Vision-Based Multimodal Interfaces: A Survey and Taxonomy for Enhanced Context-Aware System Design

Yongquan ‘Owen’ Hu
School of Computer Science and Engineering,
University of New South Wales
Sydney, Australia
yongquan.hu@unsw.edu.au

Jingyu Tang*
School of Computer Science and Technology,
Huazhong University of Science and Technology
Wuhan, China
u202215423@hust.edu.cn

Xinya Gong*
Department of Computer Science and Engineering
South University of Science and Technology
Shenzhen, China
gongxinya123@gmail.com

Zhongyi Zhou
The University of Tokyo
Tokyo, Japan
zhongyi.zhou.work@gmail.com

Shuning Zhang
Tsinghua University
Beijing, China
zsn23@mails.tsinghua.edu.cn

Don Samitha Elvitigala
Exertion Games Lab, Department of Human-Centred Computing,
Monash University
Melbourne, Australia
don.elvitigala@monash.edu

Florian ‘Floyd’ Mueller
Exertion Games Lab, Department of Human-Centred Computing,
Monash University
Melbourne, Australia
floyd@exertiongameslab.org

Wen Hu
School of Computer Science and Engineering,
University of New South Wales
Sydney, Australia
wen.hu@unsw.edu.au

Aaron J. Quigley
CSIRO’s Data61 &
University of New South Wales
Sydney, Australia
aquigley@acm.org

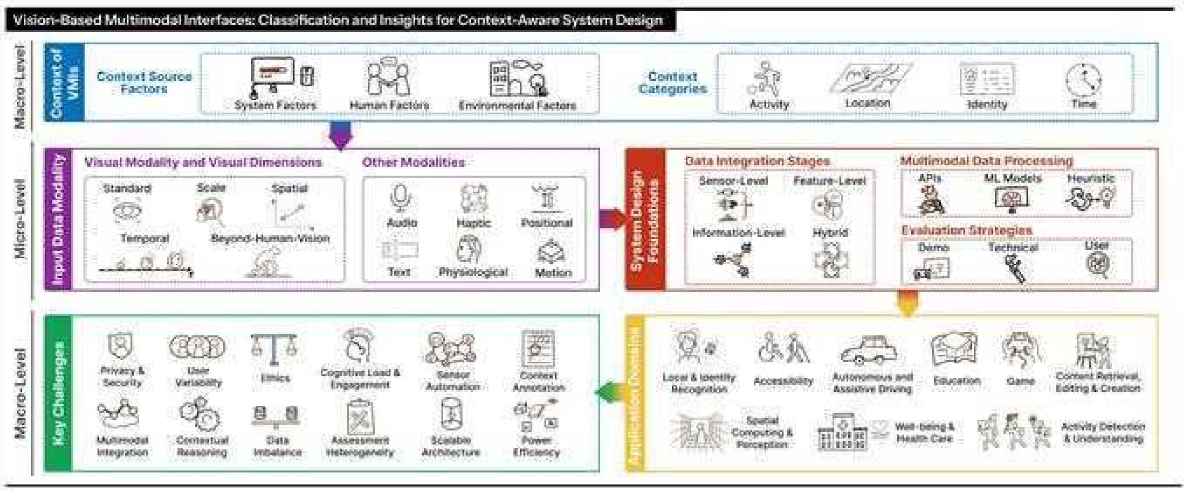


Figure 1: We review and categorize VMIs aimed at enhancing context awareness. Our key contribution is a *Macro-Micro-Macro level* (whole-detail-whole) system design framework, providing actionable references from a *Data Modality-Driven* perspective: (1) Macro-level contextual factors: considerations for context understanding (Section 3); (2) Micro-level system foundations: input data modality (visual + other modalities), data integration stages, multimodal data processing and evaluation strategies (Sections 4, 5); (3) Macro-level design synthesis: application domains, design considerations and key challenges (Sections 6, 7).

Abstract

The recent surge in artificial intelligence, particularly in multimodal processing technology, has advanced human-computer interaction, by altering how intelligent systems perceive, understand, and respond to contextual information (i.e., context awareness). Despite such advancements, there is a significant gap in comprehensive reviews examining these advances, especially from a multimodal data perspective, which is crucial for refining system design. This paper addresses a key aspect of this gap by conducting a systematic survey of data modality-driven Vision-based Multimodal Interfaces (VMIs). VMIs are essential for integrating multimodal data, enabling more precise interpretation of user intentions and complex interactions across physical and digital environments. Unlike previous task- or scenario-driven surveys, this study highlights the critical role of the visual modality in processing contextual information and facilitating multimodal interaction. Adopting a design framework moving from the whole to the details and back, it classifies VMIs across dimensions, providing insights for developing effective, context-aware systems.

CCS Concepts

• **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**; • **Computer systems organization** → *Sensors and actuators.*

Keywords

survey; system; context aware; computer vision; vision based interface; visual data; multimodal; artificial intelligence

ACM Reference Format:

Yongquan 'Owen' Hu, Jingyu Tang, Xinya Gong, Zhongyi Zhou, Shuning Zhang, Don Samitha Elvitigala, Florian 'Floyd' Mueller, Wen Hu, and Aaron J. Quigley. 2025. Vision-Based Multimodal Interfaces: A Survey and Taxonomy for Enhanced Context-Aware System Design. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 31 pages. https://doi.org/10.1145/3706598.3714161

1 Introduction

Context awareness is essential in Human-Computer Interaction (HCI), enabling systems to detect, interpret, and respond to contextual information [3, 58, 135], thereby facilitating adaptive and seamless interactions [173]. Vision-based interfaces (VIs), such as camera-based gesture recognition [59, 151], excel in interpreting complex visual data for tasks like fine-grained gesture recognition or spatial context analysis [65, 84]. VIs also enable unobtrusive interactions, supporting calm computing by reducing cognitive load [35, 136, 174]. Applications include smart homes and immersive environments such as Virtual Reality (VR) and Augmented Reality (AR) [55, 74, 109]. Vision-based Multimodal Interfaces (VMIs) enhance context awareness by integrating visual inputs with non-visual modalities, creating a unified understanding of the environment [123, 145]. For instance, VEmotion combines visual, GPS, and auditory data to improve driver emotion recognition accuracy by 28.5% compared to visual-only approaches [11], demonstrating how VMIs address unimodal limitations to deliver accurate, contextually relevant responses [74, 141].

Despite ongoing advancements in visual information processing, the use of VMIs as interactive tools remains nascent. The rapid development of technology, especially in multimodal Artificial Intelligence (AI), has outpaced the design principles and paradigms of interactive systems, creating a gap in intuitive systems that can be easily utilized by non-experts [138, 188]. As a result, VMIs have gained increasing attention in the HCI community for their ability to address the challenges of integrating multimodal data and enabling effective interactions within this evolving paradigm. As shown in Figure 2, there is a growing recognition of the value of research in this area, as evidenced by the increasing volume of related work (details in Section 2.3). Significantly, the expanding intersection of VMIs with context awareness reflects an emerging trend toward integrating these interfaces with a deeper understanding of the contextual factors influencing user interactions. Our study builds upon previous research, aligning with the growing interest in integrating multimodal data for context-aware systems, and provides a thorough and up-to-date analysis with practical guidance and systematic insights for advancing VMI design.

There exist related VMI-related surveys, particularly focusing on their applications in fields where emerging technologies enhance system capabilities, such as augmented reality enabling more immersive experiences [158], Generative AI (GenAI) improving content personalization, or mobile computing facilitating real-time interactions in dynamic environments [68]. In practice, building VMIs requires systematic efforts and there is a lack of actionable frameworks with practical insights in the community to guide the HCI practitioners to prototype interactive systems. For example, while Dumas et al. [40] provide theoretical foundations for multimodal interfaces, they offer limited practical guidance for complex, context-aware scenarios. This type of practical guidance, which is currently lacking in the field, is essential for bridging the gap between theory and practice and offering clear pathways for designing effective context-aware systems. Additionally, many studies emphasize specific perspectives, such as task-driven (e.g., gesture recognition [59]), scenario-driven (e.g., mobile and AR [51, 84, 158]), or technology-driven (e.g., GenAI [144]) approaches, which could inadvertently overlook the potential of a broader data perspective. These perspectives provide unified frameworks for diverse modalities, addressing dynamic real-world environments, characterized by rapidly changing contexts, user behaviors, and system requirements [140]. For instance, Bolchini et al. [14] demonstrated how data tailoring enhances adaptability, such as adjusting museum guides for low-vision users or personalizing content based on user interests. It underscores the value of a data-oriented approach in enabling systems to flexibly combine and adapt information for

[*] Jingyu Tang and Xinya Gong contributed equally to this work.

CHI '25, April 26-May 1, 2025, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3714161

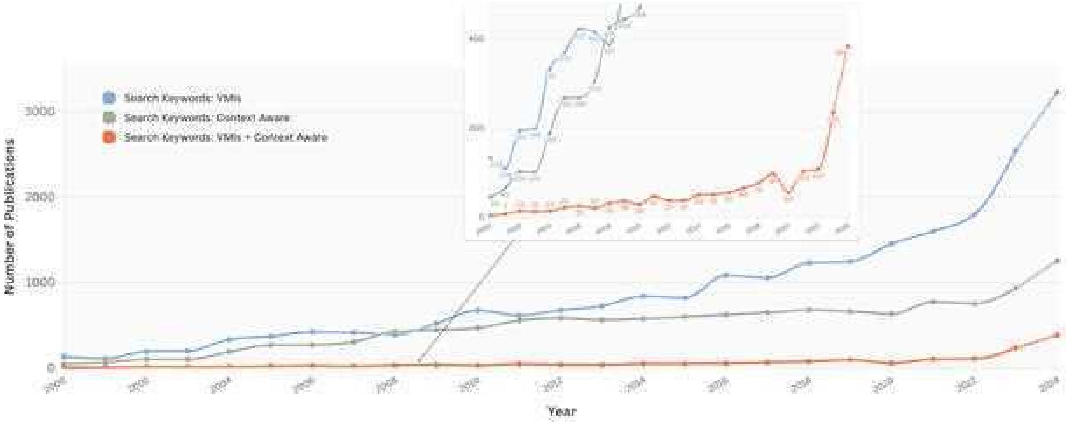


Figure 2: The publication growth trend for vision-based multimodal interface and context awareness in the ACM Digital Library (details of the search keywords are provided in Appendix A).

more effective and personalized interactions. Building on this foundation, our study adopts a ***data modality-driven*** lens to evaluate strategies for integrating visual modalities with other data streams, explore innovative methods, and assess their roles in enhancing interactive experiences within context-aware systems.

Our study offers a systematic review, organizing VMIs into a taxonomy using a ***Macro-Micro-Macro (3M) level*** system design framework, which seeks to address the aforementioned limitations. This framework aligns with the top-down and bottom-up design philosophy [15, 36] of moving from the holistic perspective (Section 3) to finer details (Section 4, 5) and then synthesizing them back into an integrated whole (Section 6, 7), as shown in Figure 1. This structure is intended not only to bridge existing gaps in usability but also to serve as an iterative ***step-by-step manual organized by sections*** for practitioners. For instance, we visualizes information flow across dimensions with the help of ***a Sankey diagram*** and ***an interactive website***, while Appendix B provides ***detailed literature statistics*** *(categories, counts, and citations) to enhance practical usability.* Moreover, by adopting a data modality-driven perspective, our study highlights its adaptability and flexibility in integrating diverse data modalities. Unlike task-driven or scenario-driven perspectives, this approach provides a systematic framework for addressing challenges such as cross-modal semantic alignment and data integration. For instance, Cai et al. [18] demonstrate how integrating multimodal data streams in smart healthcare systems enhances diagnostic accuracy and supports robust decision-making processes. These findings suggest that a data modality-driven perspective informs VMI design by structuring visual and non-visual modality integration, especially in complex scenarios. This perspective bridges theoretical insights with practical applications, enabling more adaptive and scalable interaction designs.

