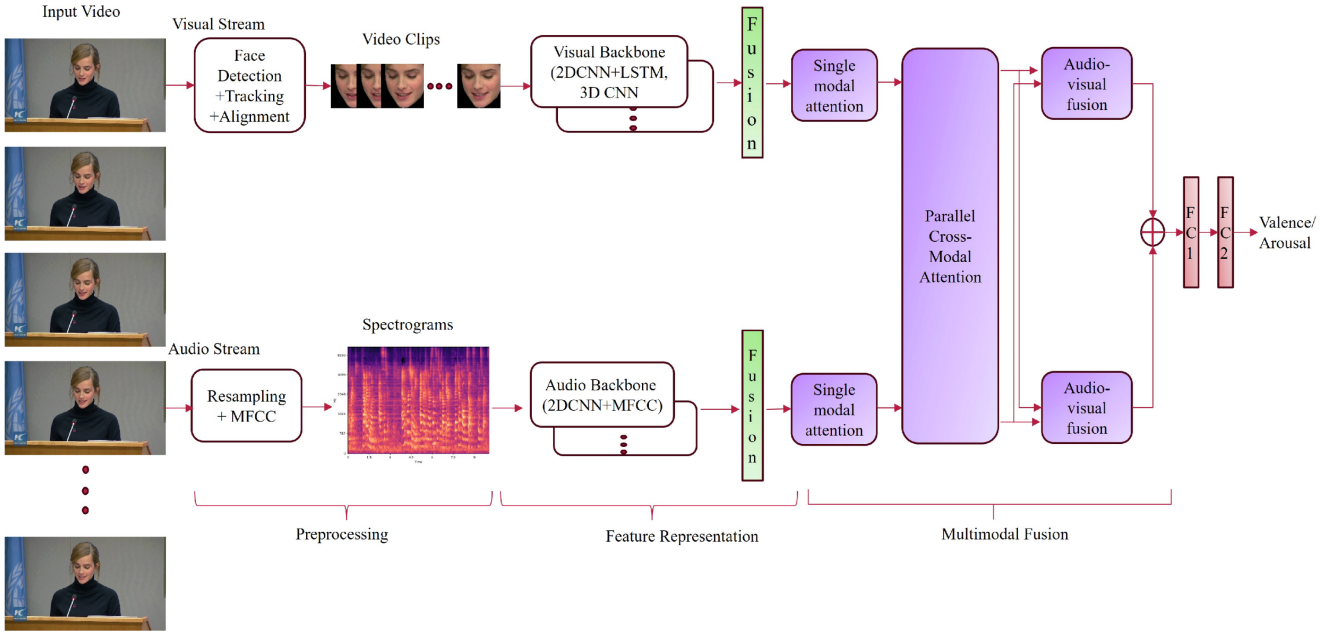


Figure 5. Cross-modal Attention Workflow [15].

Table 3. Classification of Attention Mechanisms in Multimodal Audio–Visual Systems and their Functional Trade-offs.

Attention Type	Mechanism Description	Functional Role	Strengths	Limitations	Representative References
Self-attention	Computes dependencies within a single modality using query–key–value interactions	Models intra-modal temporal/spatial relationships	Captures long-range dependencies; flexible representation learning	Does not model cross-modal interactions; may ignore complementary cues	[6], [26], [28]
Cross-attention	One modality attends to another via query–key–value mapping	Enables directed inter-modal alignment	Explicitly models cross-modal relationships; supports fine-grained alignment	Sensitive to modality misalignment and noise; computational cost increases with sequence length	[11], [15], [25]
Co-attention	Bidirectional or joint attention across modalities	Learns mutual interdependence between modalities	Captures bidirectional dependencies; preserves modality interactions	Increased model complexity; risk of overfitting with limited data	[26], [30]
Hierarchical attention	Multi-level attention across temporal/spatial scales	Aggregates local and global contextual information	Captures multi-scale dependencies; improves contextual representation	Higher computational complexity; requires careful design	[15]
Residual/Correction attention	Adds corrective or residual connections to attention outputs	Stabilizes attention learning and reweights unreliable features	Improves robustness under noisy or imbalanced modalities	May obscure interpretability of attention weights	[31]

a) **Modality imbalance:** When one modality dominates, attention may suppress useful information from weaker modalities.
b) **Noise sensitivity:** Attention can amplify noisy or irrelevant features, especially in misaligned or corrupted inputs.
c) **Missing data:** Standard attention mechanisms do not explicitly handle absent modalities and may produce unreliable weights.
d) **Non-complementary signals:** When modalities provide conflicting information, cross-attention may degrade performance.

