Multimodal Technologies and Interaction

MDPI

*Review*

# Multimodal Interaction, Interfaces, and Communication: A Survey

**Elias Dritsas [1], Maria Trigka [1], Christos Troussas [2,*] and Phivos Mylonas [2]**

[1] Industrial Systems Institute (ISI), Athena Research and Innovation Center, 26504 Patras, Greece; dritsas@isi.gr (E.D.); trigka@isi.gr (M.T.)
[2] Department of Informatics and Computer Engineering, University of West Attica, Egaleo Park Campus, 12243 Athens, Greece; mylonasf@uinwa.gr
* Correspondence: ctrouss@uinwa.gr

**Abstract:** Multimodal interaction is a transformative human-computer interaction (HCI) approach that allows users to interact with systems through various communication channels such as speech, gesture, touch, and gaze. With advancements in sensor technology and machine learning (ML), multimodal systems are becoming increasingly important in various applications, including virtual assistants, intelligent environments, healthcare, and accessibility technologies. This survey concisely overviews recent advancements in multimodal interaction, interfaces, and communication. It delves into integrating different input and output modalities, focusing on critical technologies and essential considerations in multimodal fusion, including temporal synchronization and decision-level integration. Furthermore, the survey explores the challenges of developing context-aware, adaptive systems that provide seamless and intuitive user experiences. Lastly, by examining current methodologies and trends, this study underscores the potential of multimodal systems and sheds light on future research directions.

**Keywords:** multimodal interaction; human-computer interaction; adaptive systems; multimodal fusion

check for updates

Academic Editors: Jan Auernhammer, Takumi Ohashi and Wei Liu

Received: 7 October 2024
Revised: 8 January 2025
Accepted: 10 January 2025
Published: 14 January 2025

**Citation:** Dritsas, E.; Trigka, M.; Troussas, C.; Mylonas, P. Multimodal Interaction, Interfaces, and Communication: A Survey. *Multimodal Technol. Interact.* **2025**, *9*, 6. https://doi.org/10.3390/mti9010006

## 1. Introduction

Multimodal interaction signifies a significant change in HCI, integrating multiple communication channels or modalities to enhance user experiences with extraordinary richness, intuitiveness, and efficiency. Traditionally, computer interaction has been limited to single inputs such as keyboard, mouse, or text-based commands, which can limit the naturalness of user engagement. However, recent advancements in sensor technologies, ML, and natural language processing (NLP) have opened the door for systems to understand and respond to various inputs, including speech, gesture, gaze, touch, and haptics. This shift towards multimodal interfaces enables more seamless and adaptable interaction, making technology more akin to human communication patterns, where multiple sensory inputs are processed simultaneously to convey intent, emotion, and context [1–3].

Multimodal interaction aims to improve usability and accessibility by enabling users to engage with systems in the most natural and convenient ways. By utilizing multiple input channels, multimodal systems can provide more flexible, robust, and context-aware interactions. For instance, a voice command can complement a hand gesture to clarify the user’s intention in a noisy environment. Similarly, in hands-free or immersive environments such as augmented reality (AR), virtual reality (VR) and mixed reality (MR), gaze-based input can work with touch or speech to streamline interactions. The power of multimodal systems lies in their capacity to merge and synchronize diverse inputs, giving users greater control over their interactions with technology while enhancing the system's real-time responsiveness and accuracy [4–6].

Despite multimodal systems' enormous potential, their development presents several challenges. It is crucial to carefully coordinate the fusion of different input modalities to ensure that they complement each other rather than cause conflicts. That involves complex input synchronization, modality fusion, and context awareness, all of which demand advanced algorithms and ML techniques. Additionally, multimodal systems must be designed to adapt to diverse environments and user preferences, making them versatile across various applications, from healthcare and education to entertainment and smart environments [7–9].

Table 1 summarizes prior surveys on multimodal interaction, highlighting their focus areas. This survey advances the discourse on multimodal interaction systems by providing a holistic and integrated analysis that spans critical technologies, synchronization challenges, adaptive systems, and future research directions. Unlike [10], which primarily explores vision-based multimodal techniques, including gesture and gaze, this survey includes a broader range of modalities, such as speech, touch, and haptics, and discusses their integration in diverse applications. This comprehensive approach not only contextualizes vision-based systems but also delves into the interplay of non-visual modalities, offering insights into their practical implementation in real-world environments. The author in [11] emphasizes the cognitive and neuroscience foundations of multimodal systems, focusing on language processing and multimodal ML. While these aspects are significant, this survey extends beyond foundational theories to address applied challenges such as temporal synchronization and real-time decision-level integration, which are crucial for developing adaptive systems in dynamic contexts. Furthermore, this survey introduces a detailed exploration of emerging sensory technologies like brain-computer interfaces (BCIs), which are not deeply examined in [11].

**Table 1.** Summary of surveys and descriptions.

| Survey | Description |
|---|---|
| [10] | Overview of multimodal HCI focusing on vision-based techniques like gesture, gaze, and affective interaction, including challenges and emerging applications. |
| [11] | State-of-the-art analysis of multimodal-multisensor interfaces, emphasizing cognitive and neuroscience foundations, language processing, multimodal ML, and future research directions. |
| [12] | Focuses on challenges in designing multimodal interfaces, discussing technologies for integrating speech, gesture, and gaze recognition with robust interface designs for real-world applications. |
| [13] | Examination of multimodal interfaces and communication cues in remote collaboration, highlighting VR/AR/MR technologies and their impact on task performance and user experience. |
| [14] | Analysis of multimodal interfaces for HCI, detailing their evolution, principles, and use cases across domains like robotics, transport, and education. |
| [15] | Study of model-driven engineering approaches to multimodal interaction, focusing on simplifying the design and development of mobile multimodal applications. |
| [16] | Comprehensive discussion on multimodal interaction principles, frameworks, and architectures, including real-time processing and fusion of multiple data types. |
| [17] | Presents a comprehensive review of how interaction technologies such as VR, AR, haptics, and tracking are utilized across various domains, including medicine, cultural heritage, transportation, and industry. |

**Table 1.** *Cont.*

| Survey | Description |
|---|---|
| [18] | Explores the evolution and key aspects of multimodal interaction systems. It discusses the definition, advantages, and history of multimodal interaction, focusing on the role of input/output modalities, fusion engines, and human-centered interaction approaches. |
| [19] | Discusses the design and functionality of multimodal systems that integrate multiple input and output modalities, such as speech, gestures, and visual inputs. It introduces key concepts like multimodal messages, temporal relationships, and classes of cooperation between modalities. |
| [20] | Provides an overview of the evolution, opportunities, and challenges in multimodal HCI. It explores input/output modalities, integration methods, and key design principles, offering insights into the historical advancements and future directions in the field. |

In [12] is discussed the integration of speech, gesture, and gaze recognition, with an emphasis on robust interface designs. While this overlaps with some topics covered here, this survey provides a more extensive analysis of synchronization techniques and their application in creating context-aware systems. It also addresses challenges in multimodal fusion and adaptability in noisy or complex environments, which are not the primary focus of [12]. Moreover, [13] investigates multimodal interfaces in remote collaboration and their influence on user experience, particularly within VR, AR, and MR environments. This survey builds on such insights by exploring these technologies across a wider array of use cases, including accessibility and healthcare. The inclusion of future trajectories, like integrating BCIs, positions this work as a forward-looking contribution compared to the more domain-specific focus of [13].

In [14] is analyzed the evolution and principles of multimodal interfaces, emphasizing robotics, transportation, and education. This survey complements and expands on these discussions by presenting a unified roadmap that ties technological advancements to user-centric design principles, focusing on adaptive and context-aware interactions that are applicable across domains. The model-driven engineering approaches to multimodal interaction highlighted in [15] provide a valuable perspective on simplifying interface design. However, this survey takes a different approach by focusing on challenges in deployment, particularly modality fusion and real-time processing, offering practical solutions for enhancing usability and adaptability in real-world applications.

Furthermore, the authors in [16–18] delve into the principles, frameworks, and architectural considerations of multimodal systems. While these studies provide strong conceptual models, this survey distinguishes itself by presenting a detailed examination of modality synchronization, error handling, and the role of artificial intelligence (AI) in personalizing user experiences, thus bridging theoretical insights with actionable methodologies. Lastly, the evolution of multimodal systems and the integration of multiple modalities are outlined in [19,20]. This survey extends the aforementioned works' scope by emphasizing emerging trends and addressing gaps, such as the practical challenges of achieving seamless interaction in noisy environments and leveraging advanced ML algorithms for context-aware adaptation.

In conclusion, this survey distinguishes itself by integrating a wide range of multimodal interaction aspects, addressing both theoretical foundations and applied challenges. It offers a comprehensive roadmap for advancing the field, with a forward-looking focus on future directions like BCIs and advanced sensory technologies. This work not only synthesizes the state of the art but also provides critical insights into overcoming current limitations, setting a benchmark for future research and development. Specifically, this survey:

- Provides an analysis of the various input and output modalities used in multimodal interaction systems, such as speech, gesture, touch, and gaze.
- Discusses the challenges involved in modality fusion, synchronization, and context awareness, offering insight into current solutions and ongoing research.
- Explores the applications of multimodal systems across industries, including healthcare, VR, smart environments, and accessibility technologies.
- Highlights future directions for multimodal interaction, focusing on advancements in ML and emerging technologies like BCIs.

Figure 1 illustrates a visual overview of the key topics, major application areas, and future directions in the field. It serves as a roadmap for the following sections, guiding a structured exploration of the challenges and solutions in multimodal interaction systems and their integration into real-world applications. The figure highlights the interplay between modalities, practical implementations, and technological advancements necessary for future development, ensuring a cohesive narrative throughout the subsequent sections.

More specifically, the remaining paper is structured as follows. Section 2 defines multimodal interaction and interfaces. Moreover, Section 3 outlines the communication in multimodal systems. Section 4 analyzes interaction modalities. Besides, Section 5 notes challenges in multimodal interaction systems. Section 6 discusses applications of multimodal systems and future directions. Finally, Section 7 summarizes the findings of this research survey.

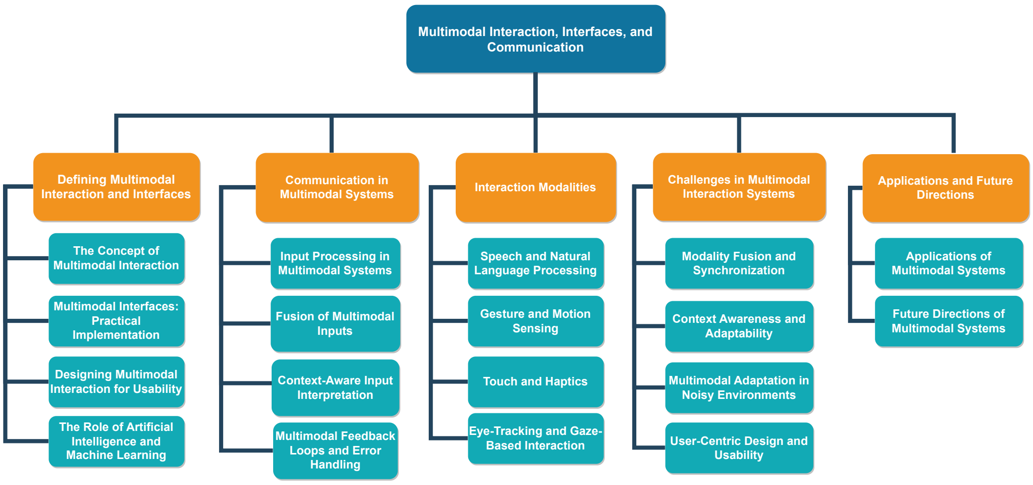
Multimodal Interaction, Interfaces, and Communication
Defining Multimodal Interaction and Interfaces
The Concept of Multimodal Interaction
Multimodal Interfaces: Practical Implementation
Designing Multimodal Interaction for Usability
The Role of Artificial Intelligence and Machine Learning
Communication in Multimodal Systems
Input Processing in Multimodal Systems
Fusion of Multimodal Inputs
Context-Aware Input Interpretation
Multimodal Feedback Loops and Error Handling
Interaction Modalities
Speech and Natural Language Processing
Gesture and Motion Sensing
Touch and Haptics
Eye-Tracking and Gaze-Based Interaction
Challenges in Multimodal Interaction Systems
Modality Fusion and Synchronization
Context Awareness and Adaptability
Multimodal Adaptation in Noisy Environments
User-Centric Design and Usability
Applications and Future Directions
Applications of Multimodal Systems
Future Directions of Multimodal Systems

**Figure 1.** Roadmap for the topics explored in this survey.

## 2. Defining Multimodal Interaction and Interfaces

The evolution of HCI has progressed toward more natural and intuitive systems, where multimodal interaction and interfaces play a pivotal role. By enabling communication through multiple sensory channels, such as speech, gesture, and touch, these systems emulate the richness of human-to-human communication. This section delves into the principles and design of multimodal interaction and interfaces, highlighting their transformative potential in creating seamless and adaptive interactions.

### 2.1. *The Concept of Multimodal Interaction*

Multimodal interaction is fundamentally rooted in the ability to use multiple sensory channels for communication between humans and machines, aiming to mimic the natural ways humans interact with the world and each other. Traditional human-computer interfaces like keyboards and mice are limited to single-modality inputs. In contrast, multimodal interaction systems integrate diverse input channels like speech, gesture, touch, and gaze, allowing users to interact more flexibly and intuitively. This approach leverages the inherent strengths of different modalities, overcoming the limitations of each and providing a richer interaction experience. By enabling the concurrent use of multiple modalities, these systems allow users to communicate more naturally, combining verbal and non-verbal cues, as humans often do in everyday conversations [11,21,22].

Complementarity is a central concept in multimodal interaction, where different modalities convey more detailed or clarified information. For example, in a system where users give verbal commands, pointing gestures can specify the command's target, as in the instruction, "Turn on that light", while pointing to a specific fixture. This complementary use of modalities reduces ambiguities and enhances system reliability by compensating for any channel's weaknesses or uncertainties. Furthermore, redundancy strengthens interaction robustness when the same information is provided across different modalities. If one modality fails or is disrupted (such as speech recognition in a noisy environment), another modality (such as a touch gesture) can ensure the continuity of interaction. The system's capacity to handle these multimodal inputs simultaneously forms the basis for more prosperous, more effective user-system communication [23–25].

Beyond enhancing user experience, multimodal interaction systems are designed to adapt to various contexts and user preferences. They achieve this by leveraging ML algorithms and context-aware systems that dynamically adjust which modalities are prioritized based on environmental conditions or user behaviour. For instance, in a noisy room, a system may shift its focus from speech recognition to gesture or touch input to maintain efficient interaction. The underlying complexity of such systems lies in their ability to harmonize multiple modalities without increasing the cognitive load on the user. Ensuring smooth, seamless transitions between different input types and enabling users to naturally switch between them reflects the ultimate goal of multimodal interaction systems: creating interactions that feel as intuitive and adaptable as human communication itself [26–28].

### *2.2. Multimodal Interfaces: Practical Implementation*

Multimodal interfaces are the technical systems that support multimodal interaction, providing the infrastructure necessary for capturing, processing, and integrating inputs from diverse sensory channels. These interfaces interpret user inputs and coordinate multiple output modalities to deliver appropriate feedback. This interplay between input and output across modalities forms the core of effective multimodal systems [29–31].

The role of input modalities, such as speech, gestures, touch, and gaze, is critical in enabling rich interaction. Speech remains one of the most commonly used modalities due to advancements in automatic speech recognition (ASR) technologies. However, challenges such as background noise, accents, and variations in speech patterns still affect the reliability of speech recognition systems. On the other hand, gesture recognition enables users to interact with systems using natural movements captured by cameras or sensors. Gesture input is instrumental in scenarios where speech or touch may not be feasible, such as when a user's hands are occupied [32–34].

