A Survey of Multimodal AI and Sensory Integration Techniques

www.surveyx.cn

Abstract

Multimodal AI represents a significant leap in artificial intelligence, enabling systems to process and integrate information from multiple sensory modalities, such as vision, sound, and text, through advanced computational techniques. This survey examines the evolution, current state, and future directions of multimodal AI, focusing on its applications in emotion recognition, cross-modal learning, and the integration of heterogeneous data sources. The survey highlights the pivotal role of large language models (LLMs) in enhancing multimodal AI capabilities, particularly in sentiment analysis and human-computer interaction. Key challenges include optimizing computational efficiency, improving alignment and fusion processes, and developing comprehensive evaluation metrics. Innovative approaches such as dynamic fusion techniques, attention mechanisms, and neural integration strategies are explored to address these challenges. The survey also underscores the transformative impact of multimodal AI in fields such as medical imaging, where it enhances diagnostic accuracy by integrating diverse data modalities. Future research directions emphasize the need for robust benchmarking, improved data quality, and scalable integration techniques to optimize the performance and applicability of multimodal systems. Overall, the findings underscore the potential of multimodal AI to drive advancements across various domains, paving the way for more sophisticated and adaptable AI solutions capable of seamlessly processing complex sensory inputs.

1 Introduction

1.1 Significance of Multimodal AI

Multimodal AI signifies a transformative leap in technology, significantly enhancing the processing and interpretation of complex data by integrating various sensory modalities. This capability is essential for sentiment analysis, where the combination of visual, verbal, and acoustic inputs from videos allows for a deeper understanding of human emotions [1]. Such integration is particularly vital in applications like human-computer interaction and virtual reality, where aligning lip movements with audio is crucial for creating realistic experiences [2].

In the rapidly evolving landscape of video applications, multimodal AI facilitates a comprehensive understanding by analyzing and synthesizing information from diverse sensory inputs, thereby addressing the urgent need for automated video content comprehension [3]. Its potential is further exemplified in video categorization, where leveraging both video and audio modalities significantly boosts classification accuracy [4].

Moreover, multimodal AI's ability to generate text from video and other multimodal sources underscores its advancement beyond previous methods reliant on specialized cross-modal fusion modules [5]. This progress is critical for developing robust systems capable of managing a wider array of tasks beyond traditional text and vision modalities, as highlighted in recent surveys of generalist multimodal models [6].

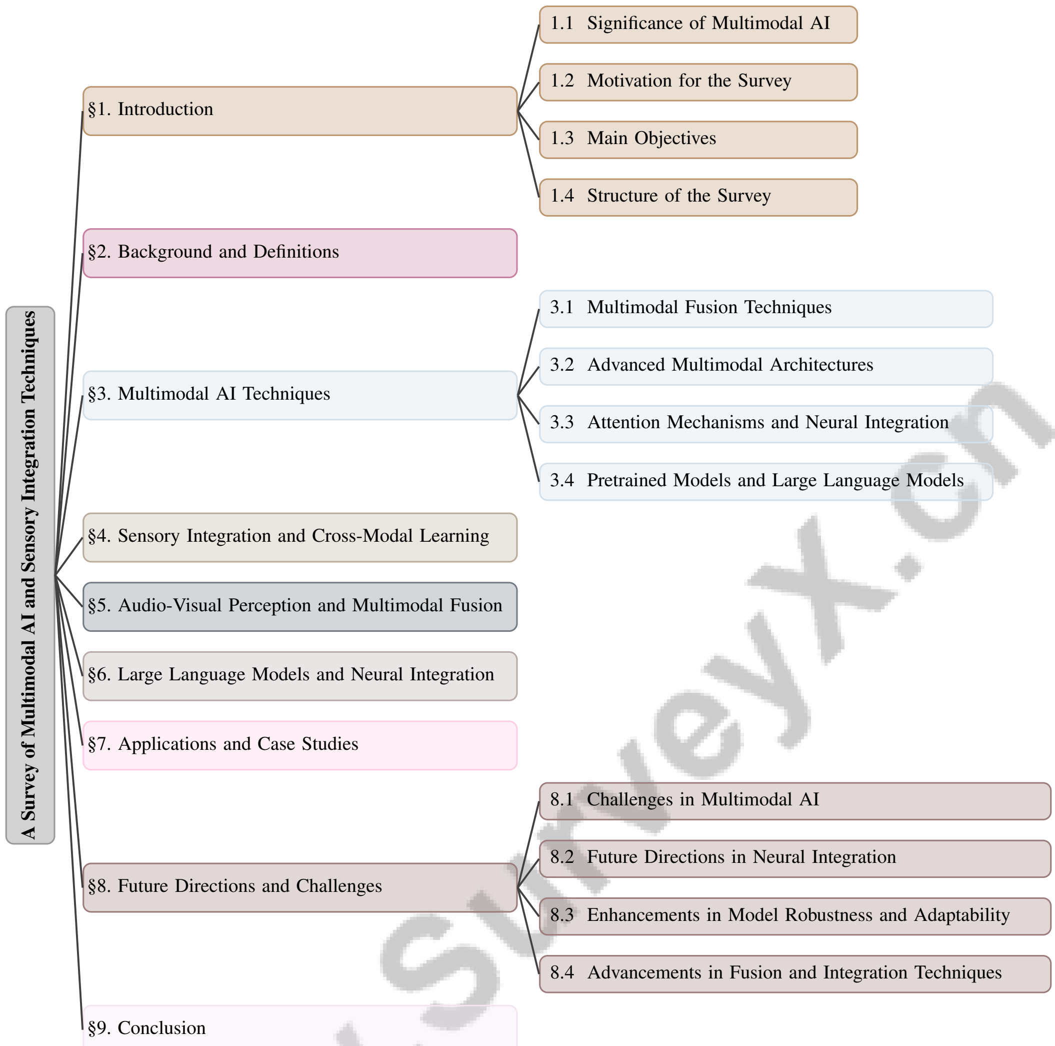


Figure 1: chapter structure

The significance of multimodal AI extends to enhancing human decision-making processes and providing nuanced insights into complex real-world scenarios. Its application in fields like education demonstrates how AI can facilitate transparent tutoring evaluations through multimodal data analysis, addressing multifaceted tasks that integrate visual and textual information. By harnessing multimodal intelligence, AI improves accuracy and effectiveness in decision-making, ultimately driving innovation across diverse applications [7, 8].

1.2 Motivation for the Survey

This survey is motivated by the need to address significant gaps in the literature surrounding multimodal AI and sensory integration. A notable challenge is the absence of effective visual attention prediction models that incorporate audio cues, which are crucial for influencing human attention in videos [9]. The survey aims to investigate methodologies that integrate audio-visual cues to enhance attention prediction models.

Additionally, the operational and computational costs associated with multimodal fusion techniques, particularly in industrial contexts, remain insufficiently explored in existing literature [10]. This survey seeks to evaluate and propose more efficient fusion strategies that reduce computational burdens while preserving performance.

AI-generated, for reference only.

In sentiment analysis, the asynchronous nature of multimodal sequences complicates the alignment and fusion process, posing a significant challenge [1]. The survey will explore innovative approaches for effectively analyzing and integrating unaligned multimodal data.

Furthermore, generating coherent and contextually relevant text from multimodal inputs is an area where current research is lacking [5]. This survey intends to address these deficiencies by reviewing methodologies that enhance text generation capabilities from diverse modalities.

In the medical domain, there is a pressing need for integrating various medical data modalities to improve clinical decision-making [11]. This survey will explore deep learning applications that facilitate this integration, advancing medical AI applications.

Developing comprehensive models capable of simultaneously understanding video, image, and language is crucial for applications such as video surveillance and content retrieval [3]. This survey aims to explore models that achieve multi-modal comprehension.

Moreover, reducing the computational cost of fine-tuning large multimodal models for downstream tasks is an urgent concern [12]. The survey will identify strategies that optimize resource utilization while maintaining model efficacy.

Existing video categorization methods face limitations, necessitating cross-modal learning approaches that leverage inter-modal correlations [4]. This survey will investigate these approaches to enhance categorization accuracy.