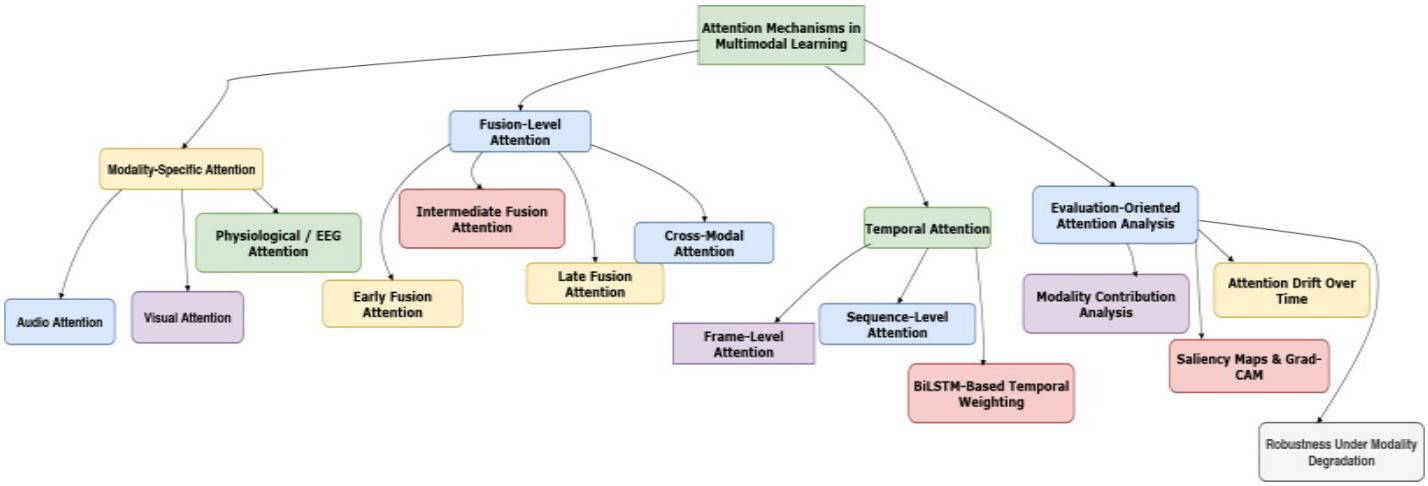


Figure 6. Functional Taxonomy of Attention Mechanisms in Multimodal Audio–Visual Systems.

e) **Computational complexity:** Dense attention over long sequences increases computational cost.
f) **Interpretability limitations:** Attention weights do not always correspond to human-interpretable importance.

These limitations highlight that attention effectiveness depends on data quality, modality balance, and alignment conditions, rather than the mechanism alone.

To provide a structured understanding of how attention mechanisms operate within multimodal audio–visual systems, it is useful to organize them according to their functional role in the processing pipeline. Rather than grouping mechanisms solely by architectural design, this survey adopts a functional taxonomy that distinguishes attention mechanisms based on where and how they are applied, including intra-modal feature modeling, inter-modal fusion, temporal dependency modeling, and evaluation-oriented analysis. This perspective facilitates clearer interpretation of their operational behavior and supports the systematic analysis of their strengths and limitations in relation to different data conditions.

As illustrated in Figure 6, attention mechanisms can be broadly categorized into four functional groups. Modality-specific attention focuses on salient features within individual modalities, while fusion-level attention governs information exchange across modalities. Temporal attention captures sequential dependencies in audio–visual signals, and evaluation-oriented attention supports interpretability and robustness analysis. This functional categorization provides a foundation for comparing different attention mechanisms and understanding their behavior under varying conditions.

3.4. Fusion Strategies and Their Relation to Attention

Fusion strategies define how modalities interact within AV systems, and attention mechanisms can be integrated at different stages of fusion.

a) **Early fusion:** Combines modality features before deep processing. Attention may be applied after concatenation to refine joint representations. While it enables dense interactions, it is sensitive to noise and alignment issues [32].
b) **Late fusion:** Combines modality-specific predictions. Attention can be used to weight modality outputs, improving robustness to missing or unreliable inputs, but inter-modal interactions remain limited [33].
c) **Hybrid/intermediate fusion:** Introduces interaction at multiple levels using attention modules such as cross-attention or co-attention. This approach balances independent processing with inter-modal alignment and is widely used in modern AV systems [6], [15].

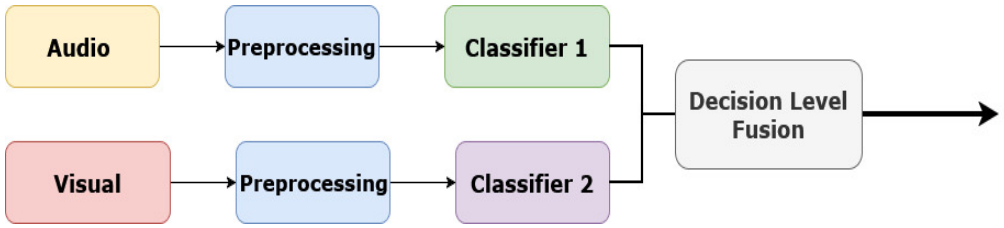
Attention therefore functions as a core mechanism for adaptive fusion, operating either before, during, or after modality interaction.