Touch and haptic feedback are essential, particularly in mobile devices and virtual environments. Touchscreens allow users to interact directly with visual interfaces, while haptic systems provide tactile feedback that enhances the user's sense of control. For example, in VR, haptic feedback can simulate physical sensations, contributing to a more immersive experience. Eye-tracking and gaze-based systems offer another dimension to multimodal interaction by allowing users to control systems with their gaze, making it particularly useful in accessibility contexts or hands-free environments [35–37].

On the output side, multimodal interfaces generate feedback through visual, auditory, and tactile channels. Visual output, typically delivered through screens or AR displays, provides users real-time information. Auditory output, such as spoken responses or sound alerts, is especially useful in hands-free environments or when visual attention is elsewhere. Haptic feedback, often in the form of vibrations or force resistance, enhances the physicality of interaction, especially in AR/VR systems [38–40].

### *2.3. Designing Multimodal Interaction for Usability*

The design of multimodal interfaces must account for user experience, balancing the complexity of multiple input channels with the need for a clear and intuitive interaction flow. User-centred design (UCD) principles are critical in guiding the development of multimodal systems, ensuring that interfaces are tailored to user needs and preferences [41–43].

A key concern in multimodal design is cognitive load. When users are presented with too many modalities simultaneously, the interaction can become overwhelming, leading to errors or frustration. Therefore, multimodal systems must be designed to offer seamless transitions between modalities, allowing users to switch between input channels as needed without increasing mental effort. Effective multimodal design also considers personalization, where systems adapt to individual user preferences, learning which modalities are most effective in different contexts [44–46].

Error handling is another crucial aspect of usability in multimodal systems. Given the complexity of interpreting multiple input modalities, errors are inevitable. A well-designed system must detect when an input has been misunderstood and ask for clarification or offer alternative responses. Providing real-time feedback through visual or auditory cues can help guide users and reduce the likelihood of repeated errors [47–49].

### *2.4. The Role of Artificial Intelligence and Machine Learning*

AI and ML are critical enablers of advanced multimodal interaction systems, providing the computational power and flexibility to interpret complex and dynamic inputs. Multimodal systems generate diverse data streams—such as speech, gestures, touch, and gaze—each requiring specialized processing techniques. AI and ML allow systems to integrate these inputs seamlessly, making interactions more intuitive, adaptive, and robust [50,51].

One of the key contributions of AI in multimodal systems is its ability to adapt dynamically to user behaviour and environmental contexts. For example, AI models can analyze real-time data to determine the most suitable modalities for a given scenario. In a smart home environment, when speech recognition is hindered by noise, the system can shift focus to touch or gesture inputs. These adaptive capabilities are made possible through ML algorithms that learn from historical user interactions, optimizing modality selection and improving interaction efficiency over time [52–54].

AI-driven models also enhance the personalization of multimodal systems. By analyzing individual user preferences and interaction patterns, ML algorithms can tailor interfaces to specific needs. For instance, a system may prioritize gesture recognition for a user who frequently uses hand motions, while another may emphasize speech-based commands for a user with limited mobility. This level of personalization improves usability and increases user satisfaction by aligning the system's behaviour with individual expectations [55–58].

Furthermore, AI is integral to managing the computational complexity of multimodal systems. Advanced algorithms such as deep learning (DL) enable efficient feature extraction, real-time decision-making, and error correction across modalities. For example, neural networks can resolve ambiguities in user input by integrating data from multiple modalities, such as combining speech and gaze to disambiguate a command. Predictive models also allow systems to anticipate user intent, reducing cognitive load and enhancing the fluidity of interactions [59–61].

Although AI and ML provide significant advantages, their implementation in multimodal systems poses challenges, such as the need for high computational resources and ensuring user privacy. Balancing the scalability of AI models with real-time performance remains a critical area of research, particularly for applications in resource-constrained environments like mobile devices. Additionally, ethical considerations, including the security and confidentiality of user data, are essential to fostering trust and ensuring the responsible deployment of AI in multimodal systems [62,63].

By leveraging AI and ML, multimodal systems are evolving to become more context-aware, personalized, and efficient, setting the stage for transformative advances in human-computer interaction. These technologies enable seamless integration across diverse modalities, making interactions more natural and adaptive to user needs [64].

Table 2 provides a clear view of the current landscape in multimodal systems and categorizes surveyed references across various domains related to multimodal interaction. It highlights critical areas such as modality fusion, synchronization challenges, user-centred design, and the role of AI and ML in enhancing system adaptability. The table serves as a foundation for understanding the complexities of integrating multiple modalities and offers insights into the key technologies driving the field forward.

**Table 2.** A classification of references on multimodal interaction and interfaces.

| References | Focus Area | Techniques/Features | Description |
|---|---|---|---|
| [11,21,22] | Multimodal integration | Conceptual integration across modalities | Theoretical models for HCI improvements with limited real-world validations. |
| [23–25] | Complementarity and redundancy | Fusion based on redundancy and complementarity | Enhances system robustness by combining modalities to overcome weaknesses. |
| [26–28] | Context-aware adaptation | ML-based dynamic prioritization | Systems adjust input modalities dynamically based on environmental contexts. |
| [29–31] | Multimodal infrastructure | Core systems for input-output integration | Highlights systems' role in enhancing multimodal feedback mechanisms. |
| [32–34] | Gesture recognition | Camera/sensor-based gesture systems | Addresses usability in speech-limited or hands-occupied scenarios. |
| [35–37] | Eye-tracking and gaze | Accessibility and hands-free systems | Utilizes gaze tracking for accessibility and dynamic interaction. |
| [38–40] | Haptic feedback | Enhancing AR/VR immersion | Simulates real-world tactile interactions for improved engagement. |
| [41–43] | UCD | Principles for user-driven system design | Guides development to align with user needs and preferences. |
| [44–46] | Cognitive load management | Modality transitions and personalization | Emphasizes seamless modality switching and learning user preferences. |
| [47–49] | Error handling and feedback | Real-time correction and clarification | Uses feedback mechanisms to mitigate input errors. |
| [50,51] | AI/ML for data fusion | General multimodal integration | AI and ML enable seamless interpretation of diverse input streams. |
| [52–54] | Adaptive learning | Context-aware smart systems | Systems adapt dynamically to real-time environmental contexts. |
| [55–58] | User personalization | Tailored interfaces for users | AI models learn individual preferences to enhance usability. |

**Table 2.** *Cont.*

| References | Focus Area | Techniques/Features | Description |
|---|---|---|---|
| [59–61] | Neural networks and predictive models | Real-time error handling and intent prediction | Resolves ambiguities and reduces cognitive load through AI integration. |
| [62,63] | Privacy and ethical AI | Secure and scalable systems | Addresses challenges like computational overhead and user data confidentiality. |
| [64] | Multimodal dialogue systems | Industry-specific AI dialogue models | Explores dialogue systems in AI-assisted industries for adaptive and effective communication. |

## 3. Communication in Multimodal Systems

Effective communication in multimodal systems hinges on the seamless integration of diverse sensory inputs and outputs to facilitate natural and intuitive user interactions. These systems rely on sophisticated processing techniques to interpret, fuse, and respond to multimodal data in real-time, ensuring accuracy and coherence across channels. This section delves into the core principles, challenges, and advancements in communication mechanisms within multimodal systems, emphasizing their role in creating adaptive and context-aware interactions.

### *3.1. Input Processing in Multimodal Systems*

In multimodal systems, the first stage of communication begins with processing user inputs across different modalities. This input processing is not only about receiving raw data but involves understanding, transforming, and interpreting these inputs in a way the system can handle [65,66].

Each modality—speech, gestures, or touch—requires specialized signal processing techniques. For instance, speech input relies on ASR, which involves several layers of signal filtering, noise reduction, and the application of sophisticated NLP models to understand user intent. In contrast, gestural inputs use motion capture technologies, often based on depth-sensing cameras or wearable sensors, to detect body parts' position, movement, and orientation. Gesture recognition algorithms then process this data to distinguish meaningful actions from irrelevant or involuntary movements [67–69].

At this stage, one of the significant challenges is dealing with the inherent variability in input signals, both within and across modalities. Speech, for example, can be affected by accents, background noise, or even the user's emotional state, while gestures can vary in speed, amplitude, or fluidity depending on the user's context. Multimodal systems must, therefore, employ robust preprocessing mechanisms to ensure that the input from each modality is accurately interpreted before it is passed on for fusion [70–72]. Additionally, these systems need to account for cultural differences and personal preferences in gesture interpretation [73–75], as well as adapt dynamically to variations in environmental conditions and user behavior to maintain accuracy and usability across diverse scenarios. This adaptability is critical for achieving reliable and seamless interaction in real-world applications.

### *3.2. Fusion of Multimodal Inputs*

The fusion of multimodal inputs is a critical component of effective communication in multimodal systems. By integrating data from diverse modalities such as speech, gestures, and gaze, these systems can achieve a coherent interpretation of user intent. The fusion process enables complementary modalities to work together, mitigating the limitations of individual channels and enhancing overall system robustness. For example, a speech command such as "Turn on that light" can be clarified by a simultaneous pointing gesture, ensuring accurate intent recognition [76–78].

Fusion can occur at different levels (as shown in Figure 2), depending on the system's requirements. Early or sensor-level fusion involves combining raw data from different modalities during initial processing. This approach captures the synergies between modalities but requires robust preprocessing algorithms to handle high-dimensional data. Feature-level fusion, on the other hand, integrates features extracted from each modality, enabling systems to focus on specific characteristics relevant to the task. Decision-level fusion, where outputs from independent classifiers are combined, is often employed in systems requiring modularity or asynchronous operation [79,80].

Synchronization is a central challenge in multimodal fusion. Inputs from different modalities frequently arrive at varying times due to differences in processing speeds or user behavior. Temporal alignment algorithms, such as dynamic time warping or recurrent neural networks, play a crucial role in ensuring that asynchronous inputs contribute to a unified interpretation. For instance, speech recognition may take longer to process than gesture recognition, necessitating sophisticated synchronization techniques to avoid resynchronization [81–83].

AI and ML have significantly advanced the capabilities of multimodal fusion. DL models, in particular, excel at handling large volumes of heterogeneous data, enabling systems to learn complex patterns and relationships between modalities. These models also resolve conflicts between modalities by dynamically assigning weights based on context and input reliability. For example, in a noisy environment, the system may prioritize gesture inputs over speech to maintain interaction continuity [84,85].

The applications of multimodal fusion are diverse, spanning virtual assistants, VR/AR systems, and smart home environments. In VR, for example, fusion enables users to interact seamlessly with virtual objects through a combination of speech, gaze, and hand gestures. Similarly, in smart home systems, multimodal fusion ensures robust operation by integrating voice commands, touch interactions, and environmental context [86–88].

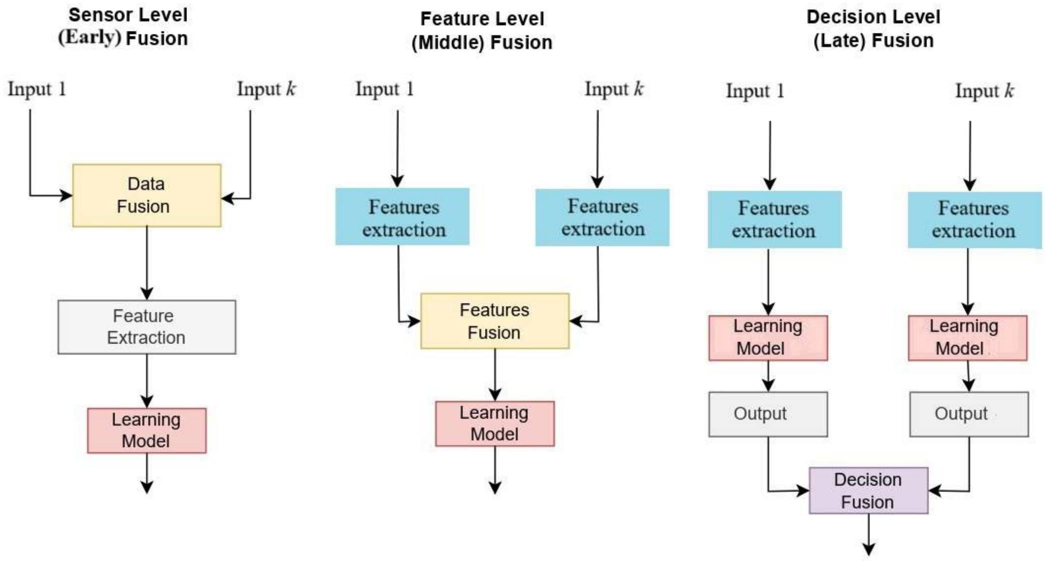

**Figure 2.** An overview of different levels of modalities fusion.

### *3.3. Context-Aware Input Interpretation*

Context-aware input interpretation is a fundamental aspect of multimodal systems, enabling them to adapt dynamically to the user's environment, tasks, and behavior. By incorporating contextual information, these systems can prioritize specific input modalities and adjust their processing strategies to enhance interaction accuracy and efficiency [89,90].

One of the core elements of context-awareness is environmental adaptability. Multimodal systems leverage sensors and algorithms to detect and interpret environmental factors such as noise, lighting, or user location. For example, in a smart home environment, a system may prioritize gesture input over voice commands when it detects high levels of background noise. Similarly, in outdoor settings with limited visibility, systems might emphasize touch or auditory feedback to ensure seamless interaction [91–93].

Another critical aspect is task and user-specific adaptability. Multimodal systems must recognize and respond to the unique requirements of a user's current activity. For instance, in a driving scenario, a system might emphasize voice commands and gaze-based input to minimize manual interaction, ensuring both safety and usability. Over time, ML models can refine this adaptability by learning individual user preferences and tailoring interactions to suit specific habits. For example, a user who frequently combines gestures with speech for smart home commands will experience a system that anticipates and supports this modality combination [94–96].

The ability to handle contextually relevant input extends to managing conflicting or incomplete data. When a user provides ambiguous or partial inputs—such as a vague gesture combined with a spoken command—the system uses contextual clues to infer intent. This process might involve considering the user's physical location, previous interactions, or environmental conditions to resolve ambiguity and deliver the appropriate response [97,98].

However, achieving robust context awareness is not without challenges. Real-time data processing and adaptation require computationally efficient algorithms, especially in mobile or embedded systems. Additionally, ensuring that context-aware interactions remain intuitive and do not overwhelm users with unnecessary complexity is a delicate balance. Effective multimodal systems must seamlessly integrate contextual cues into their input interpretation processes without imposing additional cognitive load on users [99,100].

### *3.4. Multimodal Feedback Loops and Error Handling*

Multimodal systems must use continuous feedback loops to ensure a fluid and adaptive interaction between the user and the system. These feedback loops are critical for refining user inputs, mainly when ambiguities arise. In multimodal communication, users often employ several input modalities in tandem, such as combining speech with gesture or gaze, to convey intent. A key challenge lies in the system's ability to interpret these inputs correctly and provide timely feedback when interpretation errors occur. Multimodal systems use feedback loops to confirm the success or failure of input interpretation to facilitate smoother interactions. For instance, when the system is uncertain about a command in speech recognition, it may prompt the user with clarification questions or offer options, such as "Did you mean A or B"? This process reduces the likelihood of escalating errors, enhancing the system's reliability [101–103].

Effective error handling in multimodal systems is tightly integrated with these feedback loops, enabling dynamic adjustments based on user responses. When multimodal inputs conflict or fail to meet the system's confidence threshold, real-time feedback helps the system and user recover from potential breakdowns. In complex environments like virtual or AR, haptic or visual cues may alert users when their input is misunderstood. For example, if a gesture is misinterpreted in a VR interface, the system might provide subtle visual feedback, such as highlighting the wrong object, prompting the user to modify or repeat the gesture. This ongoing feedback prevents significant disruptions and ensures the system can quickly adapt to fluctuating user behaviours and environmental contexts, creating a more resilient interaction model [104–107].