While covering a broad spectrum of modalities, this survey deliberately excludes highly specialized or niche applications that do not contribute to the general understanding of multimodal AI [6]. By addressing these diverse challenges, the survey aims to provide a comprehensive overview of the current state of multimodal AI, offering insights into potential solutions and future research directions.

1.3 Main Objectives

The primary objectives of this survey are to systematically review and synthesize advancements in multimodal AI, focusing on emotion recognition, cross-modal learning, and the integration of heterogeneous data sources. This survey seeks to investigate frameworks that enhance the integration of multimodal signals—such as audio, visual, and textual data—significantly improving the accuracy and robustness of emotion recognition systems. By examining state-of-the-art multimodal fusion techniques and their applications, this research aims to address challenges in affective computing and sentiment analysis, ultimately contributing to the development of more effective emotion recognition technologies [13, 14, 15]. It will also explore innovative methods such as TelME, which employs cross-modal knowledge distillation to bolster weaker modalities in emotion recognition tasks, enhancing overall performance.

A significant goal is to assess frameworks like Spider, which facilitates Any-to-Many Modalities Generation (AMMG). This framework enables the seamless integration and generation of diverse combinations of modalities—text, images, audio, and video—into a single cohesive output. Spider achieves this through a Base Model for basic modality processing, an Efficient Decoders-Controller for managing multimodal content generation, and a specialized Any-to-Many Instruction Template for creating signal prompts. By constructing a novel Text-formatted Many-Modal (TMM) dataset, Spider enhances its capability to produce arbitrary multimodal outputs, advancing the field of multimodal interaction and providing valuable data support for future research initiatives [16, 17]. The survey will also investigate machine learning frameworks that predict resource requirements and adjust allocations to optimize performance in multimodal systems.

This survey aims to deliver a thorough examination of multimodal intelligence, focusing on the integration of vision and natural language through advanced models and learning methods. By analyzing recent developments in multimodal representation learning, it will explore key concepts such as embedding techniques that unify diverse modalities into a single vector space, innovative architectures for signal fusion, and various applications including image-to-text caption generation, text-to-image generation, and visual question answering. This comprehensive analysis seeks to propel advancements in multimodal representation learning and enhance understanding of the complex interplay between different modalities in AI [18, 19, 20, 8, 21]. Additionally, methods like Text-centric Alignment for Multi-Modality Supervised Learning (TAMML) will be examined to improve the generalizability of multimodal systems under diverse conditions.

In real-world applications, the survey seeks to highlight the development of multimodal fusion models utilizing high-level audio and visual features for sentiment analysis, enhancing their deployability in practical scenarios. It will evaluate holistic frameworks such as the Holistic AI in Medicine (HAIM), which leverage multimodal inputs to generate and test AI systems in medical applications [11].

The survey will also explore the integration of insights from Conversation Analysis (CA) and Theory of Mind (ToM) to enhance human-machine interactions. By analyzing the linguistic, organizational, and structural elements of multimodal communication, the research aims to inform the development of AI systems that are more intuitive and capable of understanding social contexts and managing conversational flows effectively. This approach is expected to contribute to creating AI that supports human decision-making processes, particularly in educational settings, by leveraging multimodal data to improve interaction quality and decision transparency [7, 22]. It will also address the Semantic-centric Multimodal Affective Computing (SemanticMAC) framework, which enhances the learning process of multimodal representations, while tackling challenges in efficiently accessing, manipulating, and querying entities within neural network models processing multimodal data.

The survey aims to showcase enhancements in the multimodal capabilities of large language models, particularly through advanced data fusion techniques and training procedures. It will provide an extensive analysis of multimodal machine learning, highlighting challenges that extend beyond traditional fusion methods, and examining the implications of bias and fairness in multimodal fusion techniques within automated recruitment systems, utilizing the FairCVdb dataset. The effectiveness of early-fusion versus late-fusion methods in achieving accurate and equitable outcomes will be addressed, alongside proposing future research avenues to enhance fairness through alternative fusion strategies and modality-specific constraints [23, 8]. Finally, it will fill knowledge gaps regarding the taxonomy of Multimodal Sentiment Analysis (MSA) methods and recent advancements in multimodal fusion architectures, exploring the Adaptive Language-guided Multimodal Transformer (ALMT) to improve MSA effectiveness by filtering out irrelevant information.

1.4 Structure of the Survey

This survey is organized to provide a comprehensive examination of multimodal AI and sensory integration techniques, structured into distinct sections that build upon one another. The initial section introduces the topic, emphasizing the significance of multimodal AI in processing information from various sensory modalities, while outlining the motivation and objectives of the survey. Following this, the background and definitions section offers foundational explanations of key concepts such as multimodal AI, sensory integration, and related terminologies, alongside discussing their historical development and evolution.

The survey then delves into specific multimodal AI techniques, exploring methods and frameworks that enable the integration and processing of multiple sensory inputs, with a focus on large-scale neural networks and learning models. The subsequent section provides an in-depth analysis of sensory integration and cross-modal learning, highlighting specific principles and methodologies that improve decision-making and enhance user interaction in AI systems. This exploration is particularly relevant in the context of multimodal intelligence, where combining various data types—such as visual and textual information—can significantly augment human decision-making processes, as evidenced by empirical studies in educational settings. By leveraging advanced classification models that utilize multimodal data, the research demonstrates how AI can assist educators in making more transparent and accurate evaluations, thereby fostering a more effective learning environment [7, 8].

Subsequent sections analyze audio-visual perception and multimodal fusion, addressing the challenges and innovative techniques for effective data fusion. The impact of large language models on multimodal AI is discussed, alongside neural integration techniques that enhance AI capabilities.

Real-world applications and case studies illustrate the practical implementation and benefits of multimodal AI techniques across various fields. The survey concludes with a discussion of future directions and challenges, identifying emerging trends and areas for further research and development in the field. Throughout the survey, a novel taxonomy is introduced to categorize challenges in multimodal machine learning, focusing on representation, translation, alignment, fusion, and co-learning [21].The following sections are organized as shown in Figure 1.

AI-generated, for reference only.

2 Background and Definitions

2.1 Definitions of Multimodal AI and Sensory Integration

Multimodal AI is designed to integrate and process information across various sensory modalities, such as text, audio, video, and skeletal data, enhancing interaction and recognition capabilities by addressing semantic ambiguities and data heterogeneity [24]. A core aim is to merge unimodal representations into a unified framework, crucial for advancing recognition tasks [25]. Challenges include unaligned data and the need to capture long-range cross-modal dependencies, necessitating sophisticated alignment techniques [1].

Sensory integration in AI unifies diverse sensory inputs, enhancing perception and decision-making. In medical applications, integrating modalities like MRI and CT scans improves diagnostic accuracy, underscoring the importance of effective multimodal data fusion [11]. However, issues like subsampling and noise complicate data fusion, potentially leading to incomplete or unreliable information [26]. Innovations such as the Component Attention Network (CANet) address these challenges, exemplified in tasks like dance recognition [27].

Cross-modal learning, a subset of multimodal AI, leverages intermediate outputs from one modality to inform others, based on their correlations [4]. This is pivotal in applications like music transcription and speaker tracking, though it faces challenges from adversarial samples that can mislead models [28]. Despite these hurdles, advancements in multimodal AI enhance systems' cognitive emulation, as seen in applications from lip-synchronized talking head generation [2] to generalist multimodal models across various data modalities [6]. Effective use of embedding spaces is essential for distinguishing channel importance and capturing fine-grained event details [29].

2.2 Historical Development and Evolution

The evolution of multimodal AI and sensory integration marks a shift from unimodal to integrated systems, driven by the need to process complex, diverse sensory inputs. Initially, AI systems focused on individual sensory inputs, limiting their efficacy in complex tasks [14]. Early fusion methods, such as decision-level or feature-level fusion, often suffered from data loss and noise, impeding system effectiveness [26].

As the field progressed, the limitations of unimodal approaches became apparent, leading to the development of sophisticated multimodal systems that integrate audio, visual, and textual data into cohesive frameworks [24]. This integration required innovative fusion methods like tensor representations, which, despite their power, introduced significant computational costs and complexity with increasing modalities [25].