To clarify the operational differences between fusion strategies and their relationship to attention mechanisms, Table 4 summarizes when each fusion approach is most appropriate, along with their strengths and limitations.

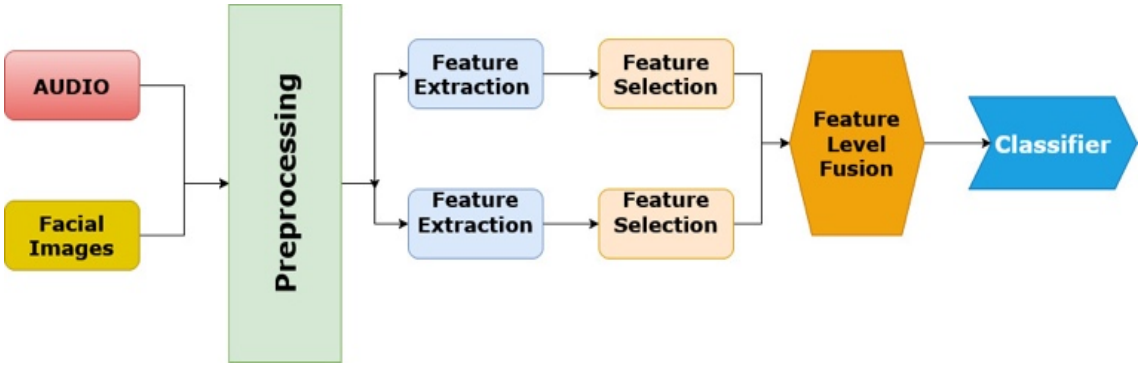
3.5. Deep Learning Architectures for Attention-Based AV Fusion

Attention mechanisms are embedded within various deep learning architectures:

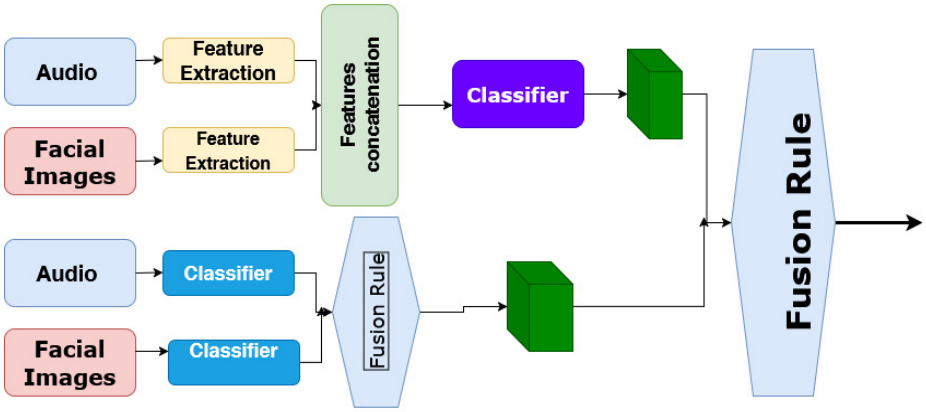
a) **CNN–RNN architectures:** CNNs extract spatial or spectral features, while RNNs or BiLSTMs model temporal dependencies. Attention is applied to emphasize relevant time steps or regions [11], [25].
b) **Transformer-based architectures:** Transformers rely on self-attention and cross-attention to model long-range dependencies and inter-modal interactions [34]. These models support parallel processing and flexible fusion strategies [24], [26].
c) **Hybrid architectures:** Combinations such as CNN–Transformer or CNN–BiLSTM integrate spatial, temporal, and attention-based fusion mechanisms within a unified framework [35].



(a) Decision Level Fusion



(b) Feature Level Fusion



(c) Hybrid Level Fusion

Figure 7. General Representation for Audio – Visual Fusion Methods.

Table 4. Comparison of Multimodal Fusion Strategies in Audio–Visual Systems.

Fusion Strategy	Description	When Effective	Strengths	Limitations	Representative References
Early fusion	Combines modality features at input or shallow layers	When modalities are well-aligned and noise levels are low	Enables direct feature interaction; simple integration	Sensitive to misalignment and noise; increased feature dimensionality	[6], [24]
Late fusion	Combines predictions from modality-specific models	When modalities differ in reliability or may be missing	Robust to missing/noisy modalities; modular design	Limited modeling of cross-modal dependencies	[15], [36]
Hybrid/intermediate fusion	Integrates modalities at multiple network stages using attention or fusion modules	When both intermodal interaction and robustness are required	Balances independent processing and interaction; flexible design	Increased architectural complexity; higher computational cost	[11], [26]