2 Scope and Methodology

2.1 Scope and Definitions

In this section, we aim to establish a clear scope and definitions of the terminology used throughout this paper.

2.1.1 *Context Awareness and Visual Data.*

Context awareness is a foundational concept in HCI, enabling systems to perceive and respond to environmental changes dynamically [109]. Initially focused on static factors like user location, it has evolved to encompass dynamic properties shaped by interactions and activities [34, 39, 135], supporting adaptive interfaces, personalized data, and smart environments [14]. Visual data has been pivotal in advancing these capabilities, as demonstrated by Schilit et al.'s navigation systems using visual feedback [135] and Dey and Abowd's applications for real-time tracking and location-based guidance [34, 35]. Over time, the role of visual data expanded through integration with other modalities in VMIs, enabling real-time spatial and gesture recognition in XR systems [10, 157]. These multimodal approaches allow systems to interpret complex contexts, such as subtle gestures or dynamic environments, and tailor interactions to diverse users, enhancing accessibility [62, 126]. To support these advancements, established frameworks like Dey and Abowd's [3] categorize context into location, identity, activity, and time, while Grubert et al. emphasize high-level factors like human and environmental elements [51]. Rather than creating a new framework, our work adapts these taxonomies to VMIs with application-specific customizations, preserving their strengths while tailoring them to the unique requirements of multimodal systems. This approach bridges theoretical foundations with practical applications, making visual data a cornerstone of adaptable and effective context-aware technologies.

2.1.2 *Vision-Based Interfaces.*

We define a VI by synthesizing insights from various sources: Kuno et al. [87] highlight the intuitive potential of visual input for user-centered interactions without rigid calibration; Sá et al. [131] demonstrate its integration with other modalities to enhance contextual awareness; Zabulis et al. [191] and Gopalan et al. [50] focus on processing visual input for capturing user behaviors and environmental factors through gesture recognition and posture analysis; and Kolsch et al. [84] emphasize system adaptability in dynamic environments. Building on these contributions, we define a VI as *a system that utilizes* ***visual input from one or more sensors (e.g., cameras),*** *capturing*

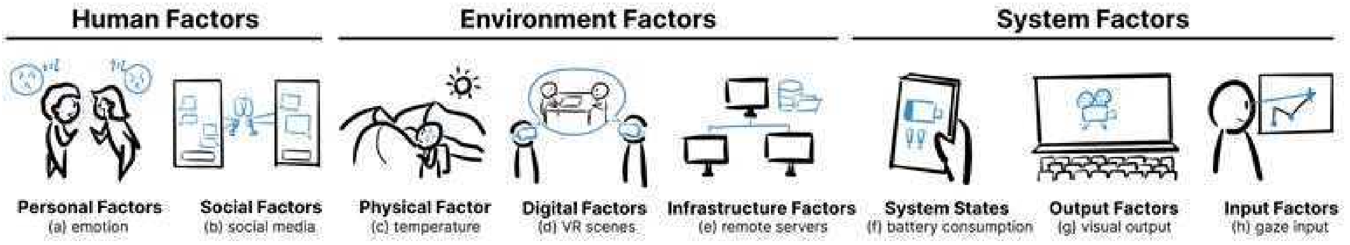


Figure 3: Examples of context source factors in VMIs with descriptions and citations (illustrative references (a): [11], (b): [129], (c): [132], (d): [22], (e): [171], (f): [37], (g): [22], (h): [92]).

user-relevant factors (e.g., human behavior, environment) to facilitate human-machine interaction. While the literature on VIs shares commonalities, it varies in descriptions that we exclude from our definition, considering them as features rather than essential criteria. For instance, real-time processing, though important for enhancing interaction [70, 84, 191], is not mandatory, as some studies prioritize algorithm accuracy over real-time capabilities [159]. Similarly, while most studies have pointed out cameras as primary sensors, some briefly mention [44, 159, 184, 198] or exclude them entirely [99].

2.1.3 Vision-based Multimodal Interface and Enhanced Context-Aware System. Our definition of the VMI builds on the foundational concept of a VI, incorporating insights from related literature [141, 145, 161]. Specifically, a VMI is *a subset of VI, characterized by the inclusion of* ***at least one non-visual modality in addition to the visual modality or a combination of two or more distinct visual dimensions within its input data*** (detailed in Section 4). This information, related to the user, is used to enhance interaction. VMIs can also enhance context awareness by integrating multimodal data, enabling systems to capture context more accurately [79]. While Salber noted the differences between multimodal systems, which rely on explicit input, and context-aware systems, which use implicit input [133], research has expanded multimodal systems to include both input types [130]. In this paper, we focus on the intersection of these definitions, with VMIs incorporating both explicit and implicit inputs. Notably, visual input in VMIs often functions as an implicit source of context-awareness, highlighting its critical role in understanding and adapting to dynamic environments.

2.2 Contributions

In this paper, we make three key contributions. First, we ***systematically review the literature*** across HCI venues, identifying trends and underexplored areas through a ***data modality-driven perspective***. We place a particular focus on the visual modality, while integrating data from other modalities to address the dynamic requirements of context awareness. Second, we propose a taxonomy structured within the ***3M framework*** for system design, providing an ***actionable reference*** that includes an iterative process and practical resources, such as ***interactive website***, to guide the development of context-aware systems. Third, we identify ***key design considerations*** and ***open research challenges***, offering future directions for advancing VMIs and multimodal interaction paradigms.

2.3 Literature Selection Methodology

2.3.1 Literature Search and Selection. We conducted a systematic literature search in digital libraries including ACM and IEEE, following the PRISMA framework [1]. Using the query `("vision-based") AND ("multimodal") AND ("context aware")` and related synonyms, we targeted English-language publications since 2018. This search was informed by factors such as research trends, advancements in the field, and paper volume. After removing duplicates, 929 papers remained, which were reviewed to exclude works outside the scope of our study, such as single-modality interfaces or non-HCI-related literature. This process resulted in 98 relevant papers. To complement the search, expert discussions added 11 significant works, yielding a curated collection of 109 papers. Further details on the search process and selection criteria are provided in Appendix A.

2.3.2 Analysis and Synthesis. The dataset was analyzed through a multi-step process. First, we conducted open coding on a small subset of our sample to identify an initial approximation of the dimensions and categories within the design space. Next, we reviewed the initial classification to assess the consistency and comprehensiveness of the categorization methods, during which categories were merged, expanded, or removed. Following this, we systematically coded the entire dataset, applying individual tags for precise categorization. Finally, we reviewed the individual tags to resolve any discrepancies and arrive at the final coding results. To minimize bias, ensure comprehensive assessment, and enhance reliability and transparency, the data was independently coded and analyzed by four co-authors, with the results subsequently consolidated [1].

3 Context of VMIs

In this section, we build on previous research to refine context classification, highlighting the ultimate goal of system design as its whole guiding factor.

3.1 Context Source Factors

To gain a deeper understanding of the factors influencing context, we build on the classification of situations proposed by Grubert

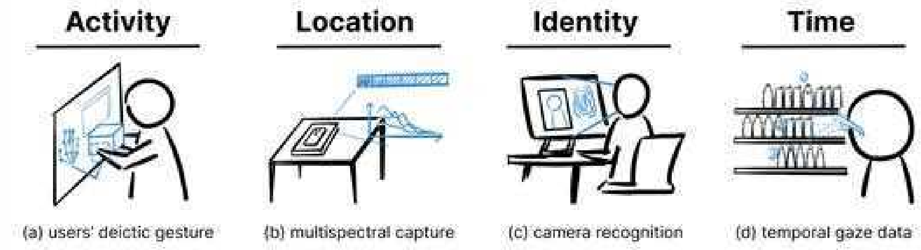


Figure 4: Examples of context categories in VMIs with descriptions and citations (illustrative references (a): [200], (b): [187], (c): [100], (d): [92]).

et al. [51]. However, we extend its application beyond AR scenarios to encompass a broader range of fields. Additionally, we refine their three major categories—human factors, environmental factors, and system factors—by introducing a more fine-grained classification, accompanied by examples for clarification and illustration, as shown in Figure 3.

Factor-1. Human Factors: Classically, *Human Factors* refer to the study of how humans interact with elements of a system or environment. This field focuses on understanding and improving how well people interact with the systems they use. Human factors can be categorized into personal and social factors.

-Personal Factors: Focusing on an individual user, personal factors such as cognitive load [19, 176, 201], emotion [11, 19, 44, 114] and user preference [19, 22, 176, 183] are frequently considered in interface design to promote a tailored user experience. Furthermore, user posture and physical movement [4, 100, 164, 165] constitute significant personal factors, underscoring the need for adaptive design strategies.

-Social Factors: In contrast to personal factors, social factors encapsulate the interactions and relational networks between multiple individuals. These elements include, but are not limited to, social media information [129], as well as social norms and social zones [139]. Together, these factors provide a richer understanding of social interplay, thereby improving the design and effectiveness of systems.

Factor-2. Environment Factors: *Environment Factors* describe the surrounding of the user and the interfaces in which interaction takes place. Within the domain of environmental factors, we distinguish between physical factors, digital factors and infrastructure factors.

-Physical Factors: Physical factors encompass environmental elements of the physical world. Raw factors, such as temperature [132, 167, 196], light levels [103, 137, 187, 196], and noise levels [103], can be directly perceived by human senses or measured via sensors. Derived factors, by contrast, are calculated by combining multiple raw factors or abstracting higher-level information from low-level data. For example, the spatial or geometric configuration of a scene can be inferred from multispectral or visual data [45, 137, 155, 185, 187]. Similarly, the presence or absence of physical artifacts, such as objects or materials, can provide contextual insights [33, 99, 104].

-Digital Factors: In contrast to physical factors, the second category of environmental factors focuses on the digital environment. This category encompasses information stored or displayed in virtual environments such as conversation logs [183], online conference data [60] and VR/AR scenes [22, 53, 159, 176].

-Infrastructure Factors: Unlike human and physical factors, which are often subject to immediate perception and direct manipulation, infrastructure factors operate behind the scenes, subtly yet powerfully influencing the system's dynamics. This category encompasses elements including network connectivity [2, 4, 24, 27, 100, 171, 176, 183, 185, 187], databases [2, 75] and remote server access [44, 113, 171].

Factor-3. System Factors: These include general system configuration, computational capabilities of the device, output and input devices, and modalities. System factors can be classified into system state, output factors, and input factors.

-System States: the system state refers to the current availability and performance of a system's computational resources, including factors such as memory usage [110], latency [110, 164], and battery consumption [37, 67, 110, 115, 116, 137, 183]. These elements collectively determine the system's capacity to handle tasks and its operational efficiency at any given time.

-Output Factors: Output factors refer to the various ways information is presented to the user, including visual output [22, 44, 45, 53, 73, 99, 159], as well as other modalities like audio [80, 103, 104, 152, 176] or haptic feedback [25, 97].

-Input Factors: Input factors describe the different methods available for users to interact with the system, such as gestures [47, 92], haptic input [4, 5, 170], mouse input [201], gaze [22, 92, 112, 116, 171, 176], or speech [11, 19, 81, 168, 171]. Depending on the input modalities available, the system can adapt its operation to best suit the user's input method.