Advancements in multimodal AI have explored various fusion techniques, including early, late, and hybrid fusion, as well as tensor and hierarchical fusion, aiming to bridge the heterogeneity gap between modalities [21]. Despite these efforts, existing benchmarks have not thoroughly evaluated the integration of different modalities throughout the scientific process, leaving gaps in understanding the limitations of current models [30].

In healthcare, the evolution of multimodal AI has seen the emergence of advanced models using deep neural networks for medical tasks, highlighting the significance of effective data fusion for diagnostic accuracy [11]. However, challenges such as limited multimodal datasets and complex modality alignment persist, requiring robust architectural designs to address diverse tasks [6].

Recent advancements focus on improving multimodal alignment in representation learning. Earlier methods, like the Dual Cross-modal Information Disentanglement (DCID) model, were limited by treating all channels equally, overlooking minor event information [29]. The ongoing evolution of multimodal AI continues to enhance the field's ability to tackle these challenges, paving the way for more robust and adaptable AI systems.

In recent years, the field of artificial intelligence (AI) has witnessed significant advancements, particularly in the realm of multimodal techniques. These innovations have paved the way for more effective integration of diverse sensory inputs, thereby enhancing the overall capabilities of AI systems. To elucidate the complexities of these advancements, Figure 2 illustrates the hierarchical categorization of multimodal AI techniques. This figure details various categories, including fusion models, advanced architectures, and attention mechanisms, while also emphasizing the pivotal role of pretrained models and large language models. Each category not only highlights specific methods but also showcases their applications, ultimately contributing to a more nuanced understanding of how AI can process and integrate information from multiple modalities.

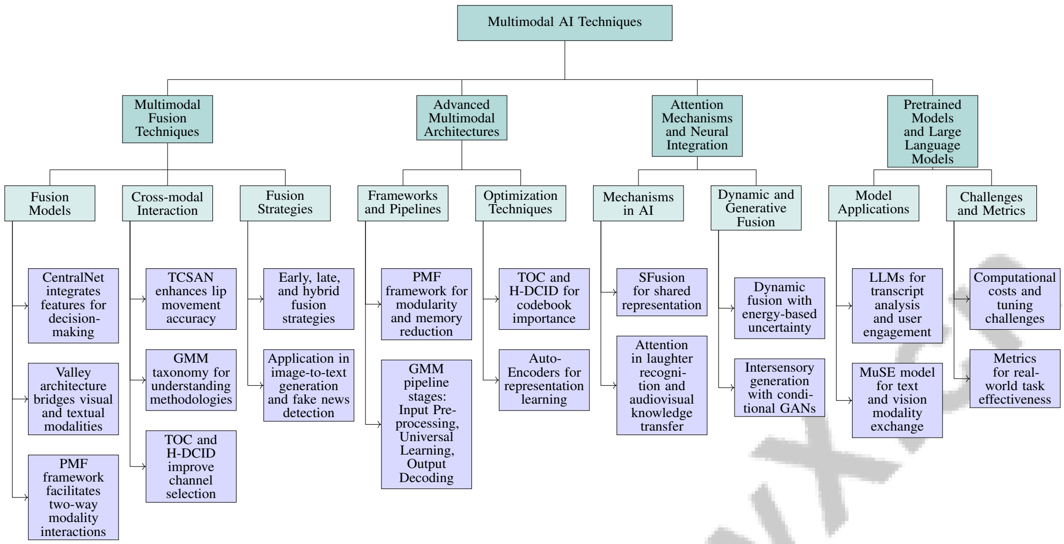


Figure 2: This figure illustrates the hierarchical categorization of multimodal AI techniques, detailing fusion models, advanced architectures, attention mechanisms, and the role of pretrained models and large language models. Each category highlights specific methods and applications that contribute to enhancing AI's ability to process and integrate diverse sensory inputs.

3 Multimodal AI Techniques

3.1 Multimodal Fusion Techniques

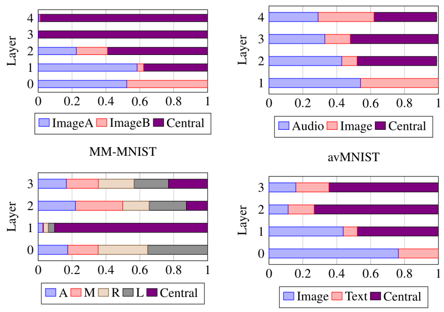
Multimodal fusion techniques synthesize information from varied sensory modalities like text, audio, and visual data, enhancing AI performance in complex tasks. These methods leverage modality complementarities to develop robust AI models. CentralNet exemplifies this by integrating features through a central network for independent and integrated decision-making [31]. The Valley architecture bridges visual and textual modalities via a temporal modeling module and projection layer, distinguishing it from task-specific models [3]. The PMF framework facilitates two-way modality interactions using interactive prompts for optimized learning [12].

The Dilated Non-Causal Temporal Convolutional Self-Attention Network (TCSAN) enhances lip movement accuracy by fusing multimodal features from audio and facial action units, illustrating cross-modal interaction benefits [2]. A survey on Generalist Multimodal Models (GMMs) introduces a taxonomy based on Unifiability, Modularity, and Adaptability, aiding in understanding current methodologies [6]. Cross-modal learning techniques, like those by Goyal et al., train a correlation tower to assess modality relevance and transform attention vectors for cross-modal influence [4]. The Training-free Optimization of Codebook (TOC) and Hierarchical Dual Cross-modal Information Disentanglement (H-DCID) methods improve channel selection and information alignment across modalities [29].

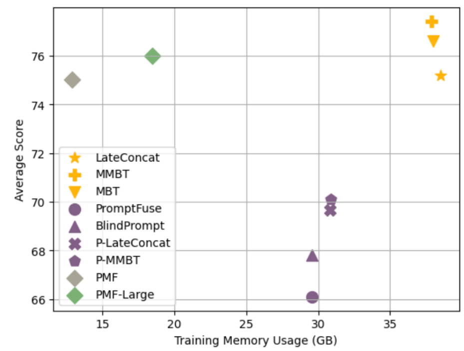
Fusion strategies—early, late, and hybrid—highlight multimodal fusion's role in enhancing AI systems. Early fusion leverages unique modality characteristics for accuracy, while late fusion yields generalized outcomes, emphasizing the need for appropriate technique selection to optimize performance in applications like image-to-text generation and fake news detection [18, 32, 23, 8].

Figure 3 illustrates the importance of visualizations in optimizing multimodal AI techniques. The "Visualization of Attention Distribution Across Layers and Tasks for Different Models" provides insights into attention mechanism variability, while the "Comparison of Average Scores and Training Memory Usage for Different Models" aids in evaluating model efficiency and performance [31, 12].

AI-generated, for reference only.



(a) Visualization of Attention Distribution Across Layers and Tasks for Different Models[31]



(b) Comparison of Average Scores and Training Memory Usage for Different Models[12]

Figure 3: Examples of Multimodal Fusion Techniques

3.2 Advanced Multimodal Architectures

Advanced multimodal architectures optimize sensory input integration, enhancing AI performance across tasks. The PMF framework exemplifies modularity, enabling parallel processing of unimodal transformers with deep-layer prompts, reducing memory usage [12]. The GMM pipeline structures multimodal AI into Input Pre-processing, Universal Learning, and Output Decoding stages, ensuring robust and adaptable AI systems [6].

TOC and H-DCID optimize architectures by enhancing codebook importance calculation and facilitating integration through hierarchical structures [29]. These techniques highlight efficient channel utilization and hierarchical structuring's significance in multimodal learning. Advanced architectures improve representation learning and vision models' capabilities in comprehending complex, text-heavy visual content. Auto-Encoders and fine-tuning with instructional data demonstrate potential for broader applications, achieving high accuracy in interconnected textual and visual tasks [20, 33].

3.3 Attention Mechanisms and Neural Integration

Attention mechanisms are pivotal in multimodal AI, enabling dynamic prioritization and synthesis of diverse inputs. These mechanisms enhance decision-making and performance in complex tasks. SFusion, a self-attention based N-to-One fusion block, exemplifies this by learning to fuse modalities into a shared representation [34]. In laughter recognition, attention mechanisms improve accuracy by focusing on specific audio and video features [35]. Audiovisual cross-modality knowledge transfer benefits from attention-enhanced feature integration via semantic alignment [10].