Additionally, adaptive feedback loops are crucial for learning and improving user interactions over time. Multimodal systems that incorporate ML or AI can use these feedback loops to fine-tune their models based on user behaviour, preferences, and context. For example, if a user frequently corrects a specific type of error in gesture recognition, the system can adjust its algorithms to interpret future inputs better, improving accuracy and efficiency. This type of adaptive error handling increases user satisfaction and allows the system to become more personalized and responsive to individual needs. By continuously learning from user feedback, multimodal systems can evolve to offer more intuitive and seamless interactions, further reducing the cognitive load on users and making human-computer communication more natural and fluid [108–110].

Table 3 provides a structured classification of the key references that address the critical aspects of multimodal communication. These topics include advanced input processing techniques, the complexities of synchronizing multiple data streams, and the development of context-aware systems that adapt dynamically to user behavior and environmental changes. Additionally, the table highlights the challenges and solutions related to output coordination, such as managing multimodal feedback and real-time error handling. This organized overview serves as a guide for understanding how various technical approaches contribute to the efficient processing and synchronization of inputs as well as the refinement of interaction flows.

**Table 3.** Summary of key topics, techniques, and descriptions of multimodal systems, covering input processing, fusion, context-aware interpretation, and feedback mechanisms.

| References | Topics Covered | Techniques/Approaches | Description |
|---|---|---|---|
| [65,66] | Input processing across modalities | Signal processing, NLP for speech, motion capture for gestures | Focuses on preprocessing inputs like speech, gestures, and touch. Describes how signal filtering and noise reduction enable accurate input interpretation. |
| [67–69] | Gesture recognition and variability | Depth-sensing cameras, wearable sensors | Discusses the variability in input signals and the use of gesture recognition to detect meaningful actions from user movement. |
| [70–72] | Robust preprocessing mechanisms | Noise filtering, robust algorithms for variability | Explains techniques to handle inherent variability in speech and gestures, ensuring reliable signal interpretation before fusion. |
| [73–75] | Gesture recognition and cultural differences | Deep neural network (DNN)—body gesture, Convolutional Neural Networks (CNNs)—hand gesture, Viterbi algorithm on Hidden Markov Models (HMM)—hand, arm | Conceptual design and early prototype for real-time translation on intercultural communication; Influence of cultural factors on freehand gesture design and recognition; Real-time rescue gesture recognition for UAV. |
| [76–78] | Fusion of diverse modalities | Decision-level fusion, feature extraction | Integrates speech, gestures, and gaze for a unified interpretation of user intent. Highlights complementary use of modalities to mitigate individual weaknesses. |
| [79,80] | Levels of fusion | Sensor-level, feature-level, decision-level | Explains different fusion strategies, from raw data combination to feature extraction and classifier outputs. |
| [81–83] | Synchronization challenges | Temporal alignment, dynamic time warping | Describes methods to synchronize inputs arriving at different times to ensure coherent interpretation. |

**Table 3.** *Cont.*

| References | Topics Covered | Techniques/Approaches | Description |
|---|---|---|---|
| [84,85] | AI-driven fusion | Context-aware DL models | Explores the use of AI to assign weights dynamically to modalities, prioritizing reliable inputs in specific contexts. |
| [86–88] | Applications in multimodal systems | Virtual assistants, VR/AR, smart homes | Demonstrates applications where fusion enables robust interaction, combining modalities like voice, gaze, and touch. |
| [89,90] | Environmental adaptability | Sensor-based environment detection | Describes how systems prioritize specific modalities (e.g., gestures over voice in noisy environments) using real-time environmental data. |
| [91–93] | User-specific task adaptability | Dynamic input prioritization | Discusses tailoring modalities like voice or gaze to specific tasks, such as driving or hands-free scenarios. |
| [94–96] | Learning user preferences | ML-based adaptive interaction | Explores how systems use ML to recognize user habits and adjust modality prioritization dynamically. |
| [97,98] | Resolving conflicting inputs | Contextual clues and past interactions | Describes techniques for inferring user intent when inputs are ambiguous or incomplete, leveraging context and historical data. |
| [99,100] | Challenges in real-time adaptability | Computational efficiency, seamless transitions | Addresses the challenges of integrating context-awareness into interactions without adding complexity or cognitive load for the user. |
| [101–103] | Feedback mechanisms for input errors | Clarification prompts, redundancy checks | Discusses the importance of real-time feedback in reducing ambiguities in multimodal inputs, ensuring smooth interaction and error reduction. |
| [104–107] | Error recovery and interaction resilience | Adaptive feedback loops, user-specific error handling | Focuses on real-time feedback and error management, using visual, auditory, or haptic cues to alert users about misinterpreted commands. Adaptive systems refine input recognition dynamically. |
| [108–110] | Adaptive learning from errors | ML-driven personalization and error correction | Explains how multimodal systems learn and improve over time through user feedback, reducing repeated errors and creating a more tailored interaction experience. |

## 4. Interaction Modalities

Multimodal systems utilize various interaction channels, each contributing to different aspects of usability and user experience. This section delves into the most prominent modalities and how they are integrated within multimodal systems.

### *4.1. Speech and Natural Language Processing*

Speech is one of the most natural and efficient modes of communication for humans, making it a critical component of multimodal interaction systems. Communicating with computers through speech significantly reduces the need for cumbersome inputs, such as typing or manual manipulation, thus enhancing the overall user experience. Recent advancements in ASR and NLP technologies have made it possible for systems to understand and interpret spoken language with increasing accuracy. State-of-the-art models powered by DL and neural networks are capable of handling complex language tasks, including recognizing accents, parsing colloquialisms, and understanding context-specific nuances. In a multimodal system, speech recognition is not isolated but is often integrated with other input modalities, such as gestures or eye movements, to provide a more robust and contextually aware user experience. This integration enables users to engage with systems more intuitively, allowing for complex, context-rich interactions [111–114].

One of the core challenges in using speech as a modality in multimodal systems is dealing with the variability and ambiguity inherent in natural language. Human speech is far from perfect—filled with disfluencies like pauses, false starts, or filler words that can confuse basic speech recognition systems. To address these challenges, modern NLP techniques utilize advanced language models, such as transformer-based architectures (e.g., GPT, BERT), capable of capturing linguistic patterns and contextual meaning from large datasets. These models improve the accuracy of word recognition and enable deeper semantic understanding, allowing the system to infer user intent more accurately. In multimodal contexts, speech is often disambiguated through other modalities; for example, a gesture or gaze can help clarify a vague spoken command. This complementary use of modalities allows systems to resolve ambiguities more effectively and ensures that users interact naturally without consciously refining their speech to match machine expectations [115–118].

Speech-driven interactions are further enhanced by integrating real-time context awareness, which allows systems to adjust their processing based on environmental factors or user behaviour. Contextual NLP systems can infer meaning from the user's physical environment, task, or previous interactions. For instance, in a smart home setting, a command like "turn it off" could be disambiguated by using information from the environment (e.g., detecting that the lights are currently on) or by integrating other inputs such as gaze direction or gesture. Moreover, speech interfaces are becoming increasingly adaptive, capable of learning user preferences and adjusting over time, which is particularly useful in handling variations in accents, speech patterns, and even emotional tone. As speech and NLP technologies continue to evolve, their integration within multimodal systems will offer even greater personalization and adaptability, pushing the boundaries of HCI into more natural and seamless territories [98,119–121].

### *4.2. Gesture and Motion Sensing*

Gesture and motion sensing technologies have become pivotal components of multimodal interaction systems, offering a natural and intuitive means for users to communicate with machines. Unlike traditional input methods, such as keyboards or touchscreens, gestures allow users to express commands through physical movements, making them especially valuable when hands-free or touchless interaction is essential. The ability of motion sensors, cameras, and depth-sensing devices to capture these movements has enabled a range of applications, from gaming and VR to smart homes and automotive interfaces. The strength of gesture-based interaction lies in its alignment with natural human behaviour; gestures can be expressive, contextually rich, and non-verbal, allowing for more seamless and instinctive interactions. However, this same flexibility introduces significant challenges, as systems must accurately distinguish between intentional gestures and unintentional body movements, often in highly variable environments [122–125].

The core of gesture recognition systems relies on sophisticated computer vision and ML algorithms to interpret user movements' spatial and temporal dynamics. These systems process data from motion sensors, depth cameras, or wearables to detect and classify gestures in real-time. In multimodal systems, gesture input is often combined with other modalities, such as speech, to refine or disambiguate user commands. For example, a user might point to an object while issuing a spoken command, providing the system with additional spatial information that helps clarify the intended action. This synergy between modalities can significantly enhance the accuracy and fluidity of interactions. However, the processing demands of gesture recognition—particularly in real-time applications—require highly efficient algorithms capable of operating under varying conditions, such as different lighting, occlusion, or changes in the user's environment [126–129].

Despite the advances in gesture recognition, the development of gesture-based systems must also address user-centred concerns, such as usability and fatigue. Natural gestures are often imprecise, and users may interpret what constitutes a particular motion differently. Therefore, systems need to be flexible enough to accommodate a wide range of gesture inputs while maintaining high accuracy. Additionally, prolonged or repetitive gesture use can lead to "gorilla arm" syndrome, where users experience fatigue or discomfort after extended periods of gesturing. To counteract this, many systems incorporate adaptive learning mechanisms that tailor the gesture recognition process to individual users, optimizing for comfort and minimizing physical strain. Future research in this field will likely focus on creating more context-aware systems that can seamlessly integrate gesture recognition with other modalities and improve the naturalness and comfort of gesture-based interactions [57,130–132].

### *4.3. Touch and Haptics*

Touch interaction has become a central modality in modern HCI, particularly with the widespread use of touchscreens in smartphones, tablets, and other consumer devices. Touch interfaces enable users to manipulate digital objects directly, providing an intuitive and natural method for interaction. The simplicity of touch gestures—such as tapping, swiping, pinching, and dragging—allows users to control systems without needing physical peripherals. Furthermore, advances in multi-touch technologies have enabled complex gestures, expanding the possibilities for interaction by recognizing multiple points of contact. However, despite its intuitive nature, touch interaction can sometimes suffer from limitations such as occlusion, where the user's finger blocks the view of the interface, or the lack of tactile feedback on flat screen surfaces, leading to less satisfying user experiences [133–136].

To address the limitations of traditional touch interaction, haptic feedback has emerged as a crucial complement. Haptics uses tactile sensations to enhance interaction by providing physical feedback to the user. Through vibrations, pressure, or even force resistance, haptic technology creates the illusion of interacting with real-world objects, thus increasing the user's sense of control and immersion. In mobile devices, for example, subtle vibrations can confirm touch inputs, improving the accuracy of interactions and reducing errors. In gaming and VR environments, more advanced haptic systems can simulate textures, collisions, or resistance forces, making the virtual world feel more tangible and responsive. These tactile cues are particularly valuable in scenarios where visual or auditory feedback might be insufficient or inappropriate, such as in high-speed or visually demanding tasks [137–140].

Combining touch and haptic feedback in multimodal systems creates a powerful tool for enhancing user interaction. When paired with visual or auditory feedback, haptics can reinforce actions and provide discreet, contextually appropriate responses that guide the user without overwhelming them. For example, in an automotive interface, haptic feedback might alert the driver to changes in the vehicle's state without requiring them to take their eyes off the road. In medical training simulations, haptic technology can replicate the sensation of performing physical tasks, such as inserting a needle or making an incision, enabling realistic skill development in a safe environment. As haptic technology advances, its integration into multimodal systems will likely expand, providing more prosperous, immersive, and practical interaction experiences across various applications [141–143].

Finally, touchless haptic feedback, such as air jets and ultrasound-based systems, extends the possibilities of haptic interactions by providing tactile sensations without physical contact. These technologies enable a new level of interaction, particularly in scenarios requiring hygienic or hands-free operations while maintaining high precision and user immersion [144–146].

### *4.4. Eye-Tracking and Gaze-Based Interaction*

Eye-tracking and gaze-based interaction represent a powerful modality in multimodal systems, offering an efficient and intuitive input form. Eye-tracking systems enable hand-free control by monitoring where users are looking, providing seamless interaction with digital environments, particularly in scenarios where other modalities may be limited or impractical. This technology leverages infrared cameras or optical sensors to detect the movement and focus of a user's eyes, allowing systems to interpret gaze direction as input. When combined with other modalities, such as speech or gesture, gaze-based interaction enables highly contextual and precise interface control. For instance, a user could look at an object on a screen and issue a voice command to interact with that object, making the interaction more intuitive and reducing the need for explicit selection through touch or mouse inputs [77,147–149].

One of the significant benefits of gaze-based interaction lies in its ability to provide highly adaptive and context-aware interactions. In AR and VR, eye-tracking enhances immersion by allowing users to naturally explore and interact with the virtual space without relying solely on physical movements or controllers. By interpreting where users are looking, the system can anticipate their intentions and adjust the interface accordingly, such as highlighting objects of interest or changing the focus of the display. Additionally, gaze tracking can be particularly beneficial for accessibility applications, offering a robust alternative for individuals with motor impairments who cannot use traditional input methods like keyboards or touchscreens [150–153].

However, despite the promise of eye-tracking technology, significant challenges still need to be addressed in implementing it effectively within multimodal systems. Accuracy and precision are critical for successful interaction, but factors such as lighting conditions, user eye physiology, and movement can affect the reliability of eye-tracking sensors. Moreover, gaze-based systems need to distinguish between deliberate and passive eye movements, as not all glances or eye movements are intended to trigger actions. Sophisticated algorithms are required to filter out unintentional eye movements and improve the precision of gaze tracking. Furthermore, ethical concerns arise regarding privacy, as eye-tracking data can reveal sensitive information about users' focus and interests. Addressing these challenges requires ongoing advancements in sensor technology, algorithmic development, and privacy safeguards to ensure that gaze-based interaction becomes a reliable and integral part of multimodal systems [154–157].

Table 4 offers a breakdown of the most significant interaction modalities discussed in this section. It categorizes key studies that examine the integration of speech and gesture, haptic feedback, and gaze-based interfaces, along with their applications in fields like VR/AR. The table also highlights the role of real-time synchronization and AI-driven adaptation, which are critical to improving the fluidity and precision of interactions. This classification is a clear reference point for understanding how different modalities are combined to create richer, more immersive interactions in modern systems.

**Table 4.** Classification of references on interaction modalities highlighting their integration and applications in immersive environments.

| References | Topics Covered | Techniques/Approaches | Applications |
|---|---|---|---|
| [98,111–121] | Multimodal interaction in visual and audio systems, indigenous language speech recognition, hybrid fusion models, gesture-speech integration | Hybrid fusion models and gesture-speech interaction techniques | Effective multimodal systems addressing linguistic diversity and integration challenges. |
| [57,122–132] | HCI, multimodal communication in learning | Audio-visual integration for communication and learning | In educational and medical settings with challenges in data synchronization. |
| [133–146] | Gesture and haptic technologies in HCI, Feedback systems, Wearable tactile interfaces and surface haptics, Immersive interaction in VR/AR, Multimodal interfaces, Usability enhancement and adaptive interaction paradigms | DL for gesture recognition, Electrovibration, air-jet, ultrasonic arrays, Predictive interaction models, Dynamic tactile feedback systems, Surface and contactless haptics | HCI evaluation, Immersive AR/VR and gaming, Automotive interfaces, Surgical robotics and training, Public and touchless displays, Usability in aerial and tactile systems |
| [77,147–153] | Eye-tracking and gaze-based interfaces, real-time interaction in AR, event cameras for face and eye detection, multimodal analytics | Event-based cameras and gaze-tracking techniques | Real-time interaction systems for AR with challenges in accuracy, precision, and privacy. |
| [154–157] | Advanced sensory modalities, real-time analytics | Sensor fusion and dynamic feedback loops | Adaptive environments leveraging multimodal analytics with high computational overhead challenges. |

## 5. Challenges in Multimodal Interaction Systems

Despite significant advancements, multimodal interaction systems face several challenges that complicate their design and implementation. These issues arise from the need to integrate diverse modalities, ensure seamless synchronization, and adapt dynamically to changing user and environmental contexts. This section explores the key obstacles in developing robust, efficient, and user-friendly multimodal systems, emphasizing areas requiring further research and innovation.