3.2 Context Categories

Classifying context types is crucial for application designers to identify the most relevant aspects of context for their specific applications. Although technological advancements and changing application scenarios have led to the evolution of context classification, the general framework [3] remains a foundational reference for many studies. In this section, we adhere to their classic four-classification method (i.e., *activity*, *location*, *identity* and *time*) and provide examples to illustrate the role of VMIs in each category in Figure 4.

Category-1. Activity: VMIs in activity contexts leverage wearable technologies to recognize user actions like gestures and postures, enabling intuitive interactions. Key applications include vision-based gesture recognition in mixed reality and object annotation [92, 200], as well as gesture-based commands for robots, enhancing intent interpretation [47, 85, 163]. Gaze detection, another major use, tracks attention and assesses cognitive load using eye-tracking and neural networks [92, 176, 178]. VMIs also analyze head movements, speech, and driving behaviors across diverse contexts like video conferencing and healthcare [118, 147]. Ultimately, they enhance activity detection accuracy and expand interaction possibilities with multimodal data integration.

Category-2. Location: VMIs for location sensing use visual data directly or combine it with other modalities. Multispectral imaging detects material placement, as shown in SpeCam and SpectroPhone [137, 187], while systems like MicroCam combine RGB and IMU data for enhanced accuracy [67, 185]. Non-visual methods like GPS, IMU, and radar enable faster detection; for instance, ContextCam integrates GPS and Wi-Fi [11], while radar and Go-Pro data estimate user positions [41]. Although visual approaches provide rich contextual information, non-visual methods offer lighter, faster solutions, often complemented by visual data for broader applications like facial recognition.

Category-3. Identity: Identity recognition in VMIs answers "Who is involved?" and enhances contextual understanding. Auth+Track combines one-time authentication (e.g., iris, fingerprint) with continuous camera-based tracking to secure mobile use [100]. Khan et al.'s PAL integrates cameras and sensors for user authentication, supporting timely habit interventions [80]. For devices, AirConstellations and Kratos+ facilitate cross-device interactions and access control in multi-user environments [111, 148]. VMIs for user identity raise privacy concerns, while device identity is central to multi-device interactions.

Category-4. Time: Time in VMIs determines when systems act and tag context for later retrieval [3]. Real-time systems like GazePointAR and G-VOILA use gaze and gestures to ensure accurate, timely responses [92, 171]. Temporal data also triggers actions, such as MicroCam capturing images based on phone placement [67], or dynamically adjusting AR information based on user preferences [52]. Whether in real-time or staged systems, temporal considerations are vital for effective context-aware interactions.

4 Input Data Modality

As outlined in Section 2, VMIs are characterized by the modalities of their input data, which are typically acquired through sensing methods [40, 50]. To explore the multimodality of VMIs systematically, we begin with a detailed analysis of the visual modality, as it forms the core of our research focus. This analysis is further structured into several common dimensions, providing a foundation for a deeper understanding of VMIs from a multimodal data perspective. In addition, we briefly review other modalities, focusing on their data sources and functionalities, to offer a comprehensive view of how they complement the visual modality within multimodal systems.

4.1 Visual Modality and Visual Dimensions

Images, typically captured by cameras, are the most common medium in the visual modality. Accordingly, we categorize the visual dimension based on image concepts.

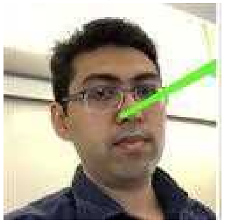
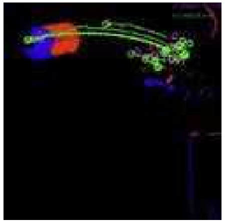
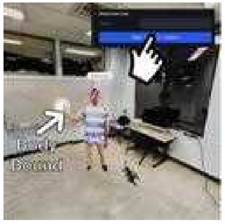
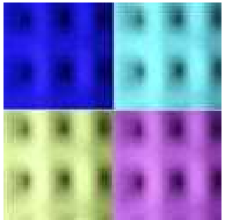
Dimension-1. Standard-Vision: The Standard Vision dimension refers to standard visible images captured by cameras, primarily in the form of RGB or grayscale images. For example, Su et al. utilized RGB images from the rear camera of a mobile phone in RASSAR to recognize and reconstruct objects, facilitating seamless navigation indoors and outdoors [153]. Similarly, Fan et al. employed the front camera of a mobile phone to detect user expressions using visible cues, enhancing the GenAI-based image creation process [44]. In VMIs, the Standard Vision dimension serves as the primary source of visual data, enabling systems to recognize objects, track movements, and detect interactions in real-time. Its ability to capture detailed and color-accurate visual information makes it indispensable for interpreting environmental cues.

Dimension-2. Scale: The scale dimension pertains to the scope or extent of the scene captured by the image, ranging from large-scale 360-degree panoramic views to microscopic images. Wide-angle cameras, for example, provide an expanded field of view, while fisheye lenses introduce specific distortions for specialized perspectives [100]. On the other end of the spectrum, microscopic imaging captures minute details beyond what the naked eye can perceive [37, 67]. The scale of an image plays a crucial role in determining its application, with panoramic images being essential for XR and mapping, while microscopic images excel in areas such as surface sensing, where detail-oriented analysis is key.

Dimension-3. Spatial: The spatial dimension captures depth perception and the three-dimensional (3D) structure of scenes, forming the basis for interpreting and interacting with complex environments. This aligns with Eriksson et al.'s "space" concept, emphasizing spatial configurations and dynamic camera spaces in creating immersive systems [43]. Depth data from RGB-D cameras or Light Detection and Ranging (LiDAR) provides precise spatial relationships, supporting environmental modeling and navigation [47, 168], while computational methods infer depth from 2D images, enabling spatial analysis without dedicated sensors [163, 167]. This dimension also incorporates the dynamic nature of "camera spaces," where user movement and sensor positioning shape spatial perception [43]. Integrating spatial data with other modalities, such as IMU or radar, enhances context-aware systems, enabling adaptive real-time interaction in applications like VR, AR, and robotics.

Dimension-4. Temporal: The temporal dimension refers to the dynamic aspect of vision, capturing changes in a scene over time. While static images offer a single moment in time, videos record movement and transitions, providing a continuous view of an evolving environment [60, 142]. This dimension enables the analysis of patterns, trends, and behaviors over time, offering essential insights for applications such as autonomous driving, where real-time monitoring of environmental changes is critical. Additionally, event cameras, which capture asynchronous brightness changes at the pixel level, provide a lightweight yet high-resolution representation of motion, enabling efficient tracking of rapid temporal dynamics [25]. By focusing on temporal changes, context-aware systems can track

Table 1: Visual modality dimensions and corresponding examples of types and uses.

Visual Modality Dimensions	Examples of Types and Uses	
Standard-Vision	RGB image for gaze capture [4] 	grayscale image of jacket texture [185]
Scale	microscopic image of plush [67] 	fisheye image for posture recognition [100]
Spatial	depth image for face detection [168] 	LiDAR image of room scaning [33]
Temporal	event image for motion tracking [25] 	video for dynamic operations [102] 
Beyond-Human-Vision	infrared image of human neck [24] 	multispectral image of breadboard [187] 

and respond to dynamic elements in the environment, allowing for more adaptive decision-making.

Dimension-5. Beyond-Human-Vision: The beyond-human-vision dimension extends perception beyond the visible spectrum, capturing electromagnetic radiation such as infrared, thermal, and ultraviolet waves. Unlike the spatial dimension, which focuses on geometric and depth information within visible light, this dimension broadens the scope by enabling systems to detect otherwise invisible properties of the environment. For instance, thermal cameras can capture heat variations, critical for applications such as object detection and environmental monitoring [27, 132]. Infrared imaging is frequently used for low-visibility scenarios or detecting surface characteristics, while multispectral imaging leverages light from multiple wavelengths to extract diverse features [56, 187]. Additionally, LiDAR technology expands beyond visual light, offering precise 3D mapping and environmental awareness [73, 167]. This dimension is distinct for its ability to capture properties invisible to the naked eye, facilitating novel applications in areas such as healthcare, security, and advanced robotics.

We provided examples in each category of data modalities and listed them in Table 1. It is important to note that the defined dimensions are not strictly orthogonal; a single image type may exhibit characteristics spanning multiple dimensions. For instance, LiDAR data [33], categorized under the spatial dimension, also possesses properties beyond human vision. However, classification is guided by the most prominent feature. Since LiDAR data is primarily used for detecting spatial parameters and often complements camera inputs, we show it in the "spatial" category. Similarly, the grayscale image example [185] in the standard-vision dimension also reflects scale characteristics, as it is captured through a microscope lens.

4.2 Other Modalities

Other modalities are categorized from the perspective of sensing devices. These are different types of data that can be collected from various sensors. In this part, we discuss several common sensing modalities that can be combined with camera data to enhance system performance.

Modality-1. Audio: Audio is the modality that is most commonly combined with vision. It serves as both an input and output medium, enhancing the scope of interaction and engagement. As an input, audio can take various forms [47, 81, 183]. For instance, the human voice can be employed in conjunction with visual information to understand human intention [19, 92, 162, 171] and human emotions [11, 105, 113, 118], which is useful in applications like video conferencing [60].

Modality-2. Text: Text is also often combined with the visual modality in VMIs. There are two primary input methods: the first involves using technology, such as software APIs, to convert speech into text, enabling the system to accurately interpret user commands and facilitate more natural, intuitive interactions. This approach is widely applied in areas such as human state understanding [183] and education [128]. The second method involves manual text input, commonly used in content recommendation, creation, and retrieval [44, 107, 114, 129, 183, 197]. Notably, we classify speech-to-text under 'text' as it represents processed textual input for system, while human voice remains under 'audio' due to its raw acoustic nature, highlighting the complementary roles of these modalities in multimodal integration.

Modality-3. Motion: Motion data typically captures parameters such as speed, direction, and acceleration of an object or person. Common sensors include gyroscopes, accelerometers, and Inertial Measurement Units (IMUs), with IMUs being integrated systems that often combine both gyroscopes and accelerometers, along with other sensors. It is often used in conjunction with camera data for tasks such as motion-tracking [73, 115, 116, 147], gesture or pose estimation [4, 163, 165], and surface sensing [67, 122].

Modality-4. Haptic: Haptic information, the sense of touch, can work in conjunction with the visual modality. As an input, it provides critical contact geometry information, enhancing the system's understanding of user interactions [5]. The sensing method of such data can come from user touch [4] or from machine manipulation based on tactile devices [170].

Modality-5. Positional: Radar and GPS are data sensors commonly used to measure position modalities. Radar has many advantages over visual modalities, such as being resistant to occlusion, high test accuracy, and good privacy and security. Therefore, it is often very complementary to visual modalities. The combination of the two is often used in the field of surveying and mapping [28] and autonomous driving [110], where multimodality promotes accurate and reliable environmental perception, which is essential for safe and efficient operation. Similarly, autonomous driving is also a typical application of the combination of GPS and vision [11]. This combination is also employed for positioning and navigation tasks, enhancing accessibility in barrier-free applications [178].

Modality-6. Physiological: Physiological data, such as heart rate [80, 196], can provide information about a user's emotional state or level of engagement. This category also includes data obtained from EEG, EMG, EDA, and PPG [53, 176]. These different types of physiological data can provide insights into various aspects of a user's physical and emotional state. This data can be combined with camera data to improve user experience or system performance.