Dynamic fusion mechanisms, like Zhang's, use energy-based uncertainty estimation to adaptively weigh modalities, ensuring focus on higher-quality inputs [36]. Multiscale Cooperative Multimodal Transformers (MCMulT) capture and synchronize diverse information through multiscale representations, enhancing learned representation quality [1]. In scenarios with missing or noisy data, generative model-driven recovery and fusion, as demonstrated by SFLR, maintain performance [26].

Attention mechanisms also facilitate intersensory generation, as seen in conditional GANs generating images from sounds and vice versa [37]. This intersensory generation highlights GANs' potential to bridge sensory modalities, enhancing system coherence and functionality. In AI chatbots, attention mechanisms enable image-text comprehension, enhancing neural integration and modality communication [38]. Recent studies emphasize attention mechanisms' critical role in optimizing diverse modality integration, improving performance in complex applications like image-to-text caption generation and visual question answering [18, 33, 8].

3.4 Pretrained Models and Large Language Models

Pretrained models and large language models (LLMs) advance multimodal AI by providing frameworks for diverse modality processing and generation. Trained on extensive datasets, these models encapsulate intricate patterns for fine-tuning in specific applications. LLMs enhance multimodal

AI-generated, for reference only.

transcript analysis, improving human behavior understanding [39]. Integrating LLMs into mobile platforms enhances multimodal interactions, supporting diverse user engagements [40]. The MuSE model, using a CrossTransformer architecture, exemplifies text and vision modality knowledge exchange, facilitating effective fusion [41].

In video comprehension, the Valley architecture utilizes LLMs with a temporal modeling module, improving instruction-following and video content comprehension [3]. The VX2TEXT model employs an autoregressive decoder for open-ended text generation, offering flexibility in generating coherent text from multimodal inputs [5]. Despite advancements, computational costs and effective tuning challenges remain, necessitating optimization for LLM efficiency and scalability in multimodal applications. Metrics reflecting real-world task effectiveness emphasize accuracy and multimodal information integration [30].

Integrating pretrained models and LLMs enhances systems' capacity to process complex sensory inputs—text, images, and audio—through advanced data fusion techniques and specialized architectures. This advancement improves performance metrics, such as accuracy and F1 scores, enabling coherent and contextually relevant output generation, paving the way for sophisticated AI solutions reflecting human cognitive understanding [42, 8].

4 Sensory Integration and Cross-Modal Learning

4.1 Principles and Methods of Sensory Integration

Sensory integration is crucial for AI systems to synthesize information from multiple modalities, enhancing perception and decision-making. Selecting the optimal fusion strategy—early, late, or hybrid—is essential to ensure complementary contributions from each modality, particularly in systems like laughter recognition that integrate audio and video cues [35]. Multiscale Cooperative Multimodal Transformers (MCMulT) leverage multi-scale mechanisms to capture diverse semantic features through cooperative interactions, allowing dynamic adjustments in modality contributions and optimizing integration [1].

Cross-modal learning principles assert that correlated modalities enhance understanding, as demonstrated by methods like the MuSE model, which regularizes embeddings into a common space for improved multimodal task performance [4, 41]. Maintaining semantic consistency amid diverse data complexities is vital, as exemplified by the VX2TEXT model, which integrates inputs into a unified semantic space for coherent text generation [5].

Temporal modeling is critical in sensory integration, as illustrated by the TCSAN method, which considers past and future information to model temporal relationships in lip movements [2]. Effective channel selection and alignment, highlighted by Huang et al., are necessary for enhanced multimodal representation learning [29].

Integrating diverse sensory inputs through multimodal intelligence allows AI systems to replicate human-like perception and cognition, effectively processing various modalities such as vision, language, and auditory signals, thus enhancing decision-making in complex applications [7, 21, 18, 8].

4.2 Cross-Modal Learning Frameworks

Cross-modal learning frameworks are pivotal in advancing AI by integrating information from various modalities, enhancing decision-making and interaction capabilities. Adaptive audiovisual saliency fusion methods offer a nuanced approach compared to traditional models, improving audio and visual data integration [9]. The AV Align framework aligns audio and visual speech representations at the frame level, enhancing tasks like lip-reading and audiovisual speech recognition [43].

To address diverse data integration challenges, frameworks like MultiFIX enhance interpretability and effectiveness, exemplifying cross-modal learning's potential to improve AI transparency [44]. CentralNet's multilayer approach emphasizes adaptability in fusion strategies, suggesting future research directions for broader multimodal task applications [45].

Future research could explore data augmentation and class-balancing strategies to enhance model performance and reduce false positives, crucial for robust AI systems in real-world applications [46]. Cross-modal frameworks are fundamental for advancing multimodal AI, integrating diverse inputs into cohesive models, and supporting sophisticated applications like image-to-text caption generation and visual question answering. By leveraging techniques like embedding representations and modality-specific networks, these frameworks address challenges such as modality mismatch and scalability, paving the way for robust AI solutions [18, 28, 8, 47, 21].

5 Audio-Visual Perception and Multimodal Fusion

5.1 Challenges in Audio-Visual Perception

Audio-visual perception poses significant challenges in AI systems, complicating the integration and interpretation of multimodal data. A primary issue is the exponential rise in computational and memory demands associated with tensor representations, especially when handling more than two modalities [48]. This restricts the deployment of sophisticated models necessary for robust integration. Inefficiencies in selecting pre-trained feature extraction networks further exacerbate this problem, often leading to suboptimal performance and increased processing time [49]. Efficient feature extraction methods are crucial for streamlining audio and visual data fusion.

Moreover, current approaches often lack mechanisms to integrate audio-visual information across multiple temporal scales, essential for navigating complex acoustic environments [50]. This is particularly detrimental in dynamic scenarios where temporal synchronization is vital. Sound localization in non-line-of-sight scenarios presents another challenge, as AI agents struggle to correlate audio signals with visual environments [51], impairing accurate environmental interpretation.

Traditional fusion methods, such as simple concatenation, fail to capture intricate dependencies between modalities necessary for nuanced perception [52]. In robotic applications, challenges include recognizing and manipulating transparent or occluded objects [53], necessitating models that handle occlusions and transparency effectively. Discrepancies between visual and audio modalities can obscure sentiment analysis, complicating interpretation [54]. The limitations of unimodal data, like skeletal data, in capturing the full expressive range further complicate perception [27]. Integrating multiple modalities is essential for comprehensive understanding in complex tasks, such as dance recognition.

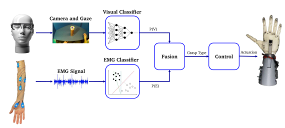
The AV Align framework exemplifies a promising approach, improving recognition accuracy by learning and exploiting audio-visual alignments [43]. These challenges highlight the need for innovative approaches to enhance audio-visual perception robustness and effectiveness in AI systems.

5.2 Innovative Fusion Techniques

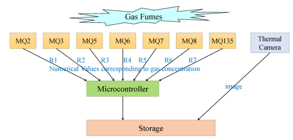
Innovative fusion techniques are critical for effective audio-visual integration, enhancing AI systems' performance and efficiency in processing multimodal data. The Multimodal Bottleneck Transformer (MBT) exemplifies advancements in this area, improving computational efficiency while maintaining high fusion performance [55]. MBT's architecture addresses audio-visual fusion complexities by using a shared bottleneck layer to process information from both modalities, enhancing cross-modal interactions and reducing data processing redundancy. This is particularly relevant in applications like automatic fake news detection, where accurate multimodal integration is crucial [18, 25, 33].

Attention-based techniques further enhance fusion by enabling models to integrate and process multi-dimensional inputs efficiently. Transformer-based architectures with fusion bottlenecks improve performance on benchmarks like Audioset, Epic-Kitchens, and VGGSound while reducing computational costs. Models like the Intra- and Inter-Attention Network (IIANet) utilize selective attention mechanisms to effectively separate audio-visual speech, outperforming previous methods in efficiency and processing time [50, 55, 56]. These techniques allow AI systems to dynamically focus on pertinent features from each modality, improving fusion accuracy and robustness.