### *5.1. Modality Fusion and Synchronization*

The integration of multiple input modalities is a cornerstone of multimodal systems, yet it presents a range of technical challenges. Modality fusion, the process of combining inputs from various sensory channels, must be performed seamlessly to ensure a coherent and responsive interaction experience. However, differences in input timing, accuracy, and the complementary or redundant nature of information across modalities often complicate this process [158,159].

One critical challenge is temporal synchronization. Inputs from modalities such as speech and gesture frequently arrive at different times due to variations in processing speeds or user behavior. For example, speech recognition might require more time to interpret a verbal command compared to the instantaneous detection of a gesture. Synchronizing these inputs to create a unified understanding of the user's intent demands advanced algorithms, such as dynamic time warping or probabilistic models, to align the inputs accurately without introducing latency [160–162].

Another aspect of modality fusion involves resolving conflicts between inputs. When modalities provide contradictory information—such as a gesture pointing to one object while speech refers to another—the system must determine which input to prioritize. ML models play a critical role in assigning weights to inputs based on contextual factors, enhancing decision-making and overall system reliability [163,164].

In addition to addressing conflicts and synchronization, robust error-handling mechanisms are essential. Multimodal systems must degrade gracefully when certain modalities fail or when inputs are noisy or incomplete. For example, a system might need to rely more heavily on gesture recognition if speech processing becomes unreliable. Effective fusion models should account for these scenarios and adjust dynamically to maintain interaction continuity [165,166].

### *5.2. Context Awareness and Adaptability*

Context awareness is fundamental to the success of multimodal systems, enabling them to adapt dynamically to the user's environment, tasks, and preferences. By leveraging contextual information, these systems can determine which modalities to prioritize and how to interpret inputs effectively. This adaptability ensures seamless interactions, particularly in scenarios where user needs or environmental conditions are constantly changing [167,168].

An essential aspect of context awareness is the system's ability to adjust its behavior based on the user's environment. For instance, in a driving scenario, the system might prioritize voice commands and eye-tracking to minimize the need for manual interaction, ensuring safety and usability. Similarly, in a hands-free environment, gestures might take precedence over touch-based inputs, offering a more natural interaction experience [169,170].

Adaptability also extends to understanding individual user preferences and behaviors. Users may favor specific modalities depending on their tasks or habits, and these preferences can vary over time. A robust multimodal system leverages ML models to learn and predict these preferences, tailoring the interaction accordingly. For instance, a user who frequently uses voice commands on a smart home system might experience a more responsive interface that prioritizes speech inputs while retaining other modalities as complementary options [171–173].

Despite its benefits, achieving context-aware adaptability is not without challenges. Continuous monitoring of environmental and user data can be computationally intensive, particularly in real-time applications. Systems must strike a balance between responsiveness and resource efficiency to remain practical across various platforms, including mobile and embedded devices. Advanced AI algorithms and sensor technologies are critical for optimizing context awareness while minimizing computational overhead [174,175].

### *5.3. Multimodal Adaptation in Noisy Environments*

Adapting multimodal systems to noisy environments is a critical challenge that significantly impacts their usability and performance. In scenarios where ambient noise interferes with speech recognition, such as in busy public spaces, kitchens, or industrial settings, alternative modalities like gestures, gaze, or touch must be prioritized to maintain seamless interaction. Effective adaptation in such conditions involves a combination of dynamic modality switching, redundancy in input channels, and advanced environmental awareness [176–178].

Dynamic modality switching is a key strategy for handling noisy environments. For instance, when speech input becomes unreliable due to high levels of acoustic interference, the system can automatically transition to gesture recognition. This transition requires real-time detection of environmental noise levels and a contextual understanding of the interaction. By employing algorithms capable of analyzing audio signals and predicting noise conditions, the system can proactively adapt its input modality to ensure uninterrupted communication [179–181].

In addition to switching between modalities, integrating multiple input channels can enhance robustness. Combining gesture recognition with gaze tracking or touch-based interactions provides redundancy, reducing the likelihood of errors caused by noisy conditions. For example, a user gesturing toward an object while gazing at it offers complementary information that the system can process to clarify intent. This multimodal integration not only compensates for individual modality failures but also enhances overall system reliability [182,183].

Adapting to noisy environments also demands sophisticated algorithms for real-time data fusion and synchronization. As inputs from different modalities may arrive asynchronously, systems must align them temporally to create a coherent understanding of user intent. Temporal alignment techniques, such as dynamic time warping or recurrent neural networks, are often employed to achieve this synchronization without introducing latency [184,185].

Despite these advancements, multimodal adaptation in noisy environments faces several challenges. The computational overhead of monitoring environmental conditions and processing multimodal inputs in real time can strain system resources, particularly in mobile or embedded platforms. Moreover, ensuring smooth transitions between modalities without disrupting user experience requires precise and efficient design [186,187].

### *5.4. User-Centric Design and Usability*

The success of any multimodal interaction system depends on its usability and how well it aligns with user expectations and cognitive capabilities. While introducing multiple modalities can enhance flexibility and provide alternative input methods, there is also a risk of overwhelming users with too many options. A well-designed multimodal system must ensure that the interaction remains intuitive and does not burden the user with unnecessary complexity. This requires a careful balance between offering diverse input channels and maintaining simplicity in the user interface [188–191].

Cognitive load is a significant concern in multimodal systems. When users are required to manage multiple input modalities simultaneously, it can lead to confusion, frustration, and reduced task performance. To mitigate this, multimodal systems should be designed to guide users naturally from one modality to another, offering clear cues and feedback. For example, if a user begins a task using speech, the system might provide a subtle visual cue or auditory confirmation that the command has been understood, allowing the user to shift seamlessly between modalities without cognitive strain. User studies and usability testing are critical in identifying friction points in the interaction flow and optimizing the system accordingly [192–195].

Personalization is another critical factor in user-centric design. Users have different preferences, abilities, and interaction styles, and a successful multimodal system must be able to adapt to these differences. Personalization can be achieved through ML models that analyze a user's past interactions and preferences, adjusting the interface to suit their needs better. For example, a system might learn that a particular user prefers touch over voice commands and modify its interface accordingly. However, implementing personalization in a way that respects user privacy and avoids intrusive behaviour is an ongoing challenge. Balancing the system's adaptability with respect for user autonomy and data security is crucial for building trust and ensuring long-term user engagement with multimodal systems [196–199].

Table 5 recaps the critical challenges discussed in this section. The listed references cover essential topics like modality fusion, synchronization, and multimodal conversations, with a focus on real-time systems. The table also highlights efforts in integrating speech, gesture, and other inputs for more natural, intuitive interactions. These studies explore solutions such as DL models for modality fusion, algorithms for temporal synchronization, and practical systems for multimodal conversations. Together, this body of research outlines the current state of multimodal systems and their ability to combine diverse inputs into seamless, coherent user interactions.

**Table 5.** Summary of key topics and techniques in multimodal interaction systems.

| References | Topics Covered | Techniques/ Approaches | Description |
|---|---|---|---|
| [158,159] | Temporal synchronization and input fusion | Dynamic time warping, probabilistic models | Discusses challenges in aligning asynchronous inputs from different modalities (e.g., speech vs. gesture) and techniques to ensure seamless integration. |
| [160–162] | Conflict resolution in multimodal inputs | ML models for contextual weighting | Explores resolving contradictory inputs from modalities by assigning weights based on context, enhancing decision-making and overall system reliability. |
| [163,164] | Error handling in multimodal fusion | Adaptive fusion models, fallback mechanisms | Focuses on systems that adjust dynamically when inputs fail or are noisy, relying on alternate modalities to maintain continuous interaction. |
| [165,166] | Robustness and adaptability in modality integration | Context-aware algorithms, redundancy in input channels | Highlights the importance of redundancy and adaptive strategies to ensure robust operation in noisy or variable environments, enhancing system reliability. |
| [167,168] | Context-aware haptic feedback | Retargeting self-haptics, dynamic response | Explores haptic feedback's role in adapting to user environments, enhancing immersion and interaction reliability, particularly in VR/AR systems. |
| [169,170] | Sensor integration for adaptability | Ultrasonic sensors, electrovibration feedback | Describes sensor-based enhancements for adaptability in multimodal systems, focusing on seamless integration and responsive feedback. |
| [171–173] | Adaptive feedback for critical applications | Real-time force feedback, AI-driven context adaptation | Highlights adaptive feedback mechanisms in surgical training and critical systems, ensuring precise interaction and skill development in dynamic scenarios. |
| [174,175] | Gaze-based interaction and context adaptation | Event-based gaze tracking, contextual prioritization | Investigates how eye-tracking and gaze technologies adapt dynamically to user contexts, improving responsiveness and interaction accuracy in smart systems. |
| [176–178] | Context-awareness in adaptive systems | AI-based context modeling, multi-sensor integration | Examines methods for leveraging AI and multi-sensor data to enhance context-awareness and system adaptability in diverse interaction environments. |
| [179–181] | Multimodal interaction in dynamic environments | Real-time fusion, environment-specific prioritization | Focuses on strategies for adapting to dynamically changing environments by prioritizing relevant modalities and maintaining interaction fluidity. |
| [182,183] | Redundancy and robustness in noisy settings | Multi-modal redundancy, failure compensation | Discusses redundancy techniques to compensate for modality failures in noisy or unpredictable environments, ensuring interaction continuity and reliability. |
| [184,185] | Synchronization in complex multimodal systems | Probabilistic models, temporal alignment algorithms | Highlights challenges in synchronizing inputs across modalities and presents solutions like probabilistic models and advanced temporal alignment techniques. |
| [186,187] | Decision-making in modality conflicts | Context-aware weighting, adaptive decision frameworks | Explores frameworks for resolving modality conflicts using contextual weighting and real-time decision-making algorithms to improve system responsiveness. |
| [188–191] | Dynamic error handling in multimodal systems | Error recovery algorithms, real-time feedback loops | Explores approaches to dynamically handle errors in multimodal interactions by employing feedback loops and recovery mechanisms for robust system operation. |

**Table 5.** *Cont.*

| References | Topics Covered | Techniques/ Approaches | Description |
|---|---|---|---|
| [192–195] | Adaptive interaction in real-time environments | ML, context-sensitive adaptation | Focuses on leveraging ML to adapt system responses dynamically based on real-time environmental and contextual changes, improving interaction efficiency. |
| [196–199] | Personalization in multimodal interfaces | User-specific modeling, preference learning | Highlights techniques to tailor multimodal systems to individual user preferences, ensuring usability and satisfaction through adaptive interaction models. |

## 6. Applications and Future Directions

The rise of multimodal interaction systems has significantly broadened the scope of applications across various industries, revolutionizing how humans interact with technology. By enabling systems to interpret and integrate multiple input modalities—such as speech, touch, gesture, and gaze—multimodal interfaces offer more natural, flexible, and efficient forms of interaction. From healthcare to VR, multimodal systems transform user experiences by enhancing accessibility, personalization, and adaptability. As the field continues to evolve, future developments are expected to push the boundaries of HCI, incorporating advanced ML, AI, and emerging sensory technologies. This section will explore both current applications of multimodal systems and the future directions that promise to shape the next generation of interactive technologies.

### *6.1. Applications of Multimodal Systems*

Multimodal systems play a transformative role across diverse industries, including healthcare, education, VR and AR, intelligent environments, and accessibility technologies. In healthcare, for instance, multimodal interfaces facilitate human-machine communication in critical situations, such as during surgeries or patient monitoring. Voice commands, gesture recognition, and touch-based inputs allow surgeons to control medical equipment without physical contact, maintaining sterile environments while enhancing precision and efficiency. In inpatient rehabilitation, multimodal systems incorporating haptics, speech, and visual feedback are used to develop personalized therapy plans, enabling more effective monitoring of motor skills and progress [200–204].

In VR and AR, multimodal interaction is essential for creating immersive environments. VR/AR systems leverage speech, gestures, and eye-tracking to allow users to navigate virtual spaces seamlessly and interact naturally with digital objects. These applications are critical for entertainment and training simulations, education, and professional development. For instance, in military or pilot training, multimodal VR systems provide realistic scenarios that closely mimic real-world conditions, allowing trainees to interact with their environments using multiple modalities simultaneously, thereby improving learning outcomes and engagement [205–209].

Intelligent environments, such as smart homes and smart cities, are another area where multimodal systems are making an impact. Voice-activated virtual assistants, like Amazon Alexa or Google Assistant, combine voice recognition with other modalities, such as gesture or touch, to control appliances, provide information, and enhance convenience. These systems can also integrate with wearable technology, using multimodal feedback to offer real-time updates on health, traffic, or environmental conditions. Accessibility technology has similarly benefited from multimodal systems, offering innovative solutions for individuals with disabilities. By combining voice recognition, gaze tracking, and haptics, these systems enable individuals with mobility impairments to interact with computers, smartphones, and other digital devices in previously inaccessible ways [210–215].

Table 6 lists references covering advances in multimodal systems applied to biomedical AI, smart gloves for rehabilitation, AR/VR in medicine, and cognitive and emotional engagement in VR. Additionally, it highlights work in command and control systems using AR, gesture recognition in smart cities, and wearable sensors for real-time interaction in AR/VR. Finally, these studies illustrate the versatility of multimodal systems, showcasing how they enhance user experiences across diverse domains by improving interaction accuracy, real-time feedback, and context-aware adaptation.

**Table 6.** Key applications of multimodal systems across diverse domains.

| References | Topics Covered | Techniques/Approaches | Applications/Use Cases & Challenges |
|---|---|---|---|
| [200–204] | Multimodal biomedical AI, Smart gloves, AR/VR in medicine | Smart tactile gloves for rehabilitation, AR/VR for surgical training, multimodal biomedical AI techniques | Enhances medical outcomes through immersive AR/VR systems and personalized rehabilitation. Challenges include sensor precision and real-time adaptability. |
| [205–209] | Command and control systems with AR, cognitive and emotional engagement in VR | AR-based systems for command control, VR environments for cognitive engagement | Improves user training and decision-making through AR/VR systems with emotional and cognitive engagement. Challenges include creating realistic and adaptive scenarios. |
| [210–215] | Voice-controlled drones, gesture recognition in smart cities, wearable sensors in AR/VR applications | Gesture recognition algorithms, wearable sensor integration for real-time interaction | Applications include voice-controlled drones, AR/VR feedback systems, and gesture-based interfaces for smart cities. Scalability and integration complexity remain challenges. |

### *6.2. Future Directions of Multimodal Systems*

The future of multimodal interaction lies in creating even more adaptive, intelligent, and context-aware systems. One of the most promising directions is the integration of AI and ML to enhance the adaptability of multimodal systems. AI-driven models will allow systems to learn from user behaviour, refining their ability to predict user intent and provide personalized interactions. As AI technologies mature, multimodal systems will become more autonomous, anticipating user needs and responding dynamically to environmental changes. For example, in smart home systems, future multimodal interfaces could adapt in real-time to lighting, sound, or user activity changes, automatically adjusting inputs and outputs to optimize the user experience [216–220].