To sum up, we classified and explored the foundational role of data modalities in defining VMIs in this section, emphasizing the centrality of visual input while briefly addressing other modalities. Table 2 further illustrates how multimodal data integration enhances context awareness. For example, the Pose-on-the-Go system combines visual modalities with motion and haptic data for activity tracking, demonstrating practical applications in mobile contexts. Similarly, VEmotion integrates visual, audio, and physiological inputs to infer drivers' emotional states, showcasing the synergy of multimodal data in real-time applications. These examples underscore the importance of visual modalities in context-aware systems while highlighting the potential of multimodal integration to address diverse challenges and enable advanced interactions.

5 System Design Foundations

With a clear understanding of multimodal data and its contextual definition, researchers need to develop systems capable of processing multimodal data captured from target contexts. In this section, we present a categorization of system design considerations for the technical implementation of VMIs. Specifically, we focus on

Table 2: An overview of visual and other modalities with examples highlighting their roles in enhancing context awareness.

Examples	Device (Data Source)	Visual Modality	Other Modalities	Role of Multimodal Data for Enhancing Context Awareness
Pose-on-the-Go [4]	smartphone	standard-vision, spatial	motion, haptic	Aiding context awareness in activity tracking, particularly in mobile applications
VEmotion [11]	GPS sensor, smartphone	standard-vision, temporal	audio, text, positional, motion, physiological	Infering the driver's emotional state indirectly by integrating diverse real-time environmental and vehicular context inputs
Blind Walking Guidance [104]	camera, microphone and laser sensor	standard-vision, spatial, beyond-human-vision	audio	Enhancing understanding of object placement and user activity, providing real-time feedback in navigation tasks
MicroCam [67]	smartphone, server	standard-vision, scale	motion	Improving surface detection of user actions in location-based services
VirtuWander [172]	VR helmet	standard-vision, spatial	audio, text	Boosting system responsiveness and precision in activity recognition while deepening understanding of user intention
EyeMU [85]	smartphone	standard-vision, temporal	audio, motion	Enhancing identity recognition through multimodal interactions such as synchronizing gestures and voice

the following three research questions that guide HCI system research: (Q1) Where is multimodal data integrated within the system for multimodal processing? (Q2) How can we process multimodal processing in VMIs? (Q3) How should the system be evaluated to understand its performance? Additionally, we faithfully reported the strengths and limitations of approaches at each stage of system design (Table 3, Table 4, and Table 5), offering high-level guidance for HCI researchers in designing and implementing VMI systems.

5.1 Data Integration Stages:

VMIs are structured around different levels of multimodal data processing, including sensor-level, feature-level, information-level, and hybrid integration systems. These architectures define how the system integrates multimodal data for better performance.

Stage-1. Sensor-level Integration: Sensor-level systems integrate raw sensing data at the very early stage of the multimodal systems. For example, in MicroCam [67], IMU and microscopic visual data are combined for real-time surface detection. PepperPose [165] fuses IMU data and visual input for full-body pose estimation, enhancing pose perception in dynamic environments and improving adaptability to various spatial directions. VEmotion [11] integrates vehicle speed, weather, and road type data to predict driver emotions. This early-stage fusion allows systems to leverage multiple perspectives to enhance performance and robustness in complex real-world scenarios.

One challenge in integrating multimodal sensor data is synchronizing and calibrating information from disparate sources. Proper alignment is essential for capturing time-dependent contexts within interactive systems that require real-time functionality. For example, VEmotion monitors a driver's behavior using multiple sensors to predict their emotional state in real time [11]. The system synchronizes data with precise timestamps from different devices, ensuring the effective analysis of temporal patterns. In MicroCam, IMU and visual data align with each other precisely for effective surface detection by leveraging temporal correlation between motion signals and visual textures to ensure robust and accurate classification [67].

Stage-2. Feature-level Integration: Feature-level systems integrate multimodal data in a later-stage of a data processing pipeline – they fuse the embedding of the raw sensor data from each modality. In an exemplary pipeline of this type, the system first encodes data from each modality using dedicated encoders. These modality-specific embeddings are then fused together using a multimodal

Table 3: A summary of key design considerations for the data integration stages in VMIs.

	Benefits	**Challenges**
Sensor-Level	+ improved raw data quality	- sensor synchronization
	+ early-stage noise reduction	- latency
Feature-Level	+ richer multimodal representation	- synchronization of high-dimensional data
	+ mitigates limitations of single modalities	- handling varying processing speeds
Information-Level	+ modular processing	- computational overhead
	+ better interpretation of complex data	- synchronization issues
Hybrid	+ flexibility for diverse tasks	- system complexity
	+ increased robustness	- modular architectures
		- latency

encoder, facilitating a multimodal representation of a target context. For instance, Saad et al. [132] combined thermal and visual features from a thermal camera, leveraging the complementary strengths of these modalities to improve robustness in varying lighting conditions, such as low visibility or shadowed environments. Similarly, Rene et al. [45] integrated depth and RGB data from a camera to construct more accurate 3D representations, which improved spatial perception and object localization. These feature-level combinations enable a more comprehensive understanding compared to using a single modality, particularly by addressing modality-specific limitations through mutual reinforcement.

Feature-level integration requires precise synchronization to ensure compatibility and alignment of data from different modalities, enabling a comprehensive understanding of the target context. Unlike sensor- or decision-level integration, feature-level systems face unique challenges due to the need to combine high-dimensional and abstract data representations from various sensors. These challenges include varying processing speeds across data encoders, the computational demands of handling multimodal features in real-time, and ensuring temporal and spatial alignment of data. For instance, Saad et al. proposed a thermal-RGB system where lighting variations and thermal noise required preprocessing steps such as filtering inconsistent thermal signatures and normalizing RGB inputs [132]. Similarly, the Velt system demonstrated the need for precise calibration and alignment when fusing depth and RGB data for accurate 3D scene representation [45]. Addressing these issues often involves lightweight feature extraction models, dynamic feature prioritization, and robust synchronization mechanisms to maintain performance under diverse conditions. These strategies enable feature-level systems to harmonize disparate data types effectively while managing computational overhead in real-time applications.

Stage-3. Information-Level Integration: We defined "information" here as human-perceivable semantics like adaptive voice feedback [176], scene descriptions [30], as opposed to those "feature" defined above (e.g., numerical embeddings [153]). Information-level systems typically conduct reasoning on such human-perceivable data to facilitate a more explainable processing pipeline. For example, Lim et al. [103] integrated environmental factors like brightness and CO2 levels with wide-angle camera data to form a cohesive scene understanding. Similarly, G-VOILA [171] merged gaze-tracking and environmental data, enhancing situational awareness by understanding both the user's interactions and their surroundings.

This architecture allows for modular processing, and it also requires careful synchronization to avoid misalignment between modalities. While offering a more abstract interpretation of complex data, information-level systems can increase computational load, necessitating efficient fusion algorithms and redundancy checks to manage inconsistencies or missing data.

Stage-4. Hybrid Integration: Hybrid integration systems combine feature-level, sensor-level, and information-level fusion, enabling flexible integration across multiple processing stages. For example, EmoTour [112] fuses audio-visual data, physiological signals, and behavioral cues like eye movements to recognize tourist emotions, capturing data at various stages from raw sensor signals to processed behavior.

These systems offer flexibility in data processing by adapting fusion strategies based on system requirements. EmoTour employs both feature-level and information-level fusion, enhancing system robustness by integrating multiple perspectives. Designers must ensure synchronization across modalities while minimizing latency, as well as implement modular architectures to manage the complexity of multi-stage fusion in dynamic environments.

5.2 Multimodal Data Processing

In addition to understanding *where* prior work integrate multimodal data, this subsection introduces *how* the existing literature processes the multimodal data to interpret a target context. Different from existing surveys on multimodal ML algorithms [6, 195], we position our scope on provide a taxonomy and strategic guidance for future HCI researchers for their the system implementations. Different from ML studies that benchmark ML models for higher

Table 4: A summary of key design considerations for the multimodal data processing in VMIs.

	Benefits	Challenges
Foundational model APIs	+ quick implementation + high accuracy + easy integration	- cloud service dependence - privacy issues - non-real-time processing
Developing dedicated ML models	+ interpretability + on-device processing	- high development effort - dataset collection
Heuristic methods	+ rapid prototyping + real-time applications	- low accuracy - low robustness

accuracy, HCI research studies the full system for better user experiences. Therefore, it is of great importance to guide HCI practitioners on how they can reasonably prototype a design concept with ML or other approaches. To this end, we categorize approaches used in the existing literature into three classes, shown as follows.

Processing-1. Rapid Solution Prototyping via Foundational Model APIs: Existing Machine Learning (ML) hubs like Hugging Face and PyTorch Hub contributed to the community with a large number of APIs so that developers can easily reuse for creative applications. Implementing the system with model APIs allows VMI researchers to rapidly implement data processing pipeline in VMIs without training a dedicated ML model. For example, MediaPipe provides off-the-shelf solutions for many real-time on-device solutions like pose detection and object tracking so that developers can build real-time apps with minimal engineering effort [108]. The community has widely utilized such APIs for prototyping new interactive systems, e.g., with facial recognition and gesture detection [106] and with LLM APIs [199]. Despite these advantages, the use of APIs brings significant concerns. Privacy issues arise when relying on cloud-based services for data processing, especially in sensitive domains. Additionally, foundational models are typically large, making them unsuitable for on-device processing and introducing latency, which reduces responsiveness in real-time applications [168].

Processing-2. Developing A Dedicated ML Models: Developing dedicated machine learning models enables tailored multimodal data processing, offering greater control over system performance and privacy. Traditional ML techniques, such as Support Vector Machines (SVMs) and Random Forests, are widely applied in structured data tasks such as surface sensing and posture estimation [11, 185]. Thanks to the powerful open-source libraries such as TensorFlow, PyTorch, and scikit-learn [1], researchers usually can quickly implement these ML models in a prototypical system. Additionally, traditional ML models are usually lightweight and explainable, implying that researchers can easily interact with, and debug, a ML-based prototype in real time to understand the new experience. However, traditional ML methods are less effective when dealing with unstructured data like images and videos.

To process such unstructured multimodal data, researchers usually incorporated a neural network to enhance the perceptual intelligence of a system. Prior work demonstrated the effectiveness of such approaches in various tasks like scene understanding and gesture recognition, which are critical to spatial-aware and interaction-aware applications [118, 139]. The main reason behind the success of such methods is their ability to automatically encode unstructured data into useful features for task-specific prediction. However, to develop these ML models, researchers need to acquire large datasets for training, and the collection of datasets in many downstream applications often requires significant effort [67]. Additionally, using deep neural networks in an interactive system will inevitably cause computation overhead. This brings a big challenge to on-device systems which need to perform real-time inference with a power consumption limit [197].

Processing-3. Heuristic Methods: Heuristic methods, relying on rule-based processing, offer a lightweight and fast alternative to machine learning. These methods are particularly used for quickly prototyping the design concepts of an interactive paradigm. This can facilitate rapid conceptual verification by taking humans in the loop in the early-stage system development process. For example, simple heuristics can involve predefined rules, such as selecting the closest matching depth map based on a straightforward similarity metric, to guide camera localization during bronchoscopic navigation. These rule-based adjustments enable rapid and lightweight prototyping without requiring extensive computational resources, aligning well with early-stage system design needs [142]. Once the system design concept is verified, researchers typically choose to invest more development efforts on a concept by, e.g., building a dedicated ML model, and utilizing simple heuristic methods are not suitable for system deployment due to its over-simplified modeling mechanism.