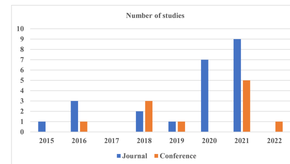
The integration of techniques like MBT and attention mechanisms underscores the importance of developing sophisticated models for handling multimodal data intricacies. Recent advancements in multimodal intelligence enhance AI systems' efficiency and effectiveness, enabling them to tackle tasks requiring seamless audio and visual input integration. This progress is evident in applications like image-to-text caption generation, text-to-image generation, and visual question answering, where deep learning unifies multimodal signals into cohesive representations. By leveraging specialized architectures and embedding methods, systems can process complex data sets, such as those in text-heavy visual content, expanding applications across domains like education, healthcare, and interactive technologies [18, 33, 8].



(a) A robotic arm and hand system for grasping objects[57]



(b) Gas Fumes Monitoring System[58]



(c) Number of Studies[59]

Figure 4: Examples of Innovative Fusion Techniques

As shown in Figure 4, innovative fusion techniques significantly enhance complex systems' efficiency and functionality in audio-visual perception and multimodal fusion. The figure illustrates three examples: a robotic arm and hand system for object grasping integrating a camera and gaze system with an EMG and visual classifier for adaptive mechanical tasks; a gas fumes monitoring system utilizing sensors and a microcontroller for effective environmental condition monitoring; and a bar chart depicting research growth in innovative fusion techniques from 2015 to 2022, highlighting increasing interest and efforts in this field. These examples demonstrate the transformative potential of combining multiple data sources and modalities to enhance system capabilities and decision-making processes [57, 58, 59].

6 Large Language Models and Neural Integration

The convergence of large language models (LLMs) and multimodal artificial intelligence (AI) represents a pivotal advancement in the field, driven by the transformative capabilities of these technologies. As LLMs progress, their integration into multimodal frameworks not only enhances the processing of diverse data types but also opens new research opportunities, particularly in multimodal sentiment analysis. This section delves into the impact of LLMs on multimodal AI, focusing on their performance enhancements and implications for future developments.

6.1 Impact of Large Language Models on Multimodal AI

LLMs play a crucial role in advancing multimodal AI systems by enhancing their ability to process and integrate varied data forms. Their integration into multimodal frameworks has propelled tasks such as sentiment analysis, as seen in the Multiscale Cooperative Multimodal Transformers (MCMulT), which excel in both aligned and unaligned sentiment analysis [1]. This underscores LLMs' potential to bolster AI systems' robustness and accuracy in complex data environments.

The Valley architecture exemplifies how LLMs enhance interaction capabilities and decoding performance, showcasing the synergy between visual and textual data processing [3]. In dynamic audiovisual contexts, the Cross-modal Attention Transformer (CAT) surpasses existing methods, demonstrating LLMs' effectiveness in boosting multimodal capabilities [60].

Moreover, the Adaptive Language-guided Multimodal Transformer (ALMT) achieves state-of-the-art results across multiple datasets in multimodal sentiment analysis, highlighting the impact of language-guided adaptive mechanisms in refining AI performance [54]. This approach exemplifies strategic LLM integration to enhance modality synergy, improving overall AI system performance.

Despite these advancements, challenges remain, particularly the computational demands of LLMs and their tuning for diverse tasks. Addressing these issues is vital for optimizing LLM efficiency and scalability in multimodal applications. Future research will focus on leveraging LLMs to further enhance multimodal AI frameworks, expanding their real-world applications. Ongoing investigations into innovative fusion techniques and advanced neural integration methods are poised to significantly advance adaptable AI solutions, especially in multimodal intelligence, where vision and natural language processing interplay is crucial for applications like image-to-text caption generation, text-to-image generation, and visual question answering [18, 33, 8].

6.2 Neural Integration Techniques

Neural integration techniques are essential for advancing multimodal AI systems, enabling effective processing and synthesis of information from various sensory modalities such as vision, speech, and text. By employing sophisticated models for multimodal representation learning and fusion, these techniques unify diverse input signals into a cohesive framework, enhancing the system's ability to perform complex tasks like image captioning, text-to-image generation, and visual question answering. This integration not only boosts application performance but also mirrors the complex processing of multisensory information in biological neural systems [18, 61, 8, 21, 62].

The MoExtend framework represents an innovative approach to neural integration by selectively adding new experts in response to data distribution shifts [63]. This method maintains the original model's integrity while expanding its capabilities, ensuring robustness and adaptability to changing input distributions. By dynamically incorporating new experts, MoExtend facilitates effective neural integration, allowing AI systems to sustain high performance amid evolving data landscapes.

Additionally, the integration of weighted Area Under the Curve (wAUC) metrics offers a nuanced evaluation of knowledge transfer (KT) models, emphasizing temporal performance [64]. This temporal focus is crucial for assessing neural integration techniques, highlighting the model's adaptability and performance improvement in response to diverse inputs. The wAUC metric serves as a valuable tool for refining neural integration strategies, ensuring efficient processing and learning from multimodal data streams.

Recent advancements in neural integration techniques underscore the critical need for developing adaptive and robust AI systems capable of seamlessly processing and integrating information from diverse modalities like text, images, and speech. This is particularly relevant as multimodal intelligence evolves, with significant research dedicated to fusing different data types to enhance applications in areas like image captioning, text-to-image generation, and visual question answering [18, 62, 20, 8]. By enhancing interactions between neural networks and multimodal inputs, these techniques pave the way for more sophisticated and effective AI solutions across a wide range of applications.

7 Applications and Case Studies

7.1 Real-World Applications and Case Studies

Multimodal AI has significantly enhanced real-world applications by integrating diverse data types to improve performance across various domains. The TelME framework exemplifies this by achieving state-of-the-art results in emotion recognition within multi-party conversations, effectively capturing complex emotional cues through multimodal fusion [65]. In concurrent speech detection, a novel audiovisual approach has demonstrated superior performance on challenging datasets like EasyCom, highlighting multimodal AI's ability to effectively integrate audio and visual data in complex auditory environments [66]. Additionally, the confirmation of double-loop feedback in multimodal interactions provides valuable insights into the statistical distributions of modalities and the dynamics of human-machine conversations [22].

Advanced models like LF-LSTM and MulT have proven effective in sentiment analysis by interpreting sentiments through diverse data sources, as shown in experiments on datasets such as CMU-MOSI and CMU-MOSEI [67]. The Deep Equilibrium Multimodal Fusion model further demonstrates versatility across tasks like BRCA classification and visual question answering, reflecting its robustness in handling diverse applications [68]. In medical diagnosis and autonomous driving, multimodal fusion techniques enhance diagnostic accuracy and decision-making processes, underscoring the importance of integrating various data forms for reliable outcomes [69]. Laughter recognition systems also find practical applications in mental health monitoring, improving human-computer interaction by accurately recognizing and responding to laughter [35].

The enhanced reasoning capabilities in multimodal AI are evident in smart vision-language reasoners, which successfully implement sophisticated and adaptable solutions across multiple domains [70].

7.2 Applications in Emotion Recognition and Interaction

Multimodal AI plays a crucial role in emotion recognition and human-computer interaction by accurately interpreting and responding to human emotions, thereby enhancing user experience. Ortega et al. demonstrated the effectiveness of a Deep Neural Network (DNN) architecture in emotion recognition tasks, achieving higher Concordance Correlation Coefficient (CCC) scores than existing methods [71]. This underscores the capability of DNNs to capture complex emotional cues from multimodal data.

By integrating audio, visual, and textual signals, AI systems gain a comprehensive understanding of emotional states, which is particularly beneficial for virtual assistants and customer service bots, where multimodal sentiment analysis can significantly enhance service quality and user satisfaction [13, 14]. In mental health monitoring, multimodal AI systems analyze inputs like facial expressions, vocal tones, and physiological signals to detect emotional distress, thereby improving assessment accuracy and interventions [13, 15, 72]. This capability is crucial for timely interventions, demonstrating multimodal AI's positive impact on mental health care.

In educational technology, AI systems with multimodal capabilities can detect student frustration and engagement levels, enabling real-time content adaptation. This personalized approach enhances learning experiences and improves educational outcomes by supporting educators in decision-making processes [33, 7, 14, 47, 73]. The integration of multimodal AI in emotion recognition and interaction marks a significant advancement in developing intuitive, responsive systems capable of analyzing user sentiments across various expression forms, paving the way for more empathetic human-computer interactions [7, 13, 18, 8].