Emerging technologies, such as BCIs, will likely become integral to future multimodal systems. BCIs offer the potential to capture neural activity directly from the brain, creating a new modality that allows users to control systems using thought alone. When integrated with existing modalities such as speech, gesture, and gaze, BCIs could enable profoundly immersive experiences in fields such as VR/AR, gaming, and accessibility. For example, BCIs could enhance accessibility for individuals with severe physical disabilities, allowing them to control devices or communicate through a combination of brain signals and other multimodal inputs [221–225].

Another exciting area of development is the enhancement of context-awareness in multimodal systems. Current systems often need help recognizing and adapting to complex, dynamic environments. In the future, multimodal interfaces will be capable of analyzing environmental data in real-time, using sensors, AI, and ML to understand the user's context and deliver contextually relevant feedback. This could be particularly valuable in industries such as automotive, where multimodal systems could combine voice commands, gesture input, and eye-tracking to provide real-time driver assistance while dynamically adjusting to traffic conditions, weather, and driver behaviour [226–230].

In summary, the future of multimodal systems will be characterized by increasingly sophisticated interactions driven by AI, ML, and new sensory technologies like BCIs. These developments will enhance user experience across a broad spectrum of applications and push the boundaries of what is possible in HCI. As multimodal systems evolve, their ability to provide natural, intuitive, and adaptive interactions will continue to grow, further bridging the gap between humans and machines [231–233]. Table 7 outlines the future paths in advancing multimodal systems. These studies emphasize the potential of AI to enhance the adaptability and personalization of multimodal interfaces, and BCIs and advanced sensory technologies to push the boundaries of HCI into entirely new dimensions.

**Table 7.** Emerging technologies and future directions in multimodal systems.

| References | Topics Covered | Techniques/Approaches | Applications/ Use Cases & Challenges |
|---|---|---|---|
| [216–220] | AI and ML for adaptive multimodal systems, real-time context adaptation | AI-driven context-aware models for real-time interaction and learning | Enhances smart home systems with adaptive and personalized multimodal interactions. Challenges include ensuring real-time adaptability and minimizing computational overhead. |
| [221–225] | BCIs & immersive technologies | BCIs integrated with gesture and speech recognition | Improves accessibility and immersive experiences in AR/VR for individuals with disabilities. Challenges include user comfort and seamless integration of modalities. |
| [226–230] | Advanced context awareness in dynamic environments | AI-driven voice, gesture, and eye-tracking integration for real-time assistance | Applications in automotive and smart environments with challenges in dynamic context management and multimodal synchronization. |
| [231–233] | Emerging sensory technologies in HCI | Innovative sensory interfaces, such as taste-based BCIs and advanced perception techniques | Expands sensory studies and multimodal scenarios, addressing emerging interaction paradigms. Challenges include developing practical and user-friendly implementations. |

## 7. Conclusions

Multimodal interaction is rapidly transforming the landscape of HCI, allowing for more natural, intuitive, and efficient engagement between users and digital systems. By leveraging multiple input and output modalities—such as speech, gestures, touch, gaze, and haptics—multimodal systems offer a robust and flexible alternative to traditional unimodal interfaces. This survey has examined the key aspects of multimodal interaction, including integrating various modalities, the technical challenges of modality fusion, and the importance of context awareness in ensuring seamless interactions. Fusing these diverse inputs allows systems to enhance user experience by offering more adaptable, contextually appropriate responses that closely mimic human communication patterns.

Despite the potential of multimodal systems, several challenges persist. Achieving efficient modality fusion and temporal synchronization remains a complex task, requiring sophisticated algorithms and ML models to ensure that different modalities complement each other rather than conflict. Additionally, the system's ability to adapt to dynamic environments, user preferences, and varying contexts is essential for the continued success of multimodal interaction. Privacy concerns, particularly in modalities like gaze tracking and voice recognition, must also be addressed to safeguard user data. Overcoming these challenges will require ongoing research and innovation, particularly in areas such as AI and ML, which are critical in making multimodal systems more intelligent and adaptive.

Looking ahead, the future of multimodal interaction promises to be shaped by emerging technologies such as BCIs, AI-driven personalization, and increasingly immersive virtual and AR environments. These developments will push the boundaries of HCI, allowing systems to learn from user behaviour, adapt in real-time, and anticipate user needs more accurately than ever before. As the field advances, the convergence of multimodal systems with AI will unlock new possibilities in fields ranging from healthcare and education to intelligent environments and accessibility, ultimately transforming how humans interact with machines and digital environments globally.

**Author Contributions:** Conceptualization, E.D. and M.T.; Methodology, E.D.; Formal analysis and Investigation, E.D. and M.T.; Writing—original draft, E.D. and M.T.; Writing—review & editing, M.T.; Visualization, M.T.; Supervision, P.M. and C.T.; Funding acquisition, C.T. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Jia, J.; He, Y.; Le, H. A multimodal human-computer interaction system and its application in smart learning environments. In Proceedings of the Blended Learning. Education in a Smart Learning Environment: 13th International Conference, ICBL 2020, Bangkok, Thailand, 24–27 August 2020; Proceedings 13; Springer: Berlin/Heidelberg, Germany, 2020; pp. 3–14.
2. Šumak, B.; Brdnik, S.; Pušnik, M. Sensors and artificial intelligence methods and algorithms for human–computer intelligent interaction: A systematic mapping study. *Sensors* **2021**, *22*, 20. [CrossRef] [PubMed]
3. Garg, M.; Wazarkar, S.; Singh, M.; Bojar, O. Multimodality for NLP-centered applications: Resources, advances and frontiers. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022; pp. 6837–6847.
4. Papadopoulos, T.; Evangelidis, K.; Kaskalis, T.H.; Evangelidis, G.; Sylaiou, S. Interactions in augmented and mixed reality: An overview. *Appl. Sci.* **2021**, *11*, 8752. [CrossRef]
5. Darin, T.; Andrade, R.; Sánchez, J. Usability evaluation of multimodal interactive virtual environments for learners who are blind: An empirical investigation. *Int. J. Hum.-Comput. Stud.* **2022**, *158*, 102732. [CrossRef]
6. Luo, Y.; Liu, F.; She, Y.; Yang, B. A context-aware mobile augmented reality pet interaction model to enhance user experience. *Comput. Animat. Virtual Worlds* **2023**, *34*, e2123. [CrossRef]
7. Yang, M.; Gao, Y.; Tang, L.; Hou, J.; Hu, B. Wearable eye-tracking system for synchronized multimodal data acquisition. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *34*, 5146–5159. [CrossRef]
8. Garagić, D.; Pelgrift, D.; Peskoe, J.; Hagan, R.D.; Zulch, P.; Rhodes, B.J. Machine Learning Multi-Modality Fusion Approaches Outperform Single-Modality & Traditional Approaches. In Proceedings of the 2021 IEEE Aerospace Conference (50100), Big Sky, MT, USA, 6–13 March 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–9.
9. Paplu, S.; Ahmed, H.; Ashok, A.; Akkus, S.; Berns, K. Multimodal Perceptual Cues for Context-Aware Human-Robot Interaction. In Proceedings of the IFToMM International Symposium on Science of Mechanisms and Machines (SYROM), Iasi, Romania, 17–18 November 2022; Springer: Cham, Switzerland, 2022; pp. 283–294.
10. Jaimes, A.; Sebe, N. Multimodal human–computer interaction: A survey. *Comput. Vis. Image Underst.* **2007**, *108*, 116–134. [CrossRef]
11. Oviatt, S. Multimodal interaction, interfaces, and analytics. In *Handbook of Human Computer Interaction*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 1–29.
12. Sebe, N. Multimodal interfaces: Challenges and perspectives. *J. Ambient. Intell. Smart Environ.* **2009**, *1*, 23–30. [CrossRef]
13. Kim, S.; Billinghurst, M.; Kim, K. Multimodal interfaces and communication cues for remote collaboration. *J. Multimodal User Interfaces* **2020**, *14*, 313–319. [CrossRef]
14. Karpov, A.; Yusupov, R. Multimodal interfaces of human–computer interaction. *Her. Russ. Acad. Sci.* **2018**, *88*, 67–74. [CrossRef]
15. Elouali, N.; Rouillard, J.; Le Pallec, X.; Tarby, J.C. Multimodal interaction: A survey from model driven engineering and mobile perspectives. *J. Multimodal User Interfaces* **2013**, *7*, 351–370. [CrossRef]
16. Dumas, B.; Lalanne, D.; Oviatt, S. Multimodal interfaces: A survey of principles, models and frameworks. In *Human Machine Interaction: Research Results of the MMI Program*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 3–26.
17. Azofeifa, J.D.; Noguez, J.; Ruiz, S.; Molina-Espinosa, J.M.; Magana, A.J.; Benes, B. Systematic review of multimodal human–computer interaction. *Informatics* **2022**, *9*, 13. [CrossRef]
18. Liang, R.; Liang, B.; Wang, X.; Zhang, T.; Li, G.; Wang, K. A Review of Multimodal Interaction. In Proceedings of the International Conference on Education, Management, Computer and Society, Shenyang, China, 1–3 January 2016; Atlantis Press: Amsterdam, The Netherlands, 2016; pp. 711–715.

19. Caschera, M.C.; Ferri, F.; Grifoni, P. Multimodal interaction systems: Information and time features. *Int. J. Web Grid Serv.* **2007**, *3*, 82–99. [CrossRef]
20. Turk, M. Multimodal interaction: A review. *Pattern Recognit. Lett.* **2014**, *36*, 189–195. [CrossRef]
21. Yin, R.; Wang, D.; Zhao, S.; Lou, Z.; Shen, G. Wearable sensors-enabled human–machine interaction systems: From design to application. *Adv. Funct. Mater.* **2021**, *31*, 2008936. [CrossRef]
22. Funk, M.; Tobisch, V.; Emfield, A. Non-verbal auditory input for controlling binary, discrete, and continuous input in automotive user interfaces. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; pp. 1–13.
23. Wang, D.; Zhao, T.; Yu, W.; Chawla, N.V.; Jiang, M. Deep multimodal complementarity learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *34*, 10213–10224. [CrossRef]
24. Mai, S.; Zeng, Y.; Hu, H. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Trans. Multimed.* **2022**, *25*, 4121–4134. [CrossRef]
25. Lee, M.; Révész, A. Promoting grammatical development through captions and textual enhancement in multimodal input-based tasks. *Stud. Second. Lang. Acquis.* **2020**, *42*, 625–651. [CrossRef]
26. Standen, P.J.; Brown, D.J.; Taheri, M.; Galvez Trigo, M.J.; Boulton, H.; Burton, A.; Hallewell, M.J.; Lathe, J.G.; Shopland, N.; Blanco Gonzalez, M.A.; et al. An evaluation of an adaptive learning system based on multimodal affect recognition for learners with intellectual disabilities. *Br. J. Educ. Technol.* **2020**, *51*, 1748–1765. [CrossRef]
27. Stefanidi, Z.; Margetis, G.; Ntoa, S.; Papagiannakis, G. Real-time adaptation of context-aware intelligent user interfaces, for enhanced situational awareness. *IEEE Access* **2022**, *10*, 23367–23393. [CrossRef]
28. Lagomarsino, M.; Lorenzini, M.; De Momi, E.; Ajoudani, A. An online framework for cognitive load assessment in industrial tasks. *Robot.-Comput.-Integr. Manuf.* **2022**, *78*, 102380. [CrossRef]
29. Rasenberg, M.; Özyürek, A.; Dingemanse, M. Alignment in multimodal interaction: An integrative framework. *Cogn. Sci.* **2020**, *44*, e12911. [CrossRef] [PubMed]
30. Chen, S.; Epps, J. Multimodal coordination measures to understand users and tasks. *ACM Trans.-Comput.-Hum. Interact. (TOCHI)* **2020**, *27*, 1–26. [CrossRef]
31. Hoggan, E. Multimodal Interaction. In *Interaction Techniques and Technologies in Human-Computer Interaction*; CRC Press: Boca Raton, FL, USA, 2024; pp. 45–63.
32. Seinfeld, S.; Feuchtner, T.; Maselli, A.; Müller, J. User representations in human-computer interaction. *Hum.–Comput. Interact.* **2021**, *36*, 400–438. [CrossRef]
33. Li, J. Recent advances in end-to-end automatic speech recognition. *APSIPA Trans. Signal Inf. Process.* **2022**, *11*, e8. [CrossRef]
34. Park, K.B.; Choi, S.H.; Lee, J.Y.; Ghasemi, Y.; Mohammed, M.; Jeong, H. Hands-free human–robot interaction using multimodal gestures and deep learning in wearable mixed reality. *IEEE Access* **2021**, *9*, 55448–55464. [CrossRef]
35. Gibbs, J.K.; Gillies, M.; Pan, X. A comparison of the effects of haptic and visual feedback on presence in virtual reality. *Int. J. Hum.-Comput. Stud.* **2022**, *157*, 102717. [CrossRef]
36. Wachowiak, L.; Tisnikar, P.; Canal, G.; Coles, A.; Leonetti, M.; Celiktutan, O. Predicting When and What to Explain From Multimodal Eye Tracking and Task Signals. *IEEE Trans. Affect. Comput.* **2024**, 1–12. [CrossRef]
37. Huang, Y.; Yao, K.; Li, J.; Li, D.; Jia, H.; Liu, Y.; Yiu, C.K.; Park, W.; Yu, X. Recent advances in multi-mode haptic feedback technologies towards wearable interfaces. *Mater. Today Phys.* **2022**, *22*, 100602. [CrossRef]
38. Cao, L.; Zhang, H.; Peng, C.; Hansberger, J.T. Real-time multimodal interaction in virtual reality-a case study with a large virtual interface. *Multimed. Tools Appl.* **2023**, *82*, 25427–25448. [CrossRef]
39. Pezent, E.; Gupta, A.; Duhaime, H.; O'Malley, M.; Israr, A.; Samad, M.; Robinson, S.; Agarwal, P.; Benko, H.; Colonnese, N. Explorations of wrist haptic feedback for AR/VR interactions with Tasbi. In Proceedings of the Adjunct Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology, Bend, OR, USA, 29 October–2 November 2022; pp. 1–5.
40. Triantafyllidis, E.; Mcgreavy, C.; Gu, J.; Li, Z. Study of multimodal interfaces and the improvements on teleoperation. *IEEE Access* **2020**, *8*, 78213–78227. [CrossRef]
41. Gong, R.; Hua, M. Designing multimodal user interfaces for hybrid collaboration: A user-centered approach. In Proceedings of the International Conference on Human-Computer Interaction, Copenhagen, Denmark, 23–28 July 2023; Springer: Cham, Switzerland, 2023; pp. 67–82.
42. Su, C.; Yang, C.; Chen, Y.; Wang, F.; Wang, F.; Wu, Y.; Zhang, X. Natural multimodal interaction in immersive flow visualization. *Vis. Inform.* **2021**, *5*, 56–66. [CrossRef]
43. Schiavo, G.; Mich, O.; Ferron, M.; Mana, N. Trade-offs in the design of multimodal interaction for older adults. *Behav. Inf. Technol.* **2022**, *41*, 1035–1051. [CrossRef]
44. Vanneste, P.; Raes, A.; Morton, J.; Bombeke, K.; Van Acker, B.B.; Larmuseau, C.; Depaepe, F.; Van den Noortgate, W. Towards measuring cognitive load through multimodal physiological data. *Cogn. Technol. Work.* **2021**, *23*, 567–585. [CrossRef]