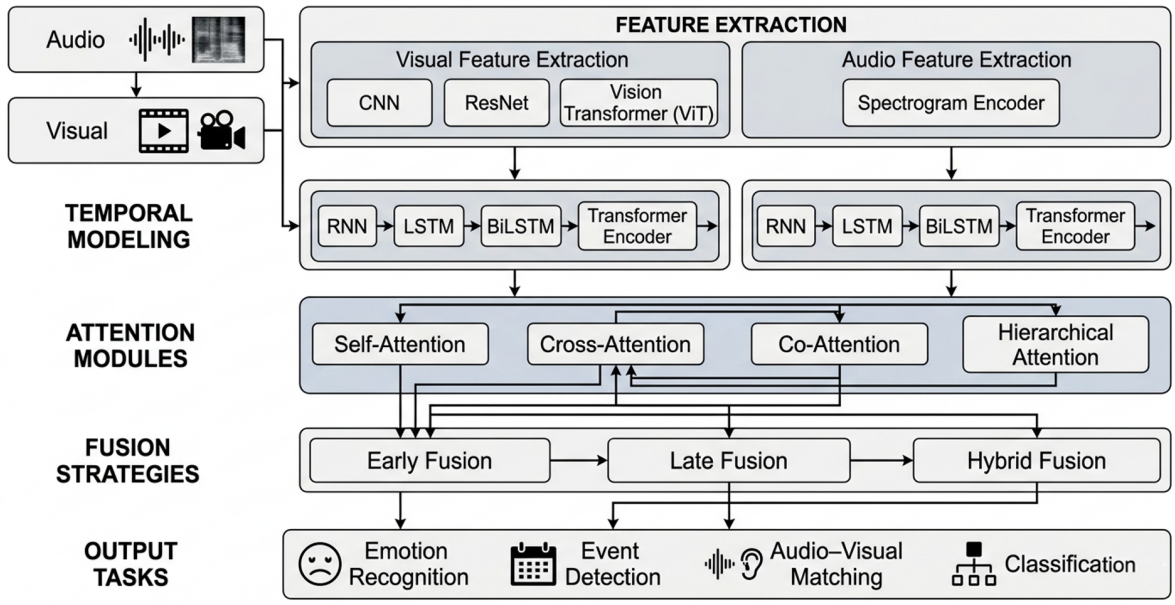


Figure 8. Representative Deep Learning Architecture for Attention Audio-Visual System.

Table 5. Mapping of Deep Learning Architectures to Modalities and Attention Mechanisms in Audio–Visual Systems.

Architecture	Input Modalities	Attention Usage	Functional Role	Limitations	Representative References
CNN-based models	Images, spectrograms	Spatial/channel attention	Extract spatial and spectral features	Limited temporal modeling capability	[11], [24]
RNN/LSTM	Audio, sequential data	Temporal attention	Model sequential dependencies	Difficulty with long-range dependencies	[25]
BiLSTM	Audio, video sequences	Bidirectional temporal attention	Captures past and future context	Higher computational cost than standard RNNs	[11], [30]
Transformer	Audio, video, multimodal tokens	Self- and cross-attention	Models long-range dependencies and inter-modal interactions	Computationally intensive for long sequences	[6], [26]

d) **Emerging architectures:** Recent approaches incorporate attention into alternative paradigms such as spiking–transformer hybrids for multimodal representation learning.

The choice of architecture determines how attention is applied, influencing both model capability (RQ1) and robustness characteristics (RQ2).

To relate attention mechanisms to their architectural implementations (RQ1) and highlight how design choices influence robustness and performance (RQ2), Table 5 maps common deep learning architectures to their input modalities and attention usage. Similarly, Figure 7 illustrates the visualization and classification of fusion methodologies applied in audio–visual systems.

Figure 8 illustrates representative deep learning pipelines used in attention-based audio–visual fusion systems. The architecture highlights the progression from modality-specific feature extraction through temporal modeling and attention-based interaction to multimodal fusion and downstream prediction tasks.

3.6. Link to Research Questions

This section establishes the conceptual foundation required to address the research questions guiding this survey. With respect to **RQ1 (identification and classification of attention mechanisms),** Section 3.2 provides formal definitions and categorization of key attention types, including self-attention, cross-attention, co-attention, hierarchical attention, and residual/correction attention. These mechanisms are further contextualized within multimodal audio–visual architectures in Section 3.5, enabling a structured understanding of how they are implemented in practice.

Regarding **RQ2 (strengths, weaknesses, and failure conditions),** Section 3.3 presents a systematic analysis of the operational characteristics of attention mechanisms, explicitly distinguishing conditions under which attention improves robustness (e.g., selective feature emphasis and adaptive fusion) and conditions under which it may degrade performance (e.g., modality imbalance, noise sensitivity, and missing data).

Together, these components provide the theoretical and structural basis for the comparative evaluation and synthesis presented in subsequent sections.