7.3 Medical Imaging and Healthcare Applications

The integration of multimodal AI in medical imaging and healthcare has revolutionized diagnostic and prognostic practices by combining diverse data modalities, such as radiology images, textual reports, and patient records. This integration enhances diagnostic accuracy and provides a comprehensive understanding of patient conditions. Notable contributions include datasets from ROCO, MedICaT, and MIMIC-CXR, which combine radiology images with captions to facilitate advanced multimodal learning [73].

The DRIFA-Net framework exemplifies effective multimodal fusion learning in medical imaging, significantly outperforming existing methods in disease classification tasks across multiple imaging datasets [74]. By leveraging the strengths of multimodal data, this framework enhances AI systems' diagnostic capabilities, demonstrating the potential for improved clinical outcomes through advanced techniques. Multimodal AI systems also support personalized treatment plans by integrating a wide range of patient data, essential for customizing healthcare interventions. Recent advancements show that these systems enhance clinical decision-making and predictive accuracy, outperforming traditional approaches by up to 33

The integration of multimodal AI in medical imaging and healthcare signifies a substantial advancement, offering enhanced diagnostic capabilities and personalized care solutions. By integrating diverse data sources, including electronic health records and multi-omics data, these advanced AI systems improve the precision, efficiency, and patient-centeredness of healthcare services, fostering a deeper understanding of individual health conditions and facilitating personalized treatment strategies [75, 73, 59].

7.4 Vision-Language Tasks and Multimodal Learning

Vision-language tasks are pivotal in advancing multimodal learning by integrating visual and textual data to enhance AI systems' interpretative and interactive capabilities. These tasks involve creating models capable of processing visual inputs, such as images and videos, alongside linguistic inputs, including text and speech. Such models perform complex applications like image captioning, visual question answering (VQA), and cross-modal retrieval, leveraging techniques that enhance the understanding of intricate relationships between visual and textual information. Recent advancements, including specialized architectures for fusing unimodal signals and utilizing large multimodal datasets, have significantly improved model performance in dynamic scenarios, enabling more accurate and contextually relevant responses [18, 8, 60, 33].

A notable advancement in this domain is the use of large-scale pretrained models that leverage extensive multimodal data to learn rich representations. Models based on the CrossTransformer architecture facilitate knowledge exchange between vision and language modalities, enhancing the system's ability to generate coherent and contextually relevant outputs [41]. This cross-modal knowledge exchange is crucial for tasks requiring deep understanding of both visual and linguistic contexts.

The integration of vision-language tasks into multimodal learning frameworks has improved interactive AI systems, such as virtual assistants and chatbots, which better understand and respond to user queries by combining visual cues with textual information. This capability is exemplified by AI chatbots utilizing multimodal models to comprehend text and images, enhancing overall interaction experiences [38]. Moreover, applying vision-language tasks in multimodal learning has expanded AI systems' potential to perform complex reasoning tasks, such as VQA, where systems interpret visual content to generate accurate answers to textual questions. These tasks underscore the importance of effective multimodal fusion techniques, like those employed in the Multimodal Bottleneck Transformer (MBT), which enhance computational efficiency while maintaining high fusion performance [55].

Vision-language tasks significantly impact multimodal learning by enhancing AI systems' ability to process and integrate diverse data forms, improving performance across a wide range of applications. Recent advancements in AI, particularly in enhancing vision models for text-heavy content and utilizing multimodal data, are leading to the development of more sophisticated and adaptable solutions. These innovations enable AI systems to comprehend and interact with complex visual and textual information similarly to human understanding. Improved vision models can accurately analyze images containing dense textual data, achieving an impressive accuracy rate of 96.71

8 Future Directions and Challenges

The advancement of artificial intelligence necessitates addressing the multifaceted challenges inherent in developing and deploying multimodal systems. These challenges not only limit current models' effectiveness but also hinder the scalability and adaptability of AI applications. A detailed understanding of these challenges, particularly in integrating diverse modalities, is crucial for fostering innovation and enhancing performance.

8.1 Challenges in Multimodal AI

Multimodal AI systems face significant challenges that impact their effectiveness and scalability. A major issue is the lack of comprehensive evaluation metrics and benchmarks, which impedes the assessment of integrated capabilities and localization of errors [6, 30]. In video and language integration, the Valley architecture's reliance on video and language inputs highlights the need for audio integration to enhance understanding [3]. Additionally, CentralNet's assumption of independent modality processing restricts its applicability [31].

Generating authentic emotional expressions in AI systems remains a challenge, as evidenced by the limitations in talking head generation models [2]. Scaling models to accommodate diverse modalities is also problematic [6]. Difficulties in integrating uncorrelated modalities, particularly in cross-modal learning with video inputs, and noise from secondary events further complicate generalization [4, 29].

These challenges necessitate innovative strategies to enhance robustness, scalability, and accessibility of multimodal AI systems, especially for complex tasks like interpreting text-heavy visual content and improving interpretability in high-stakes applications [18, 19, 8, 33]. Future research should focus on developing comprehensive benchmarks, improving data quality and diversity, and optimizing multimodal integration techniques.

8.2 Future Directions in Neural Integration

Future research in neural integration within multimodal AI will explore innovative avenues to enhance adaptability, efficiency, and robustness. Key areas include improving central network architectures like CentralNet for better fusion efficiency in complex scenarios [31]. Enhancing multilingual capabilities and incorporating audio in multimodal assistants like the Valley architecture is crucial

AI-generated, for reference only.

for robust operation across languages and modalities [3]. Improving datasets and refining evaluation frameworks are critical for better model generalization and adaptability [6].

In cross-modal learning, refining the correlation tower with stronger examples and exploring alternative loss formulations will enhance accuracy and robustness [4]. These advancements will significantly improve AI systems' sophistication, enabling seamless integration and processing of diverse sensory inputs across various applications, from image-to-text caption generation to visual question answering [18, 21, 6, 8].

8.3 Enhancements in Model Robustness and Adaptability

Enhancing the robustness and adaptability of multimodal AI models is vital for their application across diverse environments and tasks. Future research should optimize visual feature extraction to improve emotion recognition systems like the TelME framework [65]. Incorporating contextual factors such as physiological measurements can enhance model robustness [76]. Balancing intra-modal and inter-modal processing is crucial for optimizing frameworks like MAIL [77].

In biological prediction tasks, frameworks like EvoDeep should focus on enhancing robustness and adaptability [78]. Improving navigation frameworks and enhancing agents' generalization capabilities are essential for reliable performance across contexts [51]. Refining classification algorithms and expanding datasets for diverse ethical scenarios are necessary for better ethical evaluations [79]. Addressing noise challenges and expanding datasets for various conditions will improve model adaptability [80, 58]. Integrating lifelong learning capabilities will ensure AI systems remain effective over time [81].

8.4 Advancements in Fusion and Integration Techniques

Advancements in fusion and integration techniques are crucial for enhancing multimodal AI systems' performance and scalability. Efficient processing and integration of sensory inputs through advanced representation learning and information fusion methods bolster system robustness and adaptability across applications like image-to-text caption generation and visual question answering [18, 8]. The MSAF framework exemplifies efficient integration strategies by combining inter- and intra-modality relationships.

Dynamic Multimodal Fusion (DynMM) enhances efficiency by adaptively fusing data and tailoring computational paths according to modality demands, achieving significant computational cost reductions with minimal accuracy loss [32, 82, 83]. Future research should prioritize developing hybrid models that integrate various fusion techniques to enhance robustness and effectiveness [18, 33, 32, 20, 8]. Optimizing channel exchange processes and automatic registration mechanisms will improve performance in heterogeneous modality scenarios.

Refining multimodal data integration to improve prediction accuracy and engagement understanding remains a key focus. Automating parameter tuning in frameworks like AFS-MLI can improve adaptability across diverse domains [84, 19, 63]. Enhancements in handling multiple images and user interactions are vital for effective multimodal AI chatbots.