5.3 Evaluation Strategies

Given a context-aware solution by analyzing the multimodal data, how can we understand its performance? This subsection introduces three kinds of evaluation approaches commonly used in the prior work.

Evaluation-1. Prototyping and Demonstration: Evaluation through demonstration is a technique used to assess how well a

[1] Some open-source ML libraries: Tensorflow: https://www.tensorflow.org/. PyTorch: https://pytorch.org/. scikit-learn: https://scikit-learn.org/.

Table 5: A summary of key design considerations for the evaluation strategies in VMIs.

	Benefits	Limitations
Demonstration	+ rapid evaluation + collecting practical insights in the early stage	- lack of scalability - lack of standard metrics
Technical Evaluation	+ benchmarking with objective metrics + reproducible experiments + understanding technical system performance	- insufficient human factor considerations - limited applicability in the early stage development
User Evaluation	+ understanding realistic user experience + understanding usability + capturing user interaction data	- high consumption of human effort - difficulty in reproducing the experiment results

system will perform in specific scenarios. The most common approaches identified include prototypes [11, 19, 22, 73, 112, 167, 176, 185], proof-of-concept demonstrations [4, 132, 165, 171], and case studies [27, 54, 101, 107, 183]. Other approaches include programming by demonstration, where a system learns behaviors by observing human actions and replicating them [152]. For instance, in SonifyARCS, the system learns to generate auditory feedback based on user actions without explicit programming. Additionally, showing example applications [56] demonstrates the practical use of a system in real-world scenarios. Evaluation through demonstration is particularly useful during early-stage development when quick feedback on system performance is required or when the system operates in novel environments that lack established benchmarks. It is ideal for exploratory systems where proof-of-concept or prototype evaluations can reveal the system's potential in real-world contexts without needing large-scale deployments. This method is also beneficial when evaluating systems designed for highly specific use cases that require situational or contextual understanding rather than standardized testing.

Evaluation-2. Technical Evaluation: Technical evaluation primarily focuses on assessing key performance parameters of the system. The most common approaches include measuring accuracy [27, 37, 47, 56, 67, 73, 167, 168, 185, 187] and time metrics, such as response time or task completion time [22, 73, 92, 129, 164, 176]. Additionally, some works evaluate system performance by comparing their results with other systems, for instance, comparing classification algorithms [67, 115, 168]. Many studies also employ ablation studies to gain deeper insights into the interface by measuring how each component contributes to the system's overall performance [156, 168, 171, 183]. A technical evaluation is particularly suited for later stages of development when the system is relatively stable and requires precise measurements of its effectiveness and efficiency. It is essential when a system is intended to replace or outperform existing solutions, as comparative evaluations and ablation studies offer insights into the system's strengths and potential weaknesses. Moreover, technical evaluations are ideal when fine-tuning system performance is necessary, as they allow for detailed analysis of accuracy, speed, and individual system components under controlled conditions.

Evaluation-3. User Evaluation: User evaluation refers to measuring the effectiveness of a system through user studies. To quantitatively understand the system performance perceived by users, the community has introduced various Likert-scale metrics targeting different evaluation scenarios [19, 22, 44, 152, 167, 171, 183], such as SUS [22, 60, 92, 167] and NASA TLX [22, 60, 100, 162] . Understanding users' qualitative comments also plays a critical role in the evaluation. Common approaches include conducting interviews [19, 75, 103] and gathering user feedback [152, 167] through self-reported experiences [165], think-aloud studies [202], and diary studies [92].

It is important to note that it is a common practice to combine user evaluation with demonstration [75, 101, 183] or technical evaluations [56, 100, 129, 176]. The critical factor in study design is the evaluation's objective – specifically, the research questions researchers aim to explore. For example, the System Usability Scale (SUS) [17] and NASA TLX [57] are widely used examples of Likert scale-based questionnaires, focusing on measuring usability and perceived workload, respectively [78]. Interviews and think-aloud studies are effective when deeper, qualitative insights into user behavior and preferences are needed. Additionally, think-aloud or task-based studies are well-suited for systems requiring real-time interaction analysis, while surveys or interviews can capture users' overall experiences after interacting with the system. A combination of these evaluation techniques usually provides a more comprehensive evaluation but it causes extra workload for researchers.

6 Application Domains

We have identified nine key application areas for context awareness in VMIs, selected for their relevance in demonstrating the practical applications and unique capabilities of VMIs. These domains illustrate how VMIs integrate visual and multimodal data to address specific context-aware challenges, emphasizing strengths such as precision, adaptability, and real-time responsiveness. Covering a range of scenarios from location sensing to healthcare and gaming, these areas reflect the practical value of VMIs in supporting context-aware interactions. The selected domains also provide insight into how VMIs contribute to advancing HCI by addressing current challenges and enabling more effective system designs. Detailed examples and applications are shown in Figure 5.

Domain-1. Location and Identity Recognition: Location and identity recognition are key application areas for VMIs. For location recognition, various sensor-based systems have been developed, especially wearable devices like handheld systems (e.g., MicroCam [67], SpeCam [187], SpectroPhone [137]) and leg-mounted devices (e.g., HotFoot [132], RadarFoot [41]). These systems use visual and non-visual modalities to enhance accuracy and environmental awareness. For identity recognition, systems such as Auth + Track [100] leverage mobile phones and embedded cameras, effectively using fragmented "downtime" between interactions to ensure seamless experiences with minimal user burden. These applications emphasize high recognition accuracy and low latency to maintain system reliability and engagement.

Domain-2. Activity Detection and Understanding: VMIs improve activity detection through wearable technologies that recognize actions like gestures and postures, enabling intuitive interaction [165, 200]. Vision-based gesture recognition is critical, with cameras detecting gestures in mixed reality and interactive learning [200]. IMUs (e.g., accelerometers) and sensors further support gesture detection and activity recognition [49, 85]. Gaze tracking using eye-tracking or neural networks [178, 201] monitors attention and cognitive load, enabling adaptive interfaces and intuitive interactions [86, 92, 154, 176]. VMIs also detect movements [178], speech [60], and driving behaviors [101] across diverse contexts like healthcare and video conferencing [118, 147]. By integrating multimodal data, VMIs enhance detection accuracy and expand interaction possibilities, making them essential for context-aware systems.

Domain-3. Autonomous and Assistive Driving: Driver state detection is vital for safety in autonomous and assistive driving. Many systems assess the driver's state using behaviors like steering, pedal usage, and vehicle speed [93]. Recent systems integrate cameras to monitor gaze, facial expressions, and head movements, improving accuracy [38, 83, 160]. Combining these visual cues with in-car sensors detecting conditions like fatigue or intoxication enhances state classification. For example, sensors identifying slurred speech, slow reactions, or alcohol odor improve driver monitoring. External factors like road conditions, weather, and traffic further refine assessments [110]. Addressing stress due to challenging conditions rather than impairment enhances safety. Beyond detection, adaptive interventions like visual alerts [88] and voice assistants [72, 176] improve alertness, while haptic feedback (e.g., steering wheel vibrations) enhances response effectiveness. These innovations underscore VMIs' role in improving driving safety through context-aware interventions.

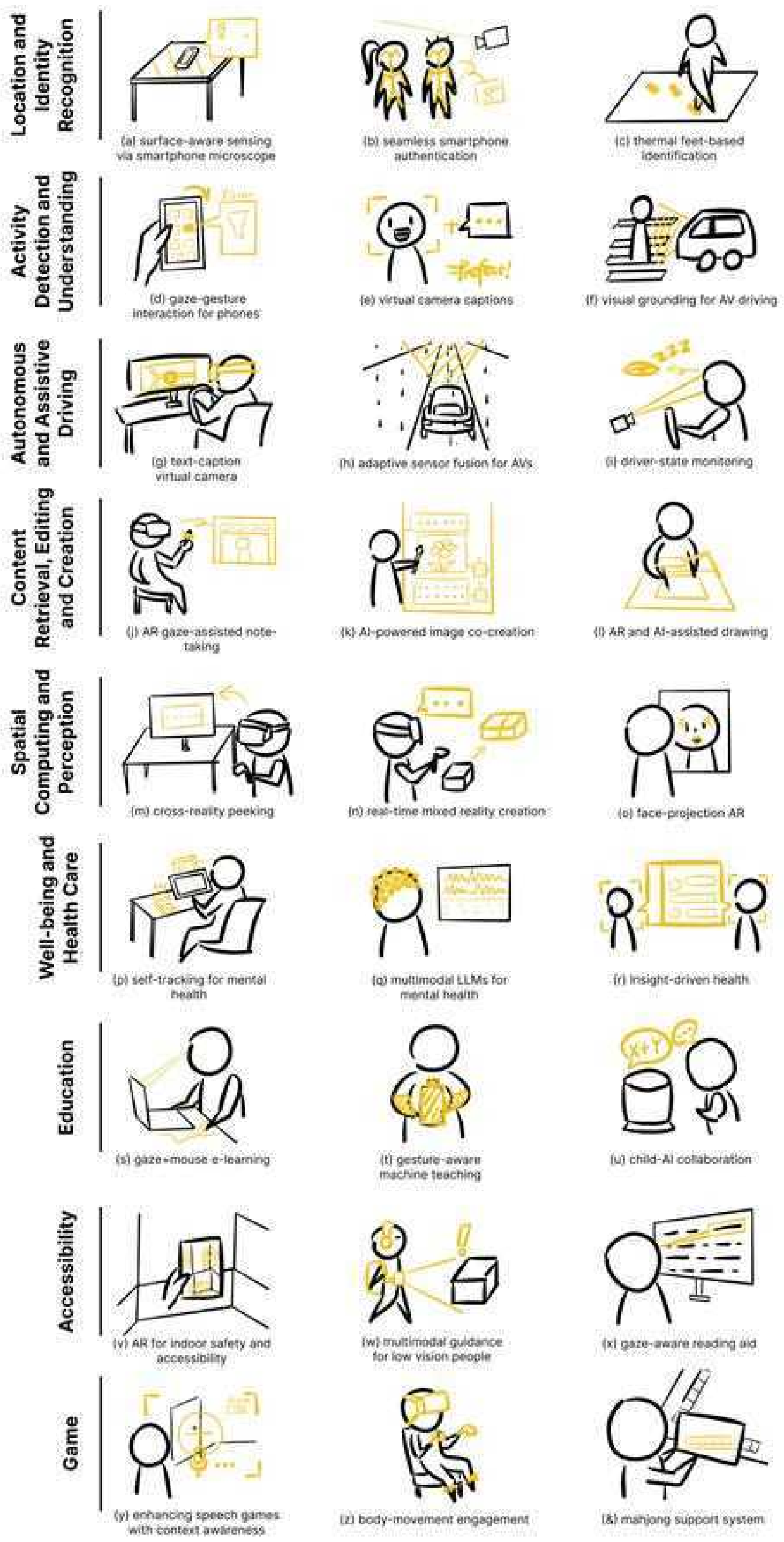
Domain-4. Content Retrieval, Editing and Creation: Multimodal interactions are increasingly essential in VMIs, enabling systems to interpret complex contexts effectively. Integrating multimodal LLMs allows richer interactions by processing inputs from visual, audio, and textual data. For example, facial expressions and gaze [162] enhance user intent interpretation, critical for creative applications where text-based inputs limit design exploration. Combining multimodal data improves context understanding, supporting adaptive interaction in domains like education and entertainment [44]. GenAI further reduces creation costs by generating initial content via Diffusion models, enabling users to refine outputs [64]. By advancing context awareness, multimodal VMIs facilitate intuitive, dynamic, and user-centered interactions, paving the way for more flexible human-computer collaboration.