45. Chan, E.; Chan, G.; Kroma, A.; Arya, A. Holistic multimodal interaction and design. In Proceedings of the International Conference on Human-Computer Interaction, Virtual, 26 June–1 July 2022; Springer: Cham, Switzerland, 2022; pp. 18–33.
46. Liu, K.; Xue, F.; Guo, D.; Wu, L.; Li, S.; Hong, R. MEGCF: Multimodal entity graph collaborative filtering for personalized recommendation. *ACM Trans. Inf. Syst.* **2023**, *41*, 1–27. [CrossRef]
47. Blake, J. Genre-specific error detection with multimodal feedback. *RELC J.* **2020**, *51*, 179–187. [CrossRef]
48. Baig, M.Z.; Kavakli, M. Multimodal systems: Taxonomy, methods, and challenges. *arXiv* **2020**, arXiv:2006.03813.
49. Andronas, D.; Apostolopoulos, G.; Fourtakas, N.; Makris, S. Multi-modal interfaces for natural Human-Robot Interaction. *Procedia Manuf.* **2021**, *54*, 197–202. [CrossRef]
50. Xu, L. Intelligence Preschool Education System based on Multimodal Interaction Systems and AI. *arXiv* **2024**, arXiv:2407.15326.
51. Alzubi, T.M.; Alzubi, J.A.; Singh, A.; Alzubi, O.A.; Subramanian, M. A multimodal human-computer interaction for smart learning system. *Int. J. Hum.-Comput. Interact.* **2023**, 1–11. [CrossRef]
52. Farooq, M.; Afraz, N.; Golpayegani, F. An Adaptive System Architecture for Multimodal Intelligent Transportation Systems. *arXiv* **2024**, arXiv:2402.08817.
53. Hu, B.; Xu, L.; Moon, J.; Yadwadkar, N.J.; Akella, A. MOSEL: Inference Serving Using Dynamic Modality Selection. *arXiv* **2023**, arXiv:2310.18481.
54. Wei, Z.; Wei, Z.; Chen, Z.; Li, R.; Xie, F.; Zheng, S. Study on the Influence of Environment on Multimodal Interaction. In Proceedings of the International Conference on Man-Machine-Environment System Engineering, Beijing, China, 20–23 October 2023; Springer: Singapore, 2023; pp. 353–361.
55. Katiyar, N.; Awasthi, M.V.K.; Pratap, R.; Mishra, M.K.; Shukla, M.N.; Tiwari, M.; Singh, R. Ai-Driven Personalized Learning Systems: Enhancing Educational Effectiveness. *Educ. Adm. Theory Pract.* **2024**, *30*, 11514–11524. [CrossRef]
56. Gaspar-Figueiredo, D.; Fernández-Diego, M.; Nuredini, R.; Abrahão, S.; Insfrán, E. Reinforcement Learning-Based Framework for the Intelligent Adaptation of User Interfaces. *arXiv* **2024**, arXiv:2405.09255.
57. Shanthakumar, V.A.; Peng, C.; Hansberger, J.; Cao, L.; Meacham, S.; Blakely, V. Design and evaluation of a hand gesture recognition approach for real-time interactions. *Multimed. Tools Appl.* **2020**, *79*, 17707–17730. [CrossRef]
58. Ascari, R.E.S.; Silva, L.; Pereira, R. Personalized gestural interaction applied in a gesture interactive game-based approach for people with disabilities. In Proceedings of the 25th International Conference on Intelligent User Interfaces, Cagliari, Italy, 17–20 March 2020; pp. 100–110.
59. Chen, J.; Seng, K.P.; Smith, J.; Ang, L.M. Situation awareness in ai-based technologies and multimodal systems: Architectures, challenges and applications. *IEEE Access* **2024**, *12*, 88779–88818. [CrossRef]
60. Al-Waisy, A.S.; Qahwaji, R.; Ipson, S.; Al-Fahdawi, S. A multimodal deep learning framework using local feature representations for face recognition. *Mach. Vis. Appl.* **2018**, *29*, 35–54. [CrossRef]
61. Wang, X.; Lyu, J.; Kim, B.G.; Parameshachari, B.; Li, K.; Li, Q. Exploring multimodal multiscale features for sentiment analysis using fuzzy-deep neural network learning. *IEEE Trans. Fuzzy Syst.* **2024**, *33*, 28–42. [CrossRef]
62. Liang, P.P.; Zadeh, A.; Morency, L.P. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Comput. Surv.* **2024**, *56*, 1–42.
63. Alwahaby, H.; Cukurova, M. The ethical implications of using Multimodal Learning Analytics: Towards an ethical research and practice framework. *EdArXiv* **2022**. [CrossRef]
64. Awasthi, V.; Verma, R.; Dhanda, N. Multimodal Dialogue Systems in the Era of Artificial Intelligence-Assisted Industry. In *Machine Vision and Industrial Robotics in Manufacturing*; CRC Press: Boca Raton, FL, USA, 2024; pp. 178–200.
65. Pellicer-Sánchez, A.; Tragant, E.; Conklin, K.; Rodgers, M.; Serrano, R.; Llanes, A. YOUNG LEARNERS'PROCESSING OF MULTIMODAL INPUT AND ITS IMPACT ON READING COMPREHENSION: AN EYE-TRACKING STUDY. *Stud. Second. Lang. Acquis.* **2020**, *42*, 577–598. [CrossRef]
66. Liang, P.P.; Lyu, Y.; Chhablani, G.; Jain, N.; Deng, Z.; Wang, X.; Morency, L.P.; Salakhutdinov, R. MultiViz: Towards User-Centric Visualizations and Interpretations of Multimodal Models. In Proceedings of the Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023; pp. 1–21.
67. Ciampelli, S.; Voppel, A.; De Boer, J.; Koops, S.; Sommer, I. Combining automatic speech recognition with semantic natural language processing in schizophrenia. *Psychiatry Res.* **2023**, *325*, 115252. [CrossRef] [PubMed]
68. Turk, M.; Athitsos, V. Gesture recognition. In *Computer Vision: A Reference Guide*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 535–539.
69. Lehmann-Willenbrock, N.; Hung, H. A multimodal social signal processing approach to team interactions. *Organ. Res. Methods* **2024**, *27*, 477–515. [CrossRef]
70. Sharma, K.; Giannakos, M. Multimodal data capabilities for learning: What can multimodal data tell us about learning? *Br. J. Educ. Technol.* **2020**, *51*, 1450–1484. [CrossRef]
71. Xiang, X.; Tan, Q.; Zhou, H.; Tang, D.; Lai, J. Multimodal fusion of voice and gesture data for UAV control. *Drones* **2022**, *6*, 201. [CrossRef]

72. Williams, A.S.; Ortega, F.R. Understanding gesture and speech multimodal interactions for manipulation tasks in augmented reality using unconstrained elicitation. *Proc. ACM Hum.-Comput. Interact.* **2020**, *4*, 1–21. [CrossRef]
73. Hasler, B.S.; Salomon, O.; Tuchman, P.; Lev-Tov, A.; Friedman, D. Real-time gesture translation in intercultural communication. *Ai Soc.* **2017**, *32*, 25–35. [CrossRef]
74. Wu, H.; Gai, J.; Wang, Y.; Liu, J.; Qiu, J.; Wang, J.; Zhang, X. Influence of cultural factors on freehand gesture design. *Int. J. Hum.-Comput. Stud.* **2020**, *143*, 102502. [CrossRef]
75. Liu, C.; Szirányi, T. Real-Time Human Detection and Gesture Recognition for On-Board UAV Rescue. *Sensors* **2021**, *21*, 2180. [CrossRef]
76. Barnum, G.; Talukder, S.; Yue, Y. On the benefits of early fusion in multimodal representation learning. *arXiv* **2020**, arXiv:2011.07191.
77. Wang, Z.; Wang, H.; Yu, H.; Lu, F. Interaction with gaze, gesture, and speech in a flexibly configurable augmented reality system. *IEEE Trans. Hum.-Mach. Syst.* **2021**, *51*, 524–534. [CrossRef]
78. Chen, F.; Luo, Z.; Xu, Y.; Ke, D. Complementary fusion of multi-features and multi-modalities in sentiment analysis. *arXiv* **2019**, arXiv:1904.08138.
79. Akalya devi, C.; Karthika Renuka, D. Multimodal emotion recognition framework using a decision-level fusion and feature-level fusion approach. *IETE J. Res.* **2023**, *69*, 8909–8920. [CrossRef]
80. Abbas, Q.; Alsheddy, A. A methodological review on prediction of multi-stage hypovigilance detection systems using multimodal features. *IEEE Access* **2021**, *9*, 47530–47564. [CrossRef]
81. Han, T.; Xie, W.; Zisserman, A. Temporal alignment networks for long-term video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022, New Orleans, LA, USA, 18–24 June 2022; pp. 2906–2916.
82. Liu, H.; Liu, Z. A multimodal dynamic hand gesture recognition based on radar–vision fusion. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–15. [CrossRef]
83. Bibi, J.; Fatima, L. Designing Intelligent Systems with Asynchronous Multimodal Data in Human-Computer Interactions. *OSFPreprints* **2023**. [CrossRef]
84. Middya, A.I.; Nag, B.; Roy, S. Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. *Knowl.-Based Syst.* **2022**, *244*, 108580. [CrossRef]
85. Khalane, A.; Makwana, R.; Shaikh, T.; Ullah, A. Evaluating significant features in context-aware multimodal emotion recognition with XAI methods. *Expert Syst.* **2023**, *42*, e13403. [CrossRef]
86. Li, X.; Chen, H.; He, S.; Chen, X.; Dong, S.; Yan, P.; Fang, B. Action recognition based on multimode fusion for VR online platform. *Virtual Real.* **2023**, *27*, 1797–1812. [CrossRef]
87. Yong, J.; Wei, J.; Lei, X.; Wang, Y.; Dang, J.; Lu, W. Intervention and Regulatory Mechanism of Multimodal Fusion Natural Interactions on AR Embodied Cognition. *Inf. Fusion* **2024**, *117*, 102910. [CrossRef]
88. Ding, J.; Wang, Y.; Si, H.; Gao, S.; Xing, J. Multimodal fusion-adaboost based activity recognition for smart home on wifi platform. *IEEE Sens. J.* **2022**, *22*, 4661–4674. [CrossRef]
89. Zhao, S.; Gong, M.; Fu, H.; Tao, D. Adaptive context-aware multi-modal network for depth completion. *IEEE Trans. Image Process.* **2021**, *30*, 5264–5276. [CrossRef] [PubMed]
90. Heck, M. Presentation Adaptation for Multimodal Interface Systems: Three Essays on the Effectiveness of User-Centric Content and Modality Adaptation. Ph.D. Thesis, Universität Mannheim, Mannheim, Germany, 2023.
91. Yang, L.; Sun, M.; Zhang, M.; Zhang, L. Multimodal motion control of soft ferrofluid robot with environment and task adaptability. *IEEE/ASME Trans. Mechatron.* **2023**, *28*, 3099–3109. [CrossRef]
92. Lu, Y.; Zhou, L.; Zhang, A.; Wang, M.; Zhang, S.; Wang, M. Research on Designing Context-Aware Interactive Experiences for Sustainable Aging-Friendly Smart Homes. *Electronics* **2024**, *13*, 3507. [CrossRef]
93. Zhang, T.; Liu, X.; Zeng, W.; Tao, D.; Li, G.; Qu, X. Input modality matters: A comparison of touch, speech, and gesture based in-vehicle interaction. *Appl. Ergon.* **2023**, *108*, 103958. [CrossRef]
94. Sun, X.; Zhang, Y. Improvement of autonomous vehicles trust through synesthetic-based multimodal interaction. *IEEE Access* **2021**, *9*, 28213–28223. [CrossRef]
95. Henderson, N.L. *Deep Learning-Based Multimodal Affect Detection for Adaptive Learning Environments*; North Carolina State University: Raleigh, NC, USA, 2022.
96. Wang, R.J.; Lai, S.C.; Jhuang, J.Y.; Ho, M.C.; Shiau, Y.C. Development of Smart Home Gesture-based Control System. *Sensors Mater.* **2021**, *33*, 3459–3471. [CrossRef]
97. Khalane, A.; Shaikh, T. Context-aware multimodal emotion recognition. In Proceedings of the International Conference on Information Technology and Applications: ICITA 2021, Dubai, United Arab Emirates, 13–14 November 2021; Springer: Singapore, 2022; pp. 51–61.
98. Calò, T.; De Russis, L. Enhancing smart home interaction through multimodal command disambiguation. *Pers. Ubiquitous Comput.* **2024**, *28*, 985–1000. [CrossRef]

99. Zhang, Z. Towards a multimodal and context-aware framework for human navigational intent inference. In Proceedings of the 2020 International Conference on Multimodal Interaction, Utrecht, The Netherlands, 25–29 October 2020; pp. 738–742.
100. Kopetz, H.; Steiner, W. *Real-Time Systems: Design Principles for Distributed Embedded Applications*; Springer Nature: Berlin/Heidelberg, Germany, 2022.
101. Di Mitri, D.; Schneider, J.; Drachsler, H. Keep me in the loop: Real-time feedback with multimodal data. *Int. J. Artif. Intell. Educ.* **2022**, *32*, 1093–1118. [CrossRef]
102. Yang, W.; Xiong, Z.; Mao, S.; Quek, T.Q.; Zhang, P.; Debbah, M.; Tafazolli, R. Rethinking generative semantic communication for multi-user systems with multi-modal LLM. *arXiv* **2024**, arXiv:2408.08765.
103. Lee, J.; Rodriguez, S.S.; Natarrajan, R.; Chen, J.; Deep, H.; Kirlik, A. What's This? A Voice and Touch Multimodal Approach for Ambiguity Resolution in Voice Assistants. In Proceedings of the 2021 International Conference on Multimodal Interaction, Montreal, QC, Canada, 18–22 October 2021; pp. 512–520.
104. Firdaus, M.; Chauhan, H.; Ekbal, A.; Bhattacharyya, P. EmoSen: Generating sentiment and emotion controlled responses in a multimodal dialogue system. *IEEE Trans. Affect. Comput.* **2020**, *13*, 1555–1566. [CrossRef]
105. Wang, Z.; Bai, X.; Zhang, S.; He, W.; Zhang, X.; Yan, Y.; Han, D. Information-level real-time AR instruction: A novel dynamic assembly guidance information representation assisting human cognition. *Int. J. Adv. Manuf. Technol.* **2020**, *107*, 1463–1481. [CrossRef]
106. Pei, S.; Chen, A.; Lee, J.; Zhang, Y. Hand interfaces: Using hands to imitate objects in ar/vr for expressive interactions. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 30 April–5 May 2022; pp. 1–16.
107. Langerak, T.; Zárate, J.J.; Vechev, V.; Lindlbauer, D.; Panozzo, D.; Hilliges, O. Optimal control for electromagnetic haptic guidance systems. In Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology, Virtual, 20–23 October 2020; pp. 951–965.
108. Sorrell, E.; Rule, M.E.; O'Leary, T. Brain–machine interfaces: Closed-loop control in an adaptive system. *Annu. Rev. Control. Robot. Auton. Syst.* **2021**, *4*, 167–189. [CrossRef]
109. Monarch, R.M. *Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI*; Simon and Schuster: New York, NY, USA, 2021.
110. Calado, A.; Roselli, P.; Errico, V.; Magrofuoco, N.; Vanderdonckt, J.; Saggio, G. A geometric model-based approach to hand gesture recognition. *IEEE Trans. Syst. Man Cybern. Syst.* **2022**, *52*, 6151–6161. [CrossRef]
111. Saktheeswaran, A.; Srinivasan, A.; Stasko, J. Touch? speech? or touch and speech? investigating multimodal interaction for visual network exploration and analysis. *IEEE Trans. Vis. Comput. Graph.* **2020**, *26*, 2168–2179. [CrossRef]
112. Romero, M.; Gómez-Canaval, S.; Torre, I.G. Automatic Speech Recognition Advancements for Indigenous Languages of the Americas. *Appl. Sci.* **2024**, *14*, 6497. [CrossRef]
113. Ye, J.; Hai, J.; Song, J.; Wang, Z. Multimodal data hybrid fusion and natural language processing for clinical prediction models. *AMIA Summits Transl. Sci. Proc.* **2024**, *2024*, 191.
114. Sweller, N.; Sekine, K.; Hostetter, A.B. Gesture-speech integration: Combining gesture and speech to create understanding. *Front. Psychol.* **2021**, *12*, 732357. [CrossRef]
115. Saito, K.; Hanzawa, K.; Petrova, K.; Kachlicka, M.; Suzukida, Y.; Tierney, A. Incidental and multimodal high variability phonetic training: Potential, limits, and future directions. *Lang. Learn.* **2022**, *72*, 1049–1091. [CrossRef]
116. Delecraz, S.; Becerra-Bonache, L.; Favre, B.; Nasr, A.; Bechet, F. Multimodal machine learning for natural language processing: disambiguating prepositional phrase attachments with images. *Neural Process. Lett.* **2021**, *53*, 3095–3121. [CrossRef]
117. Miao, H.; Cheng, G.; Gao, C.; Zhang, P.; Yan, Y. Transformer-based online CTC/attention end-to-end speech recognition architecture. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 6084–6088.
118. Dinkar, T. Computational models of disfluencies: Fillers and discourse markers in spoken language understanding. Ph.D. Thesis, Institut Polytechnique de Paris, Paris, France, 2022.
119. Zargham, N.; Fetni, M.L.; Spillner, L.; Muender, T.; Malaka, R. "I Know What You Mean": Context-Aware Recognition to Enhance Speech-Based Games. In Proceedings of the CHI Conference on Human Factors in Computing Systems 2024, Honolulu, HI, USA, 11–16 May 2024; pp. 1–18.
120. Gorman, B.M.; Crabb, M.; Armstrong, M. Adaptive subtitles: Preferences and trade-offs in real-time media adaption. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Virtual, 8–13 May 2021; pp. 1–11.
121. Zhou, H.; Wang, D.; Yu, Y.; Zhang, Z. Research progress of human–computer interaction technology based on gesture recognition. *Electronics* **2023**, *12*, 2805. [CrossRef]
122. Graichen, L.; Graichen, M. Multimodal Interaction in Virtual Reality: Assessing User Experience of Gesture-and Gaze-Based Interaction. In Proceedings of the International Conference on Human-Computer Interaction 2023, Copenhagen, Denmark, 23–28 July 2023; Springer: Cham, Switzerland, 2023; pp. 578–585.