Exploring new fusion techniques and addressing real-time analysis challenges will significantly advance the field. Applying SFusion to variable multimodal tasks will enhance adaptability and effectiveness. Techniques like Co-Attention Multimodal Knowledge Transfer (CMKT) streamline computational requirements, crucial for applications like fake news detection [32, 8, 85]. The LAM-MSC framework exemplifies advancements in semantic extraction accuracy and reduced transmission overhead.

Future research should focus on improving audio localization accuracy and deeper integration of audio-visual attention mechanisms. Recent developments in data fusion techniques, exemplified by the LAM-MSC framework, enhance semantic extraction accuracy while minimizing overhead, essential for applications in fake news detection and cross-modal retrieval [32, 19, 86, 8]. Exploring novel uncertainty estimation methods and characterizing dynamic fusion's generalization capabilities are promising research areas.

Developing suitable quantization methods and exploring optimization techniques are essential for advancing fusion and integration techniques, paving the way for sophisticated and adaptable AI systems that address dynamic environments and diverse data forms, ultimately contributing to robust and effective multimodal AI solutions.

9 Conclusion

The survey elucidates the pivotal role of multimodal AI and sensory integration in augmenting AI capabilities across diverse fields. By synthesizing multiple sensory inputs, these systems consistently outperform unimodal counterparts, particularly in emotion recognition tasks, thereby enhancing detection accuracy. The integration of large language models within these frameworks has expanded AI's scope, enabling it to tackle intricate cognitive challenges, such as visual reconstruction from neural signals, thus propelling advancements in cognitive and neurological research.

Progress in multimodal fusion techniques, exemplified by their state-of-the-art performance on datasets like IEMOCAP, highlights the significance of robust representation learning in bolstering AI accuracy and resilience. These advancements address challenges in valence-arousal and expression recognition, underscoring their potential to enhance interpretability and precision in specialized applications.

Despite these advancements, challenges remain, particularly in spatial reasoning tasks, where models reveal limitations in cognitive processing strategies. Addressing these issues is crucial for optimizing the efficiency and scalability of multimodal AI systems. The survey's findings emphasize the transformative impact of multimodal AI on system development, paving the way for future technological breakthroughs. Continued exploration of innovative fusion and integration techniques will further amplify multimodal AI capabilities, fostering progress across various domains. As the field evolves, these advancements will facilitate the development of more sophisticated and adaptable AI solutions, adept at seamlessly integrating and processing complex sensory inputs to address the demands of diverse real-world scenarios.

References

[1] Lianyang Ma, Yu Yao, Tao Liang, and Tongliang Liu. Multi-scale cooperative multimodal transformers for multimodal sentiment analysis in videos, 2022.

[2] Sen Chen, Zhilei Liu, Jiaxing Liu, and Longbiao Wang. Talking head generation driven by speech-related facial action units and audio- based on multimodal representation fusion, 2022.

[3] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Da Li, Pengcheng Lu, Tao Wang, Linmei Hu, Minghui Qiu, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability, 2023.

[4] Palash Goyal, Saurabh Sahu, Shalini Ghosh, and Chul Lee. Cross-modal learning for multi-modal video categorization, 2020.

[5] Xudong Lin, Gedas Bertasius, Jue Wang, Shih-Fu Chang, Devi Parikh, and Lorenzo Torresani. Vx2text: End-to-end learning of video-based text generation from multimodal inputs, 2021.

[6] Sai Munikoti, Ian Stewart, Sameera Horawalavithana, Henry Kvinge, Tegan Emerson, Sandra E Thompson, and Karl Pazdernik. Generalist multimodal ai: A review of architectures, challenges and opportunities, 2024.

[7] Mutlu Cukurova, Carmel Kent, and Rosemary Luckin. Artificial intelligence and multimodal data in the service of human decision-making: A case study in debate tutoring. *British Journal of Educational Technology*, 50(6):3032–3046, 2019.

[8] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications, 2020.

[9] Xiongkuo Min, Guangtao Zhai, Jiantao Zhou, Xiao-Ping Zhang, Xiaokang Yang, and Xinping Guan. A multimodal saliency model for videos with high audio-visual correspondence. *IEEE Transactions on Image Processing*, 29:3805–3819, 2020.

[10] Jiarui Xie, Mutahar Safdar, Lequn Chen, Seung Ki Moon, and Yaoyao Fiona Zhao. Audio-visual cross-modality knowledge transfer for machine learning-based in-situ monitoring in laser additive manufacturing, 2025.

[11] Daan Schouten, Giulia Nicoletti, Bas Dille, Catherine Chia, Pierpaolo Vendittelli, Megan Schuurmans, Geert Litjens, and Nadieh Khalili. Navigating the landscape of multimodal ai in medicine: a scoping review on technical challenges and clinical applications, 2024.

[12] Yaowei Li, Ruijie Quan, Linchao Zhu, and Yi Yang. Efficient multimodal fusion via interactive prompting, 2023.

[13] Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444, 2023.

[14] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion*, 37:98–125, 2017.

[15] Zhuofan Wen, Fengyu Zhang, Siyuan Zhang, Haiyang Sun, Mingyu Xu, Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Multimodal fusion with pre-trained model features in affective behaviour analysis in-the-wild, 2024.

[16] Jinxiang Lai, Jie Zhang, Jun Liu, Jian Li, Xiaocheng Lu, and Song Guo. Spider: Any-to-many multimodal llm, 2024.

[17] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*, 2024.

[18] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493, 2020.

[19] Yunkai Dang, Kaichen Huang, Jiahao Huo, Yibo Yan, Sirui Huang, Dongrui Liu, Mengxi Gao, Jie Zhang, Chen Qian, Kun Wang, Yong Liu, Jing Shao, Hui Xiong, and Xuming Hu. Explainable and interpretable multimodal large language models: A comprehensive survey, 2024.

[20] Aishwarya Jayagopal, Ankireddy Monica Aiswarya, Ankita Garg, and Srinivasan Kolumam Nandakumar. Multimodal representation learning with text and images, 2022.

[21] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[22] João Ranhel and Cacilda Vilela de Lima. On the linguistic and computational requirements for creating face-to-face multimodal human-machine interaction, 2022.

[23] Swati Swati, Arjun Roy, and Eirini Ntoutsi. Exploring fusion techniques in multimodal ai-based recruitment: Insights from faircvdb, 2024.

[24] Feibo Jiang, Li Dong, Yubo Peng, Kezhi Wang, Kun Yang, Cunhua Pan, and Xiaohu You. Large ai model empowered multimodal semantic communications. *IEEE Communications Magazine*, 2024.

[25] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors, 2018.

[26] Robert J. Piechocki, Xiaoyang Wang, and Mohammud J. Bocus. Multimodal sensor fusion in the latent representation space, 2022.

[27] Jia Fu, Jiarui Tan, Wenjie Yin, Sepideh Pashami, and Mårten Björkman. Component attention network for multimodal dance improvisation recognition, 2023.

[28] Peng Hu, Liangli Zhen, Dezhong Peng, and Pei Liu. Scalable deep multimodal learning for cross-modal retrieval. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 635–644, 2019.

[29] Hai Huang, Yan Xia, Shengpeng Ji, Shulei Wang, Hanting Wang, Jieming Zhu, Zhenhua Dong, and Zhou Zhao. Unlocking the potential of multimodal unified discrete representation through training-free codebook optimization and hierarchical alignment, 2024.

[30] Nawaf Alampara, Mara Schilling-Wilhelmi, Martiño Ríos-García, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, N. M. Anoop Krishnan, and Kevin Maik Jablonka. Probing the limitations of multimodal language models for chemistry and materials research, 2024.

[31] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[32] Multimodal fusion with co-attent.

[33] Adithya TG, Adithya SK, Abhinav R Bharadwaj, Abhiram HA, and Surabhi Narayan. Enhancing vision models for text-heavy content understanding and interaction, 2024.

[34] Zecheng Liu, Jia Wei, Rui Li, and Jianlong Zhou. Sfusion: Self-attention based n-to-one multimodal fusion block, 2023.

[35] Fuzheng Zhao and Yu Bai. Design and development of laughter recognition system based on multimodal fusion and deep learning, 2024.

[36] Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data, 2023.

[37] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chenliang Xu. Deep cross-modal audio-visual generation, 2017.

AI-generated, for reference only.