Domain-5. Spatial Computing and Perception: VMIs play a critical role in enhancing context awareness in XR environments. Advancements in tracking technologies have introduced intuitive interactions like peeking [177], body-around [13, 46], object-centric [102], bare-hand [82], audio [119], and text-based interactions [23, 30, 146, 189]. For example, modern LLMs revolutionize text-based workflows by enabling efficient and intuitive user interactions. However, challenges like noisy real-time tracking hinder precise virtual alignment [98]. Addressing these alongside improving display and tracking technologies could expand XR's precision applications, such as surgical assistance. Collaborative XR environments are advancing but require overcoming latency and network distortion to ensure seamless teamwork [77, 134]. VMIs' continued development will enhance spatial perception, enabling accurate, context-aware interactions across industries.

Domain-6. Well-being and Health Care: The integration of VMIs and LLMs holds great promise for healthcare by enhancing context awareness through multimodal data analysis. While LLMs are widely used to infer mental [181] and physical [42, 76] health from text-based data, incorporating visual information like facial expressions and video-based emotion tracking can provide more holistic patient assessments [69]. Combining visual and physiological signals (e.g., EEG, EPG) offers critical insights, especially in mental health care, where body language and expressions reveal psychological states [103]. Integrating VMIs into LLM-driven healthcare systems enables more effective diagnosis, monitoring, and personalized treatment by leveraging multimodal inputs, unlocking new healthcare innovations.

Domain-7. Education: With advances in AI, VMIs have become essential for educational applications by integrating visual and non-visual modalities like text, sound, and sensors to create personalized learning experiences. For instance, VMIs enable intuitive interactions by interpreting gestures, voice, and contextual cues [193, 200, 201]. Gesture-aware systems like LookHere use real-time visual feedback to enhance machine teaching [200], enabling learners to annotate objects via natural gestures. Multimodal LLMs further enhance these interfaces by interpreting complex data and generating contextually relevant information, improving engagement in remote learning and multimedia interactions [128]. Combining gaze and mouse data improves engagement monitoring in e-learning, outperforming single-modality methods [201], demonstrating VMIs' potential to personalize and optimize educational technologies.

Domain-8. Accessibility: VMIs have significantly advanced accessibility by combining visual and other sensory modalities to facilitate seamless interaction and navigation. For instance, integrating visual data with auditory or haptic feedback enhances navigation for visually impaired users in digital and physical environments [153, 166]. However, challenges remain in ensuring multimodal systems' reliability in real-world scenarios where sensor data may be disrupted. Addressing issues like noise, interference, and seamless integration of multiple modalities requires robust algorithms and efficient designs [104]. Solving these challenges


Location and Identity Recognition
Activity Detection and Understanding
Autonomous and Assistive Driving
Content Retrieval, Editing and Creation
Spatial Computing and Perception
Well-being and Health Care
Education
Accessibility
Game
Perfect!
X+Y

(a) surface-aware sensing via smartphone microscope

(b) seamless smartphone authentication

(c) thermal feet-based identification

(d) gaze-gesture interaction for phones

(e) virtual camera captions

(f) visual grounding for AV driving

(g) text-caption virtual camera

(h) adaptive sensor fusion for AVs

(i) driver-state monitoring

(j) AR gaze-assisted note-taking

(k) AI-powered image co-creation

(l) AR and AI-assisted drawing

(m) cross-reality peeking

(n) real-time mixed reality creation

(o) face-projection AR

(p) self-tracking for mental health

(q) multimodal LLMs for mental health

(r) insight-driven health

(s) gaze+mouse e-learning

(t) gesture-aware machine teaching

(u) child-AI collaboration

(v) AR for indoor safety and accessibility

(w) multimodal guidance for low vision people

(x) gaze-aware reading aid

(y) enhancing speech games with context awareness

(z) body-movement engagement

(&) mahjong support system

Figure 5: Examples of application domains for VMIs (illustrative references (a): [67], (b): [100], (c): [132], (d): [85], (e): [60], (f): [101], (g): [176], (h): [110], (i): [38], (j): [162], (k): [44], (l): [64], (m): [177], (n): [30], (o): [13], (p): [103], (q): [69], (r): [42], (s): [201], (t): [200], (u): [193], (v): [153], (w): [166], (x): [166], (y): [192], (z): [149], (&): [159]).

is crucial for developing accessible systems that ensure reliable, user-friendly interactions across diverse settings.

Domain-9. Game: In gaming, VMIs enhance context awareness by integrating visual data with modalities like sound [192], motion, and physiological signals [149], creating immersive environments. In VR games, combining sensor-tracked body movements with visual data assesses user engagement and immersion [149, 192]. AR applications like Mahjong use context-aware image recognition to provide real-time feedback and improve gameplay efficiency [159]. Integrating modalities like IMUs or haptic feedback further enhances interaction by contextualizing users' physical and virtual surroundings, enabling adaptive gaming experiences that respond to real-time contexts.

7 Design Considerations and Key Challenges

Following the steps outlined above, a general context-aware system can be constructed. Revisiting the framework is necessary to ensure critical design considerations are addressed. These considerations are categorized into three aspects: user-centric considerations (challenges 1, 2, 3, and 4), data management and processing (challenges 5, 6, and 7), and system integration and resource optimization (challenges 8-12). This classification provides a structured approach to identify key aspects and guide future refinements.

Challenge-1. Privacy and Security-aware Systems: Privacy and security are critical considerations in VMI systems, as evidenced by 46 related studies, highlighting the potential sensitivity of visual data. Such data often contains rich, personal information, necessitating robust measures to protect user privacy. For example, PAL [80] used on-device deep learning for privacy-preserving, low-shot context detection, while Wang et al. [168] developed a silent-speech interaction system that avoids visual data collection by relying on depth sensing. Despite these advancements, key challenges persist, particularly in minimizing the amount of data collected and ensuring the protection of sensitive visual information. Strategies to address these challenges include dynamically adjusting the level of data detail [124], capturing only essential silhouette information [96], and using obfuscation techniques such as perturbations [186] and style transfer [180]. However, the ongoing arms race between privacy-preserving measures and increasingly sophisticated extraction attacks necessitates continuous development of more robust defense strategies.

Challenge-2. User Variability: User variability is discussed in 51 studies, reflecting its significance in the design of VMI systems. Diverse user characteristics, such as facial structure, head shape, or personal preferences, can impact system performance and accuracy. For instance, RASSAR [153] focused on incorporating varying accessibility requirements when designing systems, emphasizing the differences in users' indoor environments and the need for customized workflows to evaluate safety risks. Similarly, Ahuja et al. [4] addressed personal variance in full-body pose estimation by employing extensive sensor fusion, utilizing both front and rear cameras on smartphones. Despite these advancements, user variability remains a significant challenge for VMI systems. Variations in facial structure, head shape, and other physical features can significantly affect sensor accuracy [7, 48, 94]. Moreover, environmental factors, including lighting conditions and occlusions such as glasses or piercings, further complicate system performance [9, 71, 89, 117].

Challenge-3. Ethics: In our survey, 27 studies have specifically examined challenges related to LLMs [165] and embodied AIs [30]. These works focused on the relationship between humans and AIs, addressing concerns such as data privacy, the protection of human rights during experimentation, and the potential consequences for future AI development. For instance, De et al. [30] explored fears around job displacement for developers and creators but highlighted that their framework facilitates more effective human-AI collaboration by ensuring human involvement in the system. Despite these efforts, key ethical challenges remain, including the need to mitigate algorithmic bias, ensure fairness in AI-driven decision-making, and protect sensitive data. Furthermore, open challenges persist in addressing the societal implications of AI systems, such as ensuring transparency, accountability, and the responsible use of technology, particularly in contexts involving vulnerable populations.

Challenge-4. Cognitive Load and User Engagement: Cognitive load and user engagement are closely interconnected, mentioned in 58 work, often exhibiting a negative correlation. These factors are crucial considerations in the design of action-intensive systems, particularly within immersive environments like VR. One of the primary challenges highlighted is the difficulty in quantifying and managing cognitive load, which directly impacts user engagement. Users often face barriers to sustained interaction due to the overwhelming complexity of VR environments. The integration of multimodal data and enhanced context awareness present valuable opportunities to address this challenge. For example, Somarathna et al. [149] introduced body movements as a novel indicator of user engagement in VR gaming, offering an alternative to traditional methods. Wen et al. [176] developed a cognitively adaptive voice interface that adjusts information delivery based on varying levels of urgency and cognitive demand, aiming to optimize the balance between system responsiveness and user cognitive load. However, despite these advancements, key challenges persist. These include the need for more precise methods to measure cognitive load in real-time and across diverse users, as well as the challenge of maintaining user engagement without leading to mental fatigue. Additionally, the development of systems that can dynamically adapt to individual cognitive states throughout an interaction remains an unresolved issue.

Challenge-5. Automated Sensor Configuration: Sensing is fundamental to capturing vision-based multimodal data, reflected in 32 studies. The rapid growth of the IoT, which connects billions of sensors, has made manual sensor configuration impractical [127]. To address this, several works have focused on automating or semi-automating sensor connections to applications. For example, Kong et al. [85] explored the automatic integration of gaze-tracking sensors and IMUs to recognize gestures, while Ahuja et al.[4] used dynamic deployment of sensors and cameras in conjunction with inverse kinematics algorithms to estimate full-body poses. These efforts laid the foundation for automating sensor configuration, particularly in systems designed for human recognition. Despite these advances, several key challenges remain. These include the need for more robust methods for automatic sensor discovery, seamless integration of diverse sensor types, and ensuring real-time data synchronization across large-scale sensor networks.

Challenge-6. Context Discovery: This aspect was mentioned by 48 studies, focusing on how to automatically interpret and annotate sensor data within diverse application domains. As sensor data is generated, it must be contextualized to make it meaningful and actionable. Several approaches have been proposed to automate this process. For instance, Zargham et al. [192] explored context-aware speech recognition, where environmental and action-based context from the game enhanced the accuracy of speech recognition in interactive gaming. Similarly, Su et al. [152] utilized event context during AR interactions, applying LLMs and audio models for context-based sound acquisition and sonification. These efforts highlight the potential of integrating context awareness into sensor systems, but challenges remain. Specifically, there is the difficulty of automating context discovery in highly heterogeneous environments, where sensor data may vary significantly across different domains. While advances in semantic technologies and linked data [29, 63, 90, 150, 169] offer promising avenues for future development, key open challenges include improving the accuracy and scalability of context annotation across diverse applications and enabling real-time context awareness in dynamic environments.

Challenge-7. Semantic Multimodal Data Integration: A total of 54 papers have discussed this related aspect. Multimodal systems often require the semantic alignment of diverse sensor modalities, necessitating the development of advanced semantic modeling frameworks [169]. For example, Wang et al. [164] developed a companion bot with a visual interface that semantically integrates sensor data to provide enhanced feedback for rehabilitation, while Xu et al. used multimodal fusion of visual and audio signals through LLMs to create evolving user profiles in conversational agents [183]. Despite these efforts, semantic integration of multiple modalities often remained a case-specific manner, which scalability of the fusion algorithms across diverse sensor types and contexts, or the development of adaptive fusion algorithms worth of future exploration.