123. Hang, C.Z.; Zhao, X.F.; Xi, S.Y.; Shang, Y.H.; Yuan, K.P.; Yang, F.; Wang, Q.G.; Wang, J.C.; Zhang, D.W.; Lu, H.L. Highly stretchable and self-healing strain sensors for motion detection in wireless human-machine interface. *Nano Energy* **2020**, *76*, 105064. [CrossRef]
124. Streli, P.; Jiang, J.; Rossie, J.; Holz, C. Structured Light Speckle: Joint Ego-Centric Depth Estimation and Low-Latency Contact Detection via Remote Vibrometry. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, San Francisco, CA, USA, 29 October–1 November 2023; pp. 1–12.
125. Ben Mocha, Y.; Burkart, J.M. Intentional communication: Solving methodological issues to assigning first-order intentional signalling. *Biol. Rev.* **2021**, *96*, 903–921. [CrossRef] [PubMed]
126. Mujahid, A.; Awan, M.J.; Yasin, A.; Mohammed, M.A.; Damaševičius, R.; Maskeliūnas, R.; Abdulkareem, K.H. Real-time hand gesture recognition based on deep learning YOLOv3 model. *Appl. Sci.* **2021**, *11*, 4164. [CrossRef]
127. Pushpakumar, R.; Sanjaya, K.; Rathika, S.; Alawadi, A.H.; Makhzuna, K.; Venkatesh, S.; Rajalakshmi, B. Human-Computer Interaction: Enhancing User Experience in Interactive Systems. *E3S Web Conf.* **2023**, *399*, 04037.
128. Satybaldina, D.; Kalymova, G.; Glazyrina, N. Application development for hand gestures recognition with using a depth camera. In Proceedings of the International Baltic Conference on Databases and Information Systems 2020, Tallinn, Estonia, 16–19 June 2020; Springer: Cham, Switzerland, 2020; pp. 55–67.
129. Neethu, P.; Suguna, R.; Sathish, D. An efficient method for human hand gesture detection and recognition using deep learning convolutional neural networks. *Soft Comput.* **2020**, *24*, 15239–15248. [CrossRef]
130. Adib, M.I. Fatigue Predictive Model for Mid-Air Gesture Interaction. Master's Thesis, University of Malaya (Malaysia), Kuala Lumpur, Malaysia, 2021.
131. Kowdiki, M.; Khaparde, A. Adaptive hough transform with optimized deep learning followed by dynamic time warping for hand gesture recognition. *Multimed. Tools Appl.* **2022**, *81*, 2095–2126. [CrossRef]
132. Kang, H.; Zhang, Q.; Huang, Q. Context-aware wireless-based cross-domain gesture recognition. *IEEE Internet Things J.* **2021**, *8*, 13503–13515. [CrossRef]
133. Al Said, N.; Al-Said, K. *Assessment of Acceptance and User Experience of Human-Computer Interaction with a Computer Interface*; LearnTechLib: Waynesville, NC, USA, 2020.
134. Rodriguez-Conde, I.; Campos, C. Towards customer-centric additive manufacturing: Making human-centered 3D design tools through a handheld-based multi-touch user interface. *Sensors* **2020**, *20*, 4255. [CrossRef]
135. Ikematsu, K.; Kato, K. ShiftTouch: Extending Touchscreens with Passive Interfaces Using Small Occluded Area for Discrete Touch Input. In Proceedings of the Seventeenth International Conference on Tangible, Embedded, and Embodied Interaction, Warsaw, Poland, 26 February–1 March 2023; pp. 1–15.
136. Zhao, L.; Liu, Y.; Ye, D.; Ma, Z.; Song, W. Implementation and evaluation of touch-based interaction using electrovibration haptic feedback in virtual environments. In Proceedings of the 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Atlanta, GA, USA, 22–26 March 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 239–247.
137. Salvato, M.; Heravi, N.; Okamura, A.M.; Bohg, J. Predicting hand-object interaction for improved haptic feedback in mixed reality. *IEEE Robot. Autom. Lett.* **2022**, *7*, 3851–3857. [CrossRef]
138. Cui, D.; Mousas, C. Evaluating wearable tactile feedback patterns during a virtual reality fighting game. In Proceedings of the 2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), Bari, Italy, 4–8 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 328–333.
139. Yang, T.H.; Kim, J.R.; Jin, H.; Gil, H.; Koo, J.H.; Kim, H.J. Recent advances and opportunities of active materials for haptic technologies in virtual and augmented reality. *Adv. Funct. Mater.* **2021**, *31*, 2008831. [CrossRef]
140. Fang, C.M.; Harrison, C. Retargeted self-haptics for increased immersion in VR without instrumentation. In Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology, Virtual, 10–14 October 2021; pp. 1109–1121.
141. Breitschaft, S.J.; Pastukhov, A.; Carbon, C.C. Where's my button? Evaluating the user experience of surface haptics in featureless automotive user interfaces. *IEEE Trans. Haptics* **2021**, *15*, 292–303. [CrossRef] [PubMed]
142. Lelevé, A.; McDaniel, T.; Rossa, C. Haptic training simulation. *Front. Virtual Real.* **2020**, *1*, 3. [CrossRef]
143. Patel, R.V.; Atashzar, S.F.; Tavakoli, M. Haptic feedback and force-based teleoperation in surgical robotics. *Proc. IEEE* **2022**, *110*, 1012–1027. [CrossRef]
144. Fan, L.; Song, A.; Zhang, H. Development of an integrated haptic sensor system for multimodal human–computer interaction using ultrasonic Array and cable robot. *IEEE Sens. J.* **2022**, *22*, 4634–4643. [CrossRef]
145. Freeman, E. Ultrasound haptic feedback for touchless user interfaces: Design patterns. In *Ultrasound Mid-Air Haptics for Touchless Interfaces*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 71–98.
146. Terao, Y.; Mizushina, H.; Yamamoto, K. Evaluation of usability improvement of contactless human interface with visual, auditory, and tactile sensation for aerial display. *Opt. Rev.* **2024**, *31*, 126–133. [CrossRef]
147. Valtakari, N.V.; Hooge, I.T.; Viktorsson, C.; Nyström, P.; Falck-Ytter, T.; Hessels, R.S. Eye tracking in human interaction: Possibilities and limitations. *Behav. Res. Methods* **2021**, *53*, 1592–1608. [CrossRef]

148. Neogi, D.; Das, N.; Deb, S. Eye-Gaze Based Hands Free Access Control System for Smart City Public Interfaces. In *AI and IoT for Smart City Applications*; Springer: Singapore, 2022; pp. 139–156.
149. Ryan, C.; O'Sullivan, B.; Elrasad, A.; Cahill, A.; Lemley, J.; Kielty, P.; Posch, C.; Perot, E. Real-time face & eye tracking and blink detection using event cameras. *Neural Netw.* **2021**, *141*, 87–97.
150. Rivu, R.; Abdrabou, Y.; Pfeuffer, K.; Esteves, A.; Meitner, S.; Alt, F. Stare: Gaze-assisted face-to-face communication in augmented reality. In Proceedings of the ACM Symposium on Eye Tracking Research and Applications 2020, Stuttgart Germany, 2–5 June 2020; pp. 1–5.
151. Ugwitz, P.; Kvarda, O.; Juříková, Z.; Šašinka, Č.; Tamm, S. Eye-tracking in interactive virtual environments: Implementation and evaluation. *Appl. Sci.* **2022**, *12*, 1027. [CrossRef]
152. Bektas, K. Toward a pervasive gaze-contingent assistance system: Attention and context-awareness in augmented reality. In Proceedings of the ACM Symposium on Eye Tracking Research and Applications 2020, Stuttgart Germany, 2–5 June 2020; pp. 1–3.
153. Gardony, A.L.; Lindeman, R.W.; Brunyé, T.T. Eye-tracking for human-centered mixed reality: Promises and challenges. In *Optical Architectures for Displays and Sensing in Augmented, Virtual, and Mixed Reality (AR, VR, MR)*; SPIE: Bellingham, WA USA, 2020; Volume 11310; pp. 230–247.
154. Cukurova, M.; Giannakos, M.; Martinez-Maldonado, R. The promise and challenges of multimodal learning analytics. *Br. J. Educ. Technol.* **2020**, *51*, 1441–1449. [CrossRef]
155. Sidenmark, L.; Parent, M.; Wu, C.H.; Chan, J.; Glueck, M.; Wigdor, D.; Grossman, T.; Giordano, M. Weighted pointer: Error-aware gaze-based interaction through fallback modalities. *IEEE Trans. Vis. Comput. Graph.* **2022**, *28*, 3585–3595. [CrossRef]
156. Niehorster, D.C.; Zemblys, R.; Holmqvist, K. Is apparent fixational drift in eye-tracking data due to filters or eyeball rotation? *Behav. Res. Methods* **2021**, *53*, 311–324. [CrossRef] [PubMed]
157. Kröger, J.L.; Lutz, O.H.M.; Müller, F. What does your gaze reveal about you? On the privacy implications of eye tracking. In Proceedings of the IFIP International Summer School on Privacy and Identity Management, Windisch, Switzerland, 19–23 August 2019; Springer: Cham, Switzerland, 2020; pp. 226–241.
158. Han, W.; Chen, H.; Gelbukh, A.; Zadeh, A.; Morency, L.p.; Poria, S. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In Proceedings of the 2021 International Conference on Multimodal Interaction, Montreal, QC, Canada, 18–22 October 2021; pp. 6–15.
159. Wenderoth, L. Exploring Multi-Modality Dynamics: Insights and Challenges in Multimodal Fusion for Biomedical Tasks. *arXiv* **2024**, arXiv:2411.00725.
160. Chakraborty, S.; Timoney, J. Multimodal Synchronization in Musical Ensembles: Investigating Audio and Visual Cues. In Proceedings of the Companion Publication of the 25th International Conference on Multimodal Interaction, Paris, France, 9–13 October 2023; pp. 76–80.
161. Liang, C.; Yang, D.; Liang, Z.; Wang, H.; Liang, Z.; Zhang, X.; Huang, J. Unsupervised Multi-modal Feature Alignment for Time Series Representation Learning. *arXiv* **2023**, arXiv:2312.05698.
162. Stednitz, S.J.; Lesak, A.; Fecker, A.L.; Painter, P.; Washbourne, P.; Mazzucato, L.; Scott, E.K. Probabilistic modeling reveals coordinated social interaction states and their multisensory bases. *bioRxiv* **2024**.
163. Zhang, C.; Yang, Z.; He, X.; Deng, L. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 478–493. [CrossRef]
164. Bian, J.; Wang, L.; Xu, J. Prioritizing modalities: Flexible importance scheduling in federated multimodal learning. *arXiv* **2024**, arXiv:2408.06549.
165. Constantin, S.; Eyiokur, F.I.; Yaman, D.; Bärmann, L.; Waibel, A. Multimodal Error Correction with Natural Language and Pointing Gestures. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2023, Paris, France, 1–6 October 2023; pp. 1976–1986.
166. Chumachenko, K.; Iosifidis, A.; Gabbouj, M. MMA-DFER: MultiModal Adaptation of unimodal models for Dynamic Facial Expression Recognition in-the-wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2024, Seattle, WA, USA, 17–21 June 2024; pp. 4673–4682.
167. Cha, M.C.; Ji, Y.G. Context Matters: Understanding the Effect of Usage Contexts on Users' Modality Selection in Multimodal Systems. *Int. J. Hum.-Comput. Interact.* **2024**, *40*, 6287–6302. [CrossRef]
168. Valada, A.; Mohan, R.; Burgard, W. Self-supervised model adaptation for multimodal semantic segmentation. *Int. J. Comput. Vis.* **2020**, *128*, 1239–1285. [CrossRef]
169. Avetisyan, L.; Yang, X.J.; Zhou, F. Towards Context-Aware Modeling of Situation Awareness in Conditionally Automated Driving. *arXiv* **2024**, arXiv:2405.07088.
170. Hsu, H.C.; Brône, G.; Feyaerts, K. When gesture "takes over": Speech-embedded nonverbal depictions in multimodal interaction. *Front. Psychol.* **2021**, *11*, 552533. [CrossRef]
171. Lei, F.; Cao, Z.; Yang, Y.; Ding, Y.; Zhang, C. Learning the user's deeper preferences for multi-modal recommendation systems. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 1–18. [CrossRef]