[38] Min Young Lee. Building multimodal ai chatbots, 2023.

[39] Cheng Charles Ma, Kevin Hyekang Joo, Alexandria K. Vail, Sunreeta Bhattacharya, Álvaro Fernández García, Kailana Baker-Matsuoka, Sheryl Mathew, Lori L. Holt, and Fernando De la Torre. Multimodal fusion with llms for engagement prediction in natural conversation, 2024.

[40] J. Bieniek, M. Rahouti, and D. C. Verma. Generative ai in multimodal user interfaces: Trends, challenges, and cross-platform adaptability, 2024.

[41] Renyu Zhu, Chengcheng Han, Yong Qian, Qiushi Sun, Xiang Li, Ming Gao, Xuezhi Cao, and Yunsen Xian. Exchanging-based multimodal fusion with transformer, 2023.

[42] Farizal Hamzah and Nuraini Sulaiman. Multimodal integration in large language models: A case study with mistral llm. 2024.

[43] George Sterpu, Christian Saam, and Naomi Harte. How to teach dnns to pay attention to the visual modality in speech recognition, 2020.

[44] Mafalda Malafaia, Thalea Schlender, Peter A. N. Bosman, and Tanja Alderliesten. Multifix: An xai-friendly feature inducing approach to building models from multimodal data, 2024.

[45] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: a multi-layer approach for multimodal fusion, 2018.

[46] Lucia Gordon, Nico Lang, Catherine Ressijac, and Andrew Davies. Multimodal fusion strategies for mapping biophysical landscape features, 2024.

[47] Yun-Da Tsai, Ting-Yu Yen, Pei-Fu Guo, Zhe-Yan Li, and Shou-De Lin. Text-centric alignment for multi-modality learning, 2024.

[48] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*, 2018.

[49] Jianning Wu, Zhuqing Jiang, Shiping Wen, Aidong Men, and Haiying Wang. Rethinking the constraints of multimodal fusion: case study in weakly-supervised audio-visual video parsing, 2021.

[50] Kai Li, Runxuan Yang, Fuchun Sun, and Xiaolin Hu. Iianet: An intra- and inter-modality attention network for audio-visual speech separation, 2024.

[51] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9701–9707. IEEE, 2020.

[52] Faisal Mahmood, Ziyun Yang, Thomas Ashley, and Nicholas J. Durr. Multimodal densenet, 2018.

[53] Guanqun Cao and Shan Luo. Multimodal perception for dexterous manipulation, 2021.

[54] Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis, 2023.

[55] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems*, 34:14200–14213, 2021.

[56] Attention bottlenecks for multimodal fusion.

[57] Mehrshad Zandigohar, Mo Han, Mohammadreza Sharif, Sezen Yagmur Gunay, Mariusz P. Furmanek, Mathew Yarossi, Paolo Bonato, Cagdas Onal, Taskin Padir, Deniz Erdogmus, and Gunar Schirner. Multimodal fusion of emg and vision for human grasp intent inference in prosthetic hand control, 2024.

[58] Parag Narkhede, Rahee Walambe, Shruti Mandaokar, Pulkit Chandel, Ketan Kotecha, and George Ghinea. Gas detection and identification using multimodal artificial intelligence based sensor fusion, 2021.

[59] Farida Mohsen, Hazrat Ali, Nady El Hajj, and Zubair Shah. Artificial intelligence-based methods for fusion of electronic health records and imaging data, 2022.

[60] Qilang Ye, Zitong Yu, Rui Shao, Xinyu Xie, Philip Torr, and Xiaochun Cao. Cat: Enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios, 2024.

[61] Dave Braines, Alun Preece, and Dan Harborne. Multimodal explanations for ai-based multi-sensor fusion. *NATO SET-262 RSM on artificial intelligence for military multisensor fusion engines. NATO*, 2018.

[62] Shiv Shankar. Neural dependency coding inspired multimodal fusion, 2021.

[63] Shanshan Zhong, Shanghua Gao, Zhongzhan Huang, Wushao Wen, Marinka Zitnik, and Pan Zhou. Moextend: Tuning new experts for modality and task extension, 2024.

[64] Xinyi Ding, Tao Han, Yili Fang, and Eric Larson. An approach for combining multimodal fusion and neural architecture search applied to knowledge tracing, 2021.

[65] Taeyang Yun, Hyunkuk Lim, Jeonghwan Lee, and Min Song. Telme: Teacher-leading multi-modal fusion network for emotion recognition in conversation, 2024.

[66] Amit Eliav and Sharon Gannot. Audio-visual approach for multimodal concurrent speaker detection, 2025.

[67] Saurav Sahay, Eda Okur, Shachi H Kumar, and Lama Nachman. Low rank fusion based transformers for multimodal sequences, 2020.

[68] Jinhong Ni, Yalong Bai, Wei Zhang, Ting Yao, and Tao Mei. Deep equilibrium multimodal fusion, 2023.

[69] Qingyang Zhang, Yake Wei, Zongbo Han, Huazhu Fu, Xi Peng, Cheng Deng, Qinghua Hu, Cai Xu, Jie Wen, Di Hu, and Changqing Zhang. Multimodal fusion on low-quality data: A comprehensive survey, 2024.

[70] Denisa Roberts and Lucas Roberts. Smart vision-language reasoners, 2024.

[71] Juan DS Ortega, Mohammed Senoussaoui, Eric Granger, Marco Pedersoli, Patrick Cardinal, and Alessandro L Koerich. Multimodal fusion with deep neural networks for audio-video emotion recognition. *arXiv preprint arXiv:1907.03196*, 2019.

[72] Sharmeen M Saleem Abdullah Abdullah, Siddeeq Y Ameen Ameen, Mohammed AM Sadeeq, and Subhi Zeebaree. Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends*, 2(01):73–79, 2021.

[73] Zhihong Chen, Guanbin Li, and Xiang Wan. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge, 2022.

[74] Joy Dhar, Nayyar Zaidi, Maryam Haghighat, Puneet Goyal, Sudipta Roy, Azadeh Alavi, and Vikas Kumar. Multimodal fusion learning with dual attention for medical imaging, 2024.

[75] Luis R Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M Wiberg, Michael L Li, Ignacio Fuentes, and Dimitris Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ digital medicine*, 5(1):149, 2022.

[76] Kenneth Ooi, Karn N. Watcharasupat, Bhan Lam, Zhen-Ting Ong, and Woon-Seng Gan. Autonomous soundscape augmentation with multimodal fusion of visual and participant-linked inputs, 2024.

[77] Junnan Dong, Qinggang Zhang, Huachi Zhou, Daochen Zha, Pai Zheng, and Xiao Huang. Modality-aware integration with large language models for knowledge-based visual question answering, 2024.

[78] Xi Victoria Lin, Akshat Shrivastava, Liang Luo, Srinivasan Iyer, Mike Lewis, Gargi Ghosh, Luke Zettlemoyer, and Armen Aghajanyan. Moma: Efficient early-fusion pre-training with mixture of modality-aware experts, 2024.

[79] Alexis Roger, Esma Aïmeur, and Irina Rish. Towards ethical multimodal systems, 2024.

[80] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. Multimodal sentiment analysis with word-level fusion and reinforcement learning, 2018.

[81] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Audio visual language maps for robot navigation. In *International Symposium on Experimental Robotics*, pages 105–117. Springer, 2023.

[82] Zihui Xue and Radu Marculescu. Dynamic multimodal fusion, 2023.

[83] Laura Wenderoth. Exploring multi-modality dynamics: Insights and challenges in multimodal fusion for biomedical tasks, 2024.

[84] Soyeon Caren Han, Feiqi Cao, Josiah Poon, and Roberto Navigli. Multimodal large language models and tunings: Vision, language, sensors, audio, and beyond, 2024.

[85] Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloe Clavel. Improving multimodal fusion via mutual dependency maximisation, 2021.

[86] Tianyi Bai, Hao Liang, Binwang Wan, Yanran Xu, Xi Li, Shiyu Li, Ling Yang, Bozhou Li, Yifan Wang, Bin Cui, Ping Huang, Jiulong Shan, Conghui He, Binhang Yuan, and Wentao Zhang. A survey of multimodal large language model from a data-centric perspective, 2024.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.