Challenge-8. Multimodal Contextual Reasoning: The reasoning-related processing was involved in 73 tasks. In VMI systems, contextual reasoning enables the system to interpret complex, dynamic interactions, enhancing its ability to understand and respond to evolving situations. However, reasoning about the relationships is inherently context and task-dependent. Koch et al. [83] developed a system for inferring blood alcohol concentration in real-time based on gaze and head movement data, while Fan et al. [44] integrated contextual reasoning into a human-AI co-creation system for generating artistic images. Although there were early attempts and early advancements, significant challenges remain particularly in achieving real-time reasoning with high computational efficiency, which however is often essential for VMI systems' deployment. Additionally, most algorithms faced the challenges of improving their adaptability on evolving new contexts and accurately and proactively reason about users' intentions.

Challenge-9. Imbalanced Data: 27 papers have explored the issue of class imbalance in VMIs, Data are often imbalanced in real-world scenarios, especially for those detection-based VMI systems, resulting in biased performance [61] or even failure to deploy. Researchers proposed several sampling and training methods, specifically targeted at maintaining VMI systems' accuracy. For instance, random oversampling and undersampling have been applied to balance classes, though oversampling can lead to overfitting, and undersampling may discard valuable data [91]. More advanced techniques, such as SMOTE [21], have been used to generate synthetic examples by interpolating between instances, though careful feature normalization is required to avoid introducing noise. Additionally, Generative Adversarial Networks have been used to augment multimodal datasets, such as adding missing text data paired with visual inputs [125]. However, as VMI systems often operate in dynamic environments with evolving data distributions, further research is needed to explore how these techniques can adapt to such changing contexts over time.

Challenge-10. Assessment Heterogeneity: A total of 11 papers have discussed the challenge of assessment heterogeneity in VMI systems, particularly in comparing the performance of different sensing modalities. Inconsistent evaluation protocols and metrics across different sensing modalities often hinder progress in this area, as they prevent meaningful comparisons between studies and affect the ability to benchmark VMI systems accurately [120]. This variability directly impacts VMI systems, where multiple modalities (e.g., vision, speech, and motion) need to be integrated and evaluated cohesively. While studies like those by Matsuda et al.[112] and Sun et al. [155] have made strides toward aligning emotion recognition performance metrics across modalities and ensuring consistency in segmentation, the absence of unified evaluation standards remains a fundamental barrier. Establishing clear, standardized protocols for VMI system evaluation is essential to facilitate accurate assessments, foster meaningful comparisons, and drive further innovations in the field.

Challenge-11. Scalable Architecture: Scalable architecture is essential for evaluating a system, especially in the context of VMIs where different modalities greatly increased the complexity of the system. This aspect is highlighted by 63 papers, where many researchers specifically examined the trade-off between on-device and cloud-based processing [113], or explored the scability in specific fields such as autonomous driving or tasks such as 3D object detection [33]. With the increasing integration of VMIs within IoT ecosystem, scalable, distributed architectures are crucial for managing heterogeneous sensor modalities, the high computational demands and real-time data processing requires. Cloud-edge hybrid models, which balance resource allocation across local and cloud systems, hold promise for addressing these needs. However, open challenges remain in optimizing these architectures for large-scale deployments, ensuring real-time data handling, and maintaining adaptability as IoT devices evolve.

Challenge-12. Power Consumption: 28 papers have addressed power consumption issues in context-aware systems, which are critical due to the high computational demands of real-time visual processing. Continuous tasks like image recognition or large-scale data analysis are common in VMIs, however require significant power, particularly for mobile or wearable systems constrained by battery life. Techniques such as dynamic frame rate adjustments and energy-efficient image sensors accordingly targeted at reducing power drain during data collection. For example, Chen et al. [25] utilized event-based visual sensing, inspired by biological systems, to achieve low data rates and reduced power consumption. Zhang et al. [194] proposed a subject-aware vocal activity sensing method that reduces power usage by avoiding unnecessary system

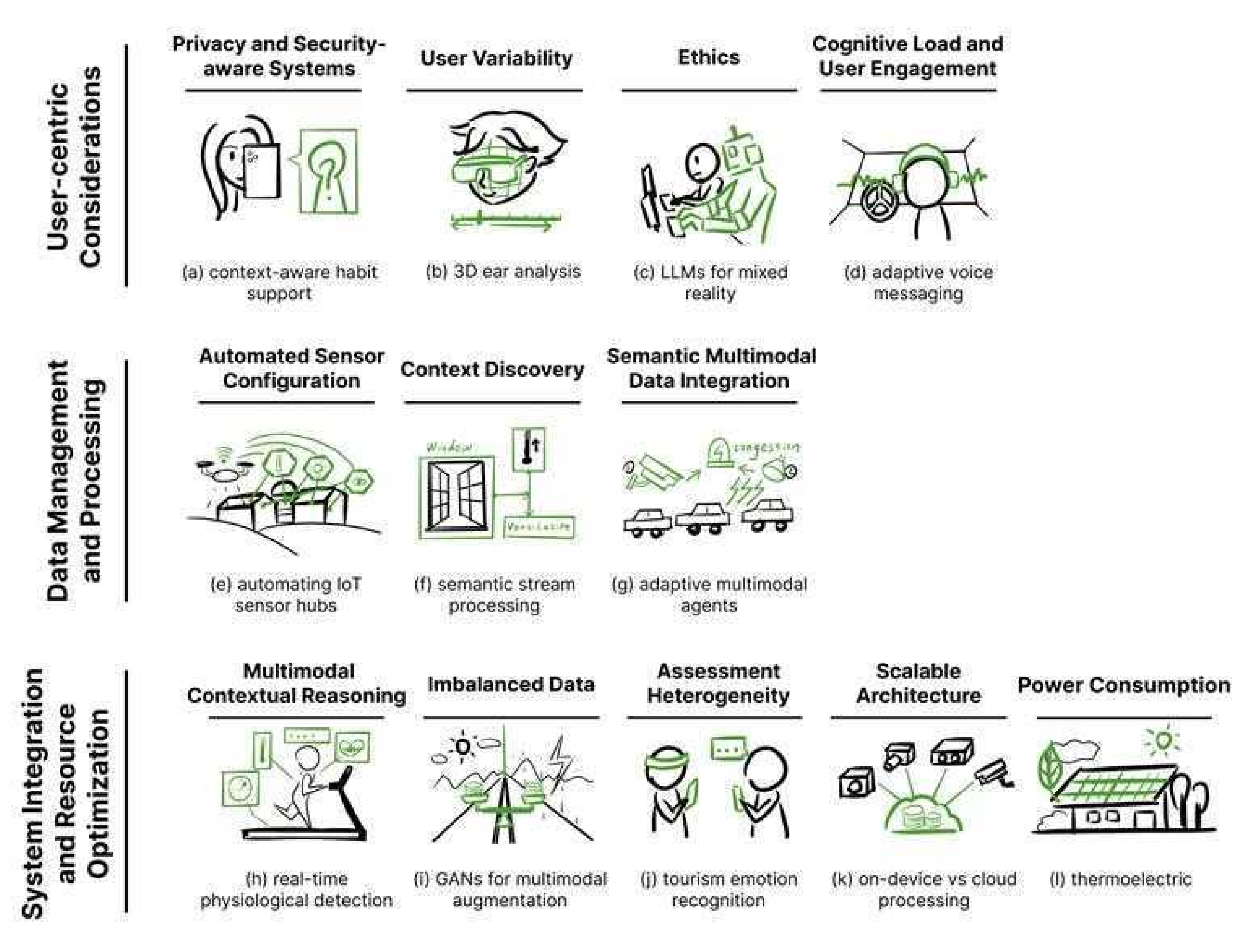

User-centric Considerations
Privacy and Security-aware Systems
User Variability
Ethics
Cognitive Load and User Engagement
(a) context-aware habit support
(b) 3D ear analysis
(c) LLMs for mixed reality
(d) adaptive voice messaging
Data Management and Processing
Automated Sensor Configuration
Context Discovery
Semantic Multimodal Data Integration
Window
Ventilation
congestion
(e) automating IoT sensor hubs
(f) semantic stream processing
(g) adaptive multimodal agents
System Integration and Resource Optimization
Multimodal Contextual Reasoning
Imbalanced Data
Assessment Heterogeneity
Scalable Architecture
Power Consumption
(h) real-time physiological detection
(i) GANs for multimodal augmentation
(j) tourism emotion recognition
(k) on-device vs cloud processing
(l) thermoelectric

Figure 6: Examples of design considerations and key challenges for VMIs (illustrative references (a): [80], (b): [94], (c): [30], (d): [176], (e): [127], (f): [150], (g): [101], (h): [83], (i): [125], (j): [112], (k): [113], (l): [16].

wake-ups. Additionally, energy harvesting technologies, such as thermoelectric or motion-based generators [16, 31], show potential for extending battery life in real-world applications. Despite these progress, ongoing challenges remain in optimizing energy efficiency without compromising performance, especially in systems requiring continuous operation.

8 Findings and Discussions

We conducted a literature survey to statistically analyze the application of VMIs in context-aware systems, presenting the results in a Sankey diagram (Figure 7). An interactive version is also available in HTML format[2]. This visualization facilitates detailed exploration and querying of the data. It intuitively illustrates the flow of contextual information across critical dimensions, offering insights into relationships among taxonomy elements.

8.1 Node Analysis

The Sankey diagram highlights key nodes. For instance, “System Factors” in Context Source Factors is referenced in 106 studies, representing 97% of the surveyed literature. This underscores its essential role in achieving resource-efficient and scalable VMIs. For example, MicroCam [67] integrates visual and IMU data to enhance context awareness in resource-constrained settings. Similarly, the “Environmental Factors” node, cited in 92 references, emphasizes the need for adaptive algorithms to address environmental variability across dimensions like “Activity,” “Location,” “Identity,” and “Time.” For instance, systems designed to dynamically respond to environmental changes, such as lighting or noise conditions [13], can improve alignment between virtual and physical environments. Practitioners should prioritize adaptive capabilities to enhance the robustness of multimodal interactions under real-world conditions.

In terms of application domains, we identified four prominent areas: “Activity Detection and Understanding,” “Content Retrieval,

[2] Source files and an interactive Sankey diagram are accessible at: https://drive.google.com/drive/folders/18dNSB9JuftudTCZss_yeJsvwflYYSJvN?usp=sharing.