4. Critical Synthesis & Research Gaps

The literature on multimodal audio–visual deep learning shows a clear shift toward hybrid fusion strategies and transformer-based architectures augmented with attention mechanisms. These approaches are widely adopted because they enable more flexible modeling of temporal dependencies and cross-modal interactions compared with traditional fusion strategies [11], [37], [38]. In particular, cross-modal and hierarchical attention mechanisms have demonstrated effectiveness in capturing complex relationships between modalities, especially in tasks requiring temporal alignment and semantic consistency [15], [38], [39].

Despite these advances, several methodological and practical limitations remain unresolved. The effectiveness of attention-based fusion depends on data quality, modality alignment, and model design. In well-aligned and data-rich environments, attention mechanisms can improve feature integration and representation learning. In contrast, their performance may degrade under modality imbalance, noisy inputs, or missing data, where attention weights may become unstable or biased.

Interpretability remains a central limitation. Although attention weights are often used to visualize model focus, they do not consistently provide reliable explanations of model decisions. Existing approaches, including attention visualization and correction-based attention modules, offer partial insights but lack standardized validation frameworks, particularly under noisy or adversarial conditions [28], [31], [40], [41].

Robustness is also inconsistently reported. While some studies demonstrate improved resilience through adaptive attention and hybrid designs, others highlight sensitivity to modality dominance and dataset bias [42], [43]. These findings indicate that attention mechanisms do not inherently guarantee robustness but require careful architectural and training considerations.

Evaluation practices further limit comparability across studies. Most works rely on accuracy and F1-score, with limited use of modality-aware or interpretability-oriented metrics. As a result, cross-study comparisons of robustness, generalization, and modality contribution remain difficult [11], [38]. In addition, the computational cost of transformer-based architectures remains a practical constraint. While attention enables parallel computation, the overall computational complexity—particularly for long sequences—can limit scalability and real-world deployment [44], [45].

These observations highlight several key research gaps:

1) Limited interpretability and lack of validated explanation frameworks for attention mechanisms
2) Inconsistent robustness under modality imbalance, noise, and missing data
3) Absence of standardized evaluation protocols incorporating modality-aware metrics
4) Limited stress-testing under real-world and adverse conditions
5) High computational cost affecting scalability and deployment

Addressing these gaps is essential for advancing multimodal audio–visual systems toward more reliable, interpretable, and practically deployable solutions.

5. Evaluation Metrics and Benchmarks

The evaluation of multimodal attention mechanisms requires metrics that extend beyond predictive performance to capture how models allocate and adapt attention across modalities. While accuracy and F1-score remain the most commonly reported metrics—particularly on benchmark datasets such as RAVDESS, CREMA-D, and AVE—they provide only a partial assessment of multimodal behavior [46], [47], they provide only a partial view of model behavior in multimodal settings.

Recent studies have introduced modality-aware evaluation approaches that provide additional insight into model behavior. Modality contribution scores quantify the relative influence of each modality within fusion processes and are useful for identifying modality dominance or imbalance [11]. Similarly, saliency drift metrics measure how attention distributions change over time or under perturbations, offering a way to assess stability and consistency in multimodal representations [38].

Another important direction involves cross-modal consistency evaluation, which examines whether models maintain stable predictions when one modality is degraded, noisy, or absent. This is particularly relevant for real-world applications, where multimodal inputs are often imperfect [42].

Despite these developments, most existing studies still rely predominantly on traditional performance metrics, with limited adoption of modality-aware or interpretability-focused evaluation methods. This lack of comprehensive evaluation frameworks makes it difficult to assess robustness, generalization, and modality interaction in a consistent manner across studies [40], [46], [47].

Table 6 summarizes commonly used datasets and evaluation metrics in multimodal audio–visual research, highlighting the limited adoption of modality-aware and saliency-based evaluation approaches. As shown in Table 6, most studies rely on accuracy-based metrics, while only a small number incorporate modality contribution or interpretability-oriented measures, reinforcing the need for more comprehensive evaluation frameworks.

Table 6. Datasets, Metrics, and Modality/Saliency Awareness.

Paper (Year) & Ref	Datasets Used	Metrics Reported	Modality/Saliency Metrics
[11]	RAVDESS, CREMA-D	Accuracy	No
[48]	RAVDESS, SAVEE	Accuracy	No
[49]	CMU-MOSEI	F1, MAE	No
[50]	5 AV datasets	Accuracy, Energy	No
[51]	AVEB (custom)	Accuracy	No
[36], [52]	AVE, UCF51, Kinetics-Sounds	Accuracy	No
[36], [53]	PISA	Accuracy	No
[53], [54]	IEMOCAP, AFEW	Accuracy	No
[55]	6 VL tasks	MM-SHAP (modality contrib.)	Yes (MM-SHAP)
[56]	MOSEI, SNLI	SHAPE (modality contrib.)	Yes (SHAPE)

5.1. Illustrative Case Studies on Attention-Aware Evaluation Metrics

Recent studies demonstrate that attention-aware evaluation can provide insights into multimodal fusion that are not captured by aggregate performance metrics. These approaches highlight how attention is distributed across modalities and how it adapts under different conditions.