172. Barange, M.; Rasendrasoa, S.; Bouabdelli, M.; Saunier, J.; Pauchet, A. Impact of adaptive multimodal empathic behavior on the user interaction. In Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents, Faro, Portugal, 6–9 September 2022; pp. 1–8.
173. Wolniak, R.; Grebski, W. The Usage of Smart Voice Assistant in Smart Home. In *Zeszyty Naukowe. Organizacja i Zarządzanie/Politechnika Śląska*; Silesian University of Technology Publishing House: Gliwice, Poland, 2023; pp. 701–710.
174. Elkady, M.; ElKorany, A.; Allam, A. ACAIOT: A Framework for Adaptable Context-Aware IoT applications. *Int. J. Intell. Eng. Syst.* **2020**, *13*, 271–282. [CrossRef]
175. Lu, J.; Zhou, H. Implementation of artificial intelligence algorithm in embedded system. *J. Phys. Conf. Ser.* **2021**, *1757*, 012015. [CrossRef]
176. Mao, H.; Zhang, B.; Xu, H.; Yuan, Z.; Liu, Y. Robust-MSA: Understanding the impact of modality noise on multimodal sentiment analysis. In Proceedings of the AAAI Conference on Artificial Intelligence 2023, Washingtonn, DC, USA, 7–14 February 2023; Volume 37; pp. 16458–16460.
177. Guo, Q.; Yao, K.; Chu, W. Switch-bert: Learning to model multimodal interactions by switching attention and input. In Proceedings of the European Conference on Computer Vision 2022, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; pp. 330–346.
178. Liu, J.; Luo, D.; Fu, X.; Lu, Q.; Kang, K.Y. Design Strategy of Multimodal Perception System for Smart Environment. In *Internet of Things for Smart Environments*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 93–115.
179. Donley, J.; Tourbabin, V.; Lee, J.S.; Broyles, M.; Jiang, H.; Shen, J.; Pantic, M.; Ithapu, V.K.; Mehra, R. Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments. *arXiv* **2021**, arXiv:2107.04174.
180. Ryumin, D.; Ivanko, D.; Ryumina, E. Audio-visual speech and gesture recognition by sensors of mobile devices. *Sensors* **2023**, *23*, 2284. [CrossRef]
181. Jose, S.; Nguyen, K.T.; Medjaher, K. Enhancing industrial prognostic accuracy in noisy and missing data context: Assessing multimodal learning performance. *J. Intell. Manuf.* **2024**, 1–25. [CrossRef]
182. Zhao, G.; Shen, Y.; Zhang, C.; Shen, Z.; Zhou, Y.; Wen, H. RGBE-Gaze: A Large-scale Event-based Multimodal Dataset for High Frequency Remote Gaze Tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *47*, 601–615. [CrossRef]
183. Kang, P.; Li, J.; Jiang, S.; Shull, P.B. Reduce system redundancy and optimize sensor disposition for EMG–IMU multimodal fusion human–machine interfaces with XAI. *IEEE Trans. Instrum. Meas.* **2022**, *72*, 1–9. [CrossRef]
184. Yang, L.; Yan, W.; Xu, Z.; Wu, H. Robot multimodal anomaly diagnosis by learning time-lagged complex dynamics. In Proceedings of the 2021 IEEE International Conference on Real-time Computing and Robotics (RCAR), Xining, China, 15–19 July 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 509–514.
185. Yang, J.; Wang, Y.; Yi, R.; Zhu, Y.; Rehman, A.; Zadeh, A.; Poria, S.; Morency, L.P. Mtgat: Multimodal temporal graph attention networks for unaligned human multimodal language sequences. *arXiv* **2020**, arXiv:2010.11985.
186. Razzaghi, P.; Abbasi, K.; Shirazi, M.; Shabani, N. Modality adaptation in multimodal data. *Expert Syst. Appl.* **2021**, *179*, 115126. [CrossRef]
187. Wang, J.; Jiang, H.; Liu, Y.; Ma, C.; Zhang, X.; Pan, Y.; Liu, M.; Gu, P.; Xia, S.; Li, W.; et al. A comprehensive review of multimodal large language models: Performance and challenges across different tasks. *arXiv* **2024**, arXiv:2408.01319.
188. AlAbdulaali, A.; Asif, A.; Khatoon, S.; Alshamari, M. Designing multimodal interactive dashboard of disaster management systems. *Sensors* **2022**, *22*, 4292. [CrossRef] [PubMed]
189. Oppelt, M.P.; Foltyn, A.; Deuschel, J.; Lang, N.R.; Holzer, N.; Eskofier, B.M.; Yang, S.H. ADABase: A multimodal dataset for cognitive load estimation. *Sensors* **2022**, *23*, 340. [CrossRef] [PubMed]
190. Zhao, S.; Ren, X.; Deng, W.; Lu, K.; Yang, Y.; Li, L.; Fu, C. A novel transient balancing technology of the rotor system based on multi modal analysis and feature points selection. *J. Sound Vib.* **2021**, *510*, 116321. [CrossRef]
191. Gorlewicz, J.L.; Tennison, J.L.; Uesbeck, P.M.; Richard, M.E.; Palani, H.P.; Stefik, A.; Smith, D.W.; Giudice, N.A. Design guidelines and recommendations for multimodal, touchscreen-based graphics. *ACM Trans. Access. Comput. (TACCESS)* **2020**, *13*, 1–30. [CrossRef]
192. Larmuseau, C.; Cornelis, J.; Lancieri, L.; Desmet, P.; Depaepe, F. Multimodal learning analytics to investigate cognitive load during online problem solving. *Br. J. Educ. Technol.* **2020**, *51*, 1548–1562. [CrossRef]
193. Van Leeuwen, T. *Multimodality and Identity*; Routledge: London, UK, 2021.
194. Kalatzis, A.; Rahman, S.; Girishan Prabhu, V.; Stanley, L.; Wittie, M. A Multimodal Approach to Investigate the Role of Cognitive Workload and User Interfaces in Human-robot Collaboration. In Proceedings of the 25th International Conference on Multimodal Interaction, Paris, France, 9–13 October 2023; pp. 5–14.
195. Handosa, M.; Dasgupta, A.; Manuel, M.; Gračanin, D. Rethinking user interaction with smart environments—A comparative study of four interaction modalities. In Proceedings of the International Conference on Human-Computer Interaction 2020, Copenhagen, Denmark, 19–24 July 2020; Springer: Cham, Switzerland, 2020; pp. 39–57.

196. Xu, G.; Zhang, R.; Xu, S.X.; Kou, X.; Qiu, X. Personalized multimodal travel service design for sustainable intercity transport. *J. Clean. Prod.* **2021**, *308*, 127367. [CrossRef]
197. Oberste, L.; Rüffer, F.; Aydingül, O.; Rink, J.; Heinzl, A. Designing user-centric explanations for medical imaging with informed machine learning. In Proceedings of the International Conference on Design Science Research in Information Systems and Technology 2023, Pretoria, South Africa, 31 May–2 June 2023; Springer: Cham, Switzerland, 2023; pp. 470–484.
198. Yanamala, A.K.Y.; Suryadevara, S.; Kalli, V.D.R. Balancing Innovation and Privacy: The Intersection of Data Protection and Artificial Intelligence. *Int. J. Mach. Learn. Res. Cybersecur. Artif. Intell.* **2024**, *15*, 1–43.
199. Gupta, A.; Basu, D.; Ghantasala, R.; Qiu, S.; Gadiraju, U. To trust or not to trust: How a conversational interface affects trust in a decision support system. In Proceedings of the ACM Web Conference 2022, Lyon, France, 25–29 April 2022; pp. 3531–3540.
200. Snaith, M.; Conway, N.; Beinema, T.; De Franco, D.; Pease, A.; Kantharaju, R.; Janier, M.; Huizing, G.; Pelachaud, C.; op den Akker, H. A multimodal corpus of simulated consultations between a patient and multiple healthcare professionals. *Lang. Resour. Eval.* **2021**, *55*, 1077–1092. [CrossRef]
201. Acosta, J.N.; Falcone, G.J.; Rajpurkar, P.; Topol, E.J. Multimodal biomedical AI. *Nat. Med.* **2022**, *28*, 1773–1784. [CrossRef]
202. Hu, H.C.; Chang, S.Y.; Wang, C.H.; Li, K.J.; Cho, H.Y.; Chen, Y.T.; Lu, C.J.; Tsai, T.P.; Lee, O.K.S. Deep learning application for vocal fold disease prediction through voice recognition: Preliminary development study. *J. Med. Internet Res.* **2021**, *23*, e25247. [CrossRef] [PubMed]
203. Ozioko, O.; Dahiya, R. Smart tactile gloves for haptic interaction, communication, and rehabilitation. *Adv. Intell. Syst.* **2022**, *4*, 2100091. [CrossRef]
204. Bin, S.; Masood, S.; Jung, Y. Virtual and augmented reality in medicine. In *Biomedical Information Technology*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 673–686.
205. Chen, L.; Wang, W.; Qu, J.; Lei, S.; Li, T. A command and control system for air defense forces with augmented reality and multimodal interaction. *J. Phys. Conf. Ser.* **2020**, *1627*, 012002. [CrossRef]
206. Verhulst, I.; Woods, A.; Whittaker, L.; Bennett, J.; Dalton, P. Do VR and AR versions of an immersive cultural experience engender different user experiences? *Comput. Hum. Behav.* **2021**, *125*, 106951. [CrossRef]
207. Huizeling, E.; Peeters, D.; Hagoort, P. Prediction of upcoming speech under fluent and disfluent conditions: Eye tracking evidence from immersive virtual reality. *Lang. Cogn. Neurosci.* **2022**, *37*, 481–508. [CrossRef]
208. Gan, C.; Schwartz, J.; Alter, S.; Mrowca, D.; Schrimpf, M.; Traer, J.; De Freitas, J.; Kubilius, J.; Bhandwaldar, A.; Haber, N.; et al. Threedworld: A platform for interactive multi-modal physical simulation. *arXiv* **2020**, arXiv:2007.04954.
209. Dubovi, I. Cognitive and emotional engagement while learning with VR: The perspective of multimodal methodology. *Comput. Educ.* **2022**, *183*, 104495. [CrossRef]
210. Bennett, J.; Nguyen, P.; Lucero, C.; Lange, D. Towards an ambient intelligent environment for multimodal human computer interactions. In Proceedings of the Distributed, Ambient and Pervasive Interactions: 8th International Conference, DAPI 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, 19–24 July 2020; Proceedings 22; Springer: Berlin/Heidelberg, Germany, 2020; pp. 164–177.
211. Hugo, N.; Israr, T.; Boonsuk, W.; Ben Miloud, Y.; Cloward, J.; Liu, P.P. Usability study of voice-activated smart home technology. In Proceedings of the Cross Reality and Data Science in Engineering: Proceedings of the 17th International Conference on Remote Engineering and Virtual Instrumentation 17, Athens, GA, USA, 26–28 February 2020; Springer: Cham, Switzerlnad, 2021; pp. 652–666.
212. Tu, Y.; Luo, J. Accessibility Research on Multimodal Interaction for the Elderly. In Proceedings of the International Conference on Human-Computer Interaction 2024, Washington, DC, USA, 29 June–4 July 2024; Springer: Cham, Switzerlnad, 2024; pp. 384–398.
213. Kim, H.; Kwon, Y.T.; Lim, H.R.; Kim, J.H.; Kim, Y.S.; Yeo, W.H. Recent advances in wearable sensors and integrated functional devices for virtual and augmented reality applications. *Adv. Funct. Mater.* **2021**, *31*, 2005692. [CrossRef]
214. Mukherjee, J.; Azmi, Z.; Dixit, A.; Mishra, S.; Tomar, A.; Ali, K.B. Hand Gesture Recognition in Smart Cities. In *Investigations in Pattern Recognition and Computer Vision for Industry 4.0*; IGI Global: Hershey, PA, USA, 2023; pp. 215–231.
215. Lawrence, I.D.; Pavitra, A.R.R. Voice-Controlled Drones for Smart City Applications. In *Sustainable Innovation for Industry 6.0*; IGI Global: Hershey, PA, USA, 2024; pp. 162–177.
216. Zubatiuk, T.; Isayev, O. Development of multimodal machine learning potentials: Toward a physics-aware artificial intelligence. *Accounts Chem. Res.* **2021**, *54*, 1575–1585. [CrossRef]
217. Sharma, K.; Papamitsiou, Z.; Olsen, J.K.; Giannakos, M. Predicting learners' effortful behaviour in adaptive assessment using multimodal data. In Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, Frankfurt, Germany, 23–27 March 2020; pp. 480–489.
218. Augusto, J.C. Contexts and context-awareness revisited from an intelligent environments perspective. *Appl. Artif. Intell.* **2022**, *36*, 2008644. [CrossRef]
219. Atakishiyev, S.; Salameh, M.; Yao, H.; Goebel, R. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. *IEEE Access* **2024**, *12*, 101603–101625. [CrossRef]

220. Koochaki, F.; Najafizadeh, L. A data-driven framework for intention prediction via eye movement with applications to assistive systems. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2021**, *29*, 974–984. [CrossRef] [PubMed]
221. Tang, X.; Shen, H.; Zhao, S.; Li, N.; Liu, J. Flexible brain–computer interfaces. *Nat. Electron.* **2023**, *6*, 109–118. [CrossRef]
222. Kim, S.; Lee, S.; Kang, H.; Kim, S.; Ahn, M. P300 brain–computer interface-based drone control in virtual and augmented reality. *Sensors* **2021**, *21*, 5765. [CrossRef]
223. Won, S.M.; Song, E.; Reeder, J.T.; Rogers, J.A. Emerging modalities and implantable technologies for neuromodulation. *Cell* **2020**, *181*, 115–135. [CrossRef]
224. Chandler, J.A.; Van der Loos, K.I.; Boehnke, S.; Beaudry, J.S.; Buchman, D.Z.; Illes, J. Brain Computer Interfaces and Communication Disabilities: Ethical, legal, and social aspects of decoding speech from the brain. *Front. Hum. Neurosci.* **2022**, *16*, 841035. [CrossRef]
225. Wen, D.; Liang, B.; Zhou, Y.; Chen, H.; Jung, T.P. The current research of combining multi-modal brain-computer interfaces with virtual reality. *IEEE J. Biomed. Health Inform.* **2020**, *25*, 3278–3287. [CrossRef]
226. Meditskos, G.; Kontopoulos, E.; Vrochidis, S.; Kompatsiaris, I. Converness: Ontology-driven conversational awareness and context understanding in multimodal dialogue systems. *Expert Syst.* **2020**, *37*, e12378. [CrossRef]
227. Wu, C.; Fritz, H.; Bastami, S.; Maestre, J.P.; Thomaz, E.; Julien, C.; Castelli, D.M.; de Barbaro, K.; Bearman, S.K.; Harari, G.M.; et al. Multi-modal data collection for measuring health, behavior, and living environment of large-scale participant cohorts. *GigaScience* **2021**, *10*, giab044. [CrossRef]
228. Luo, F.M.; Jiang, S.; Yu, Y.; Zhang, Z.; Zhang, Y.F. Adapt to environment sudden changes by learning a context sensitive policy. In Proceedings of the AAAI Conference on Artificial Intelligence 2022, Virtual, 22 February–1 March 2022; Volume 36; pp. 7637–7646.
229. Ekatpure, R. Machine Learning Techniques for Advanced Driver Assistance Systems (ADAS) in Automotive Development: Models, Applications, and Real-World Case Studies. *Asian J. Multidiscip. Res. Rev.* **2022**, *3*, 248–304.
230. Jagnade, G.; Sable, S.; Ikar, M. Advancing Multimodal Fusion in Human-Computer Interaction: Integrating Eye Tracking, Lips Detection, Speech Recognition, and Voice Synthesis for Intelligent Cursor Control and Auditory Feedback. In Proceedings of the 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 6–8 July 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–7.
231. Jarosz, M.; Nawrocki, P.; Sniezynski, B.; Indurkhya, B. Multi-Platform Intelligent System for Multimodal Human-Computer Interaction. *Comput. Inform.* **2021**, *40*,83–103. [CrossRef]
232. Ling, Y.; Wu, F.; Dong, S.; Feng, Y.; Karypis, G.; Reddy, C.K. International Workshop on Multimodal Learning-2023 Theme: Multimodal Learning with Foundation Models. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Long Beach, CA, USA, 6–10 August 2023; pp. 5868–5869.
233. Anbarasan, R.; Gomez Carmona, D.; Mahendran, R. Human taste-perception: Brain computer interface (BCI) and its application as an engineering tool for taste-driven sensory studies. *Food Eng. Rev.* **2022**, *14*, 408–434. [CrossRef]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.