Li et al. [57] present an emotion recognition framework combining facial expressions and EEG signals. While multimodal fusion improves classification performance, additional analysis using Grad-CAM reveals spatial attention patterns over EEG regions associated with affective processing. This provides an interpretable indication of modality contribution beyond accuracy metrics.

Seijdel et al. [58] investigate attention-modulated audio–visual integration using neural measurements. Their findings show that when speech quality degrades, visual information is increasingly emphasized, reflecting adaptive modality reweighting. These dynamics illustrate how attention shifts in response to modality reliability.

Vibell et al [35] examine attentional cueing effects on audio–visual perception. Their results indicate that attention allocation across modalities changes depending on task conditions, providing behavioral evidence of modality contribution dynamics.

Across these studies, attention-aware evaluation methods reveal modality dominance, adaptive reweighting, and temporal variability that are not observable through standard performance metrics. This supports the need for evaluation frameworks that incorporate modality contribution and attention dynamics alongside accuracy-based measures.

5.2. Guidelines for Practitioners

The selection of attention mechanisms in multimodal systems should be guided by task requirements and data characteristics. Self-attention is appropriate for modeling intra-modal dependencies, particularly in sequential data, while cross-modal and co-attention mechanisms are suitable for tasks requiring explicit interaction between modalities. Hierarchical attention can be useful in scenarios involving multi-scale temporal or spatial structures.

In practice, attention mechanisms should be designed with consideration for robustness and interpretability. Models should be evaluated under conditions of noise, modality imbalance, and missing data to ensure reliable performance. Incorporating modality-aware evaluation metrics can further support the analysis of model behavior and improve system transparency [15], [59], [60], [61], [62].

6. Conclusion

This study reviewed attention mechanisms in multimodal audio–visual deep learning, focusing on their roles in fusion architectures, performance characteristics, and evaluation practices. The analysis shows that self-attention, cross-modal attention, co-attention, and hierarchical attention are widely used in modern systems to support context-aware integration of audio and visual information.

Collectively, these findings address the research questions by identifying dominant attention mechanisms, examining their integration within fusion architectures, evaluating current assessment practices, and synthesizing key limitations and future research directions.

Transformer-based and hybrid architectures have become prominent due to their ability to model complex dependencies and enable flexible fusion strategies [7], [63]. However, their effectiveness depends on factors such as data quality, modality alignment, and computational resources.

Several challenges remain. These include limited interpretability of attention mechanisms, sensitivity to modality imbalance and noise, and the high computational cost of attention-based architectures [40], [64], [65]. In addition, evaluation practices are not yet standardized, which complicates the comparison of different approaches.

Overall, this review provides a structured synthesis of attention mechanisms in audio–visual fusion, highlighting their capabilities and limitations. By emphasizing the need for improved evaluation, robustness, and interpretability, it outlines key considerations for advancing multimodal AI systems in practical applications.

7. Recommendations for Future Research

Future work should focus on developing standardized evaluation frameworks that incorporate modality contribution, attention stability, and interpretability. Metrics such as modality contribution scores and saliency-based analyses can provide more detailed insights into attention behavior [11], [27], [31], [43].

Further research is also needed to improve robustness under real-world conditions, including noisy inputs, missing modalities, and dataset variability [15], [42], [43]. In addition, exploring computationally efficient attention mechanisms and lightweight architectures may help address scalability challenges.

Expanding dataset diversity and incorporating more realistic evaluation scenarios will also be important for improving generalizability. These directions are essential for ensuring that advances in multimodal attention mechanisms translate into reliable and practical systems.

8. Declarations

8.1. Author Contributions

Rexcharles Enyinna Donatus: Conceptualization, methodology, software, validation, Formal analysis, Resources, data curation, writing—original draft preparation, writing—review and editing, supervision; **Oludele Awodele:** conceptualization, formal analysis, writing—review and editing, supervision; **Osondu Everestus Oguike:** conceptualization, methodology, formal analysis, supervision; **Amina Sambo-Magaji:** methodology, investigation, writing—review and editing, visualization.

8.2. Institutional Review Board Statement

Not applicable.

8.3. Informed Consent Statement

Not applicable, as this study did not involve human participants.

8.4. Data Availability Statement

This study is based on previously published literature. All data supporting the findings of this review are available within the cited references.

8.5. Acknowledgment

The authors gratefully acknowledge the Africa Centre of Excellence on Technology Enhanced Learning, National Open University of Nigeria, Abuja, Nigeria, for their institutional support.

8.6. Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

9. References

[1] S. Li and H. Tang, "Multimodal alignment and fusion: A survey," *arXiv Prepr. arXiv2411.17040*, 2024. https://doi.org/10.48550/arXiv.2411.17040.

[2] M. A. Manzoor, S. Albarri, Z. Xian, Z. Meng, P. Nakov, and S. Liang, "Multimodality representation learning: A survey on evolution, pretraining and its applications," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 20, no. 3, pp. 1–34, 2023. https://doi.org/10.1145/3617833.

[3] Y. Yuan, Z. Li, and B. Zhao, "A survey of multimodal learning: Methods, applications, and future," *ACM Comput. Surv.*, vol. 57, no. 7, pp. 1–34, 2025. https://doi.org/10.1145/3713070.

[4] F. Zhao, C. Zhang, and B. Geng, "Deep multimodal data fusion," *ACM Comput. Surv.*, vol. 56, no. 9, pp. 1–36, 2024. https://doi.org/10.1145/3649447.

[5] N. Che, Y. Zhu, H. Wang, X. Zeng, and Q. Du, "AFT-SAM: adaptive fusion transformer with a sparse attention mechanism for Audio–Visual speech Recognition," *Appl. Sci.*, vol. 15, no. 1, p. 199, 2024. https://doi.org/10.3390/app15010199.

[6] 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. https://dl.acm.org/doi/abs/10.5555/3540261.3541349.

[7] J. Dhanith, S. Venkatraman, V. Sharma, S. Malarvannan, and M. Narendra, "Multimodal Emotion Recognition using Audio-Video Transformer Fusion with Cross Attention," *arXiv Prepr. arXiv.2407.18552*, 2024. https://doi.org/10.48550/arXiv.2407.18552.

[8] S. A. Abdu, A. H. Yousef, and A. Salem, "Multimodal video sentiment analysis using deep learning approaches, a survey," *Inf. Fusion*, vol. 76, pp. 204–226, 2021. https://doi.org/10.1016/j.inffus.2021.06.003.

[9] S. Mai, Y. Sun, A. Xiong, Y. Zeng, and H. Hu, "Multimodal boosting: Addressing noisy modalities and identifying modality contribution," *IEEE Trans. Multimed.*, vol. 26, pp. 3018–3033, 2023. https://doi.org/10.1109/TMM.2023.3306489.

[10] R. E. Donatus, U. O. Chiedu, and I. H. Donatus, "Exploring the Impact of Convolutional Neural Networks on Facial Emotion Detection and Recognition," *Asian J. Electr. Sci.*, vol. 13, no. 1, pp. 35–45, 2024. https://doi.org/10.70112/ajes-2024.13.1.4241.

[11] 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. https://doi.org/10.1016/j.imavis.2023.104676.

[12] H. Han, Q. Zheng, M. Luo, K. Miao, F. Tian, and Y. Chen, "Noise-tolerant learning for audio-visual action recognition," *IEEE Trans. Multimed.*, vol. 26, pp. 7761–7774, 2024. https://doi.org/10.1109/TMM.2024.3371220.

[13] 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. https://doi.org/10.21437/avsp.2019-5.

[14] 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. https://doi.org/10.1109/TPAMI.2018.2798607.

[15] 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. https://doi.org/10.3390/math13071100.

[16] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, "Balanced multimodal learning via on-the-fly gradient modulation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8238–8247. https://doi.org/10.1109/cvpr52688.2022.00806.

[17] J. Fu, J. Gao, B.-K. Bao, and C. Xu, "Multimodal imbalance-aware gradient modulation for weakly-supervised audio-visual video parsing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 6, pp. 4843–4856, 2023. https://doi.org/10.1109/TCSVT.2023.3337134.

[18] R. E. Donatus, "Interpretable Speech Emotion Recognition: A Comparative Study of BiLSTM Temporal Attention and Transformer-Based," *Asian J. Electr. Sci.*, vol. 14, no. 2, pp. 21–27, 2025. https://doi.org/10.70112/ajes-2025.14.2.4286.

[19] 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. https://doi.org/10.1109/TMM.2020.3046855.

[20] X. Jiang, X. Bai, and L. Yin, "The Latest Research Progress of Attention Mechanism in Deep Learning," pp. 82–89, 2025. https://doi.org/10.26689/jera.v9i3.10597.

[21] K. Bayoudh, "A survey of multimodal hybrid deep learning for computer vision: Architectures, applications, trends, and challenges," *Inf. Fusion*, vol. 105, p. 102217, 2024. https://doi.org/10.1016/j.inffus.2023.102217.

[22] S. Kalamkar and G. M. Amalanathan, "MDA-ViT: Multimodal image fusion using dual attention vision transformer," *Multimed. Tools Appl.*, vol. 84, no. 21, pp. 23701–23723, 2025. https://doi.org/10.1007/s11042-024-19968-1.

[23] 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. https://doi.org/10.1007/s10462-022-10148-x.

[24] 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. https://doi.org/10.1145/3664815.

[25] R. G. Praveen, P. Cardinal, and E. Granger, "Audio–visual fusion for emotion recognition in the valence–arousal space using joint cross-attention," *IEEE Trans. Biometrics, Behav. Identity Sci.*, vol. 5, no. 3, pp. 360–373, 2023. https://doi.org/10.1109/TBIOM.2022.3233083.

[26] J. Li, Y. Wu, Y. Qian, and C. Li, "Unified cross-modal attention: robust audio-visual speech recognition and beyond," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 1941–1953, 2024. https://doi.org/10.1109/TASLP.2024.3375641.

[27] 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. https://doi.org/10.1109/TKDE.2021.3126456.

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

[29] H. Kumar and M. Aruldoss, "Advanced optimal cross-modal fusion mechanism for audio-video based artificial emotion recognition," *Informatica*, vol. 49, no. 12, 2025. https://doi.org/10.31449/inf.v49i12.7392.

[30] R. G. Praveen, E. Granger, and P. Cardinal, "Cross attentional audio-visual fusion for dimensional emotion recognition," *16th IEEE Int. Conf. Autom. face gesture Recognit.*, pp. 1–8, 2021. https://doi.org/10.1109/FG52635.2021.9667055.

[31] 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. https://doi.org/10.1109/TIFS.2025.3586484.

[32] 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. https://doi.org/10.1109/ACCESS.2024.3360929.

[33] 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," *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 3, pp. 1903–1917, 2023. https://doi.org/10.1007/s12652-021-03407-2.

[34] 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. https://doi.org/10.1109/ICPR56361.2022.9956730.

[35] J. Vibell, A. Lim, and S. Sinnett, "Temporal perception and attention in trained musicians," *Music Percept. An Interdiscip. J.*, vol. 38, no. 3, pp. 293–312, 2021. https://doi.org/10.1525/mp.2021.38.3.293.

[36] M. Brousmiche, J. Rouat, and S. Dupont, "Multimodal attentive fusion network for audio-visual event recognition," *Inf. Fusion*, vol. 85, pp. 52–59, 2022. https://doi.org/10.1016/j.inffus.2022.03.001.

[37] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Multimodal Transformer Fusion for Continuous Emotion Recognition" *ICASSP 2020 - 2020 IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 3507–3511, 2020. https://doi.org/10.1109/ICASSP40776.2020.9053762.

[38] 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. https://doi.org/10.1038/s41598-025-09000-3.

[39] 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. https://doi.org/10.3390/app10207263.

[40] 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. https://doi.org/10.1016/j.inffus.2024.102417.

[41] W. Song, S. Ren, and B. Hu, "Interpretable Learning Method Based on Causal Interactive Attention," *IEEE Access*, vol. 13, 2025. https://doi.org/10.1109/ACCESS.2025.3583583.

[42] 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. https://doi.org/10.1007/s11042-022-13557-w.

[43] 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. https://doi.org/10.48550/arXiv.2502.12488.

[44] 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. https://doi.org/10.1145/3746027.3761982.

[45] 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. https://doi.org/10.1016/j.knosys.2024.112925.

[46] 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. https://doi.org/10.1016/j.inffus.2023.102218.

[47] 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. https://doi.org/10.3390/rs13152965.

[48] 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. https://doi.org/10.1016/j.knosys.2022.108580.

[49] 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. https://doi.org/10.1109/ICISCN64258.2025.10934189.

[50] 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. https://doi.org/10.1145/3721981.

[51] G. Sun *et al.*, "Fine-grained audio-visual joint representations for multimodal large language models," *arXiv Prepr. arXiv2310.05863*, 2023. https://doi.org/10.48550/arXiv.2310.05863.

[52] J. Li and Y. Tian, "From waveforms to pixels: A survey on audio-visual segmentation," *arXiv Prepr. arXiv2508.03724*, 2025. https://doi.org/10.48550/arXiv.2508.03724.

[53] 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. https://doi.org/10.3390/app13137431.

[54] 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. https://doi.org/10.1109/TASLP.2021.3096037.

[55] 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. https://doi.org/10.18653/v1/2023.acl-long.223.

[56] 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. https://doi.org/10.22266/ijies2025.0531.03.

[57] D. Li *et al.*, "Emotion recognition of subjects with hearing impairment based on fusion of facial expression and EEG topographic map," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 437–445, 2022. https://doi.org/10.1109/TNSRE.2022.3225948.

[58] N. Seijdel, J.-M. Schoffelen, P. Hagoort, and L. Drijvers, "Attention drives visual processing and audiovisual integration during multimodal communication," *J. Neurosci.*, vol. 44, no. 10, 2024. https://doi.org/10.1523/JNEUROSCI.0870-23.2023.

[59] 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. https://doi.org/10.1016/j.imavis.2020.104042.

[60] 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. https://doi.org/10.1016/j.treng.2024.100228.

[61] 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. https://doi.org/10.3390/app15115823.

[62] 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. https://doi.org/10.1109/TASLP.2021.3066303.

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

[64] 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. https://doi.org/10.3390/app14178071.

[65] 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. https://doi.org/10.3390/biomimetics10070418.