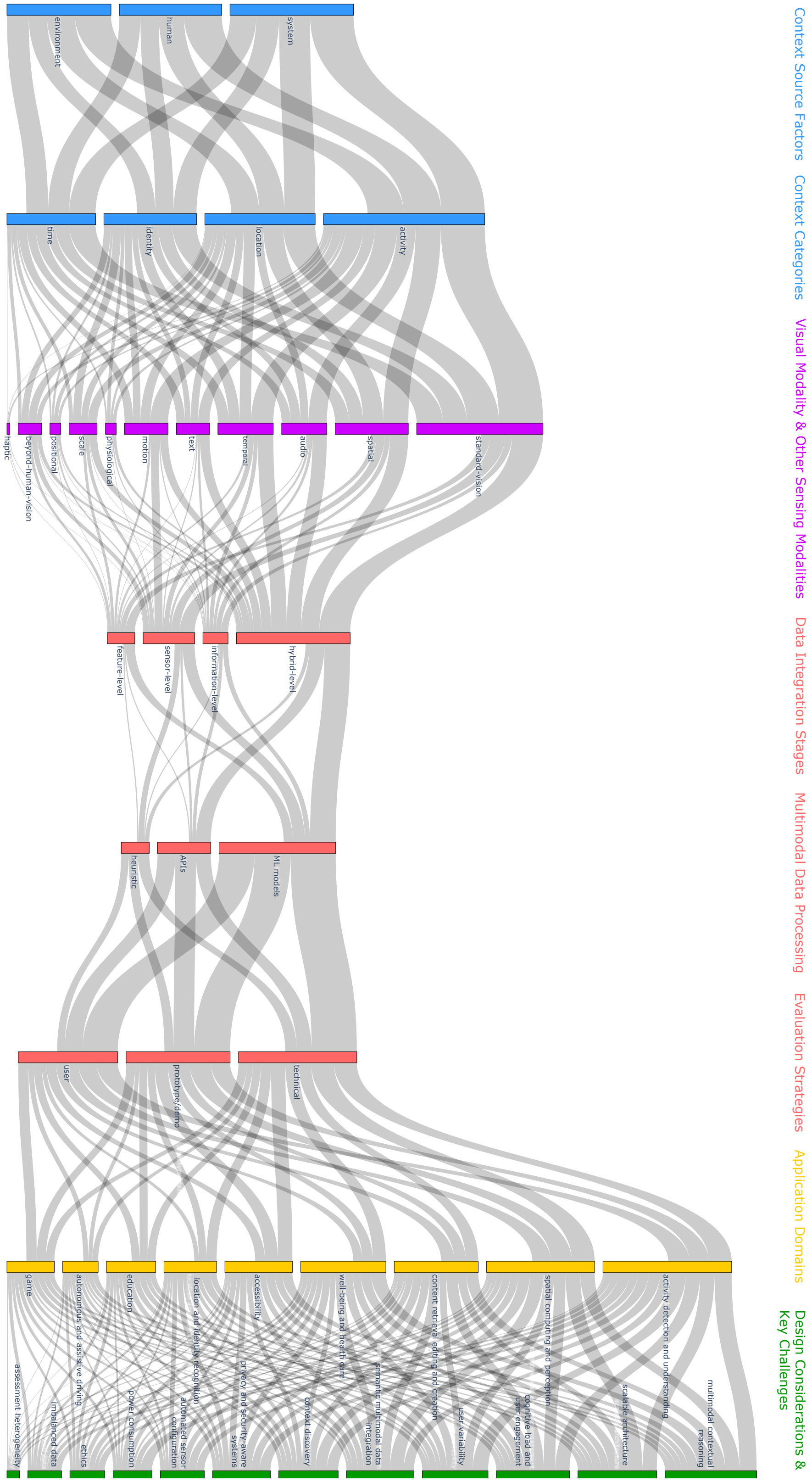


Figure 7: A Sankey diagram summarizing the overall literature counts across critical dimensions of our taxonomy. From top to bottom (from left to right), the columns represent: context source factors, context categories, visual and other sensing modalities, data integration stages, multimodal data processing, evaluation strategies, application domains, design considerations, and key challenges.

Editing, and Creation," "Spatial Computing and Perception," and "Well-being and Healthcare." Activity detection, exemplified by EyeMU [85], leverages visual and motion data for real-time user action interpretation, addressing variability across contexts. Content creation applications utilize multimodal inputs like gaze and gestures to enhance user intent interpretation and streamline workflows. In spatial computing, systems like GazePointAR [92] align virtual and physical elements, though challenges like latency and noise remain. Healthcare applications, such as VEmotion [11], integrate visual and physiological signals for holistic patient monitoring. These examples highlight the need for adaptive algorithms and robust multimodal fusion to ensure scalability and effectiveness in diverse, real-world applications.

8.2 Pathway Analysis

Figure 7 visualizes the flow of information across critical dimensions, revealing key pathways that emphasize the role of visual modalities in context-aware systems. These pathways highlight two perspectives: the standalone contributions of visual modalities to activity detection and their integration with location-based data for spatial computing.

One critical pathway, "Activity-Standard-Vision-Sensor-Level-ML Models-Activity Detection and Understanding," underscores the capabilities of visual modalities in activity recognition. For example, Pose-on-the-Go [4] combines RGB and IMU data at the sensor level to enhance real-time motion recognition. Similarly, Zhu et al. [201] demonstrate the integration of visual sensing and machine learning for robust activity detection in dynamic environments. These examples illustrate how vision-based pathways can independently enhance system responsiveness and accuracy, addressing variability in user behaviors while maintaining high performance.

Another pathway, "Location-Standard-Vision-Hybrid-APIs-Spatial Computing and Perception," highlights the integration of visual modalities with location-based data for cross-domain applications. GazePointAR [92] combines visual and positional inputs within a hybrid framework to improve spatial computing accuracy, enabling seamless virtual-physical alignment in augmented reality environments. Zimmerer et al. [202] further explore hybrid processing of LiDAR and RGB data for precise environmental mapping and navigation. These cases demonstrate the value of cross-modal integration, where visual data complements other contextual streams to address challenges like noise and latency. Designing scalable and robust integration methods is crucial to supporting diverse, multimodal applications.

8.3 Usage and Implications

Our taxonomy, including the Sankey diagram and interactive website, serves as both a practical tool and a guiding framework for designing robust VMIs. By visualizing connections across dimensions and adopting a data modality-driven perspective, it enables customized strategies for addressing real-world challenges. For example, the connection between System Factors and Activity Detection highlights the importance of hybrid data stream integration, as demonstrated by MicroCam [67], which fuses motion and visual data for surface detection. Similarly, RASSAR [153] showcases how multimodal integration improves accessibility and safety evaluations. Together, these tools facilitate efficient resource allocation, scalable designs, and systematic solutions.

The findings provide both macro and micro-level insights for designing context-aware systems. At the macro level, they identify key nodes, such as System Factors, and critical pathways shaping adaptive, scalable architectures. At the micro level, the categorization in Appendix B informs specific decisions, such as prioritizing adaptive algorithms to handle environmental variability or employing multimodal synchronization techniques. For example, EyeMU [116] demonstrates how synchronized visual and motion data enhance gesture recognition, addressing user variability. Applications like GazePointAR [92] reveal how VMIs align spatial computing with physical contexts, overcoming challenges like latency and noise. These insights emphasize the importance of robust synchronization, adaptive algorithms, and user-centric designs. Future work could explore the integration of GenAI and large language models to further enrich multimodal interactions for increasingly complex scenarios.

9 Conclusion

This research presents a taxonomy of VMIs aimed at enhancing context awareness. By synthesizing recent findings, it identifies key trends in multimodal data integration, system design, and context-aware applications. The taxonomy categorizes existing approaches across domains such as education, healthcare, accessibility, and gaming, with a focus on integrating visual modality with other inputs like audio, physiological signals, and motion. It also examines system design considerations, highlighting how VMIs process complex contextual information to improve user interactions. Open challenges, including real-time processing, data synchronization, and new interaction paradigms, are discussed to inform future research. This framework provides a foundation for developing adaptive, context-aware systems that support intuitive human-computer interactions in diverse applications.

Acknowledgments

This work was supported by the Australian Research Council (Grant Nos. DP190102068, DP200102612, LP210200656) and the Core Industry IT Convergence Program (Grant No. 20016243), facilitated by the Ministry of Trade, Industry & Energy (MOTIE, Korea).

References

[1] 2009. The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *Annals of Internal Medicine* 151, 4 (2009), W–65–W–94. https://doi.org/10.7326/0003-4819-151-4-200908180-00136 PMID: 19622512.

[2] Maythem K. Abbas, Bie Tong, and Raid Abdulla. 2018. A Hybrid Alert System for Deaf People using Context-Aware Computing and Image Processing. In *2018 4th International Conference on Computer and Information Sciences (ICCOINS)*. IEEE, 1–6. https://doi.org/10.1109/ICCOINS.2018.8510584

[3] Gregory D. Abowd, Anind K. Dey, Peter J. Brown, Nigel Davies, Mark Smith, and Pete Steggles. 1999. Towards a Better Understanding of Context and Context-Awareness. In *Handheld and Ubiquitous Computing*, Hans-W. Gellersen (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 304–307.

[4] Karan Ahuja, Sven Mayer, Mayank Goel, and Chris Harrison. 2021. Pose-on-the-Go: Approximating User Pose with Smartphone Sensor Fusion and Inverse Kinematics. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 9, 12 pages. https://doi.org/10.1145/3411764.3445582

[5] Sheeraz Athar, Gaurav Patel, Zhengtong Xu, Qiang Qiu, and Yu She. 2023. VisTac Toward a Unified Multimodal Sensing Finger for Robotic Manipulation. *IEEE Sensors Journal* 23 (2023), 25440–25450. https://api.semanticscholar.org/CorpusID:261599688

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

[7] Kimin Ban and Eui S Jung. 2020. Ear shape categorization for ergonomic product design. *International Journal of Industrial Ergonomics* 80 (2020), 102962.

[8] Hyuntae Bang, Jiyoung Min, and Haemin Jeon. 2021. Deep Learning-Based Concrete Surface Damage Monitoring Method Using Structured Lights and Depth Camera. *Sensors (Basel, Switzerland)* 21 (2021). https://api.semanticscholar.org/CorpusID:233396089

[9] Abdelkareem Bedri, Richard Li, Malcolm Haynes, Raj Prateek Kosaraju, Ishaan Grover, Temiloluwa Prioleau, Min Yan Beh, Mayank Goel, Thad Starner, and Gregory Abowd. 2017. EarBit: using wearable sensors to detect eating episodes in unconstrained environments. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 1, 3 (2017), 1–20.

[10] Hrvoje Benko, Andrew D. Wilson, and Ravin Balakrishnan. 2008. Sphere: multitouch interactions on a spherical display. In *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology* (Monterey, CA, USA) *(UIST '08)*. Association for Computing Machinery, New York, NY, USA, 77–86. https://doi.org/10.1145/1449715.1449729

[11] David Bethge, Thomas Kosch, Tobias Grosse-Puppendahl, Lewis L. Chuang, Mohamed Kari, Alexander Jagaciak, and Albrecht Schmidt. 2021. VEmotion: Using Driving Context for Indirect Emotion Prediction in Real-Time. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 638–651. https://doi.org/10.1145/3472749.3474775

[12] Anuraag Bodi, Samuel Berweger, Raied Caromi, Jihoon Bang, Jelena Senic, and Camillo Gentile. 2024. AI-Based Environment Segmentation Using a Context-Aware Channel Sounder. *2024 18th European Conference on Antennas and Propagation (EuCAP)* (2024), 1–5. https://api.semanticscholar.org/CorpusID:269389651

[13] Panagiotis-Alexandros Bokaris, Benjamin Askenazi, and Michael Haddad. 2019. Light me up: An augmented-reality projection system. In *SIGGRAPH Asia 2019 XR*. 21–22.

[14] Cristiana Bolchini, Carlo A Curino, Elisa Quintarelli, Fabio A Schreiber, and Letizia Tanca. 2007. A data-oriented survey of context models. *ACM Sigmod Record* 36, 4 (2007), 19–26.

[15] Eran Borenstein, Eitan Sharon, and Shimon Ullman. 2004. Combining top-down and bottom-up segmentation. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*. IEEE, 46–46.

[16] Jacob Bouchard-Roy, Aidin Delnavaz, and Jérémie Voix. 2020. In-ear energy harvesting: Evaluation of the power capability of the temporomandibular joint. *IEEE Sensors Journal* 20, 12 (2020), 6338–6345.

[17] John Brooke. 2013. SUS: a retrospective. *Journal of Usability Studies* 8 (01 2013), 29–40.

[18] Qiong Cai, Hao Wang, Zhenmin Li, and Xiao Liu. 2019. A survey on multimodal data-driven smart healthcare systems: approaches and applications. *IEEE Access* 7 (2019), 133583–133599.

[19] Runze Cai, Nuwan Janaka, Yang Chen, Lucia Wang, Shengdong Zhao, and Can Liu. 2024. PANDALens: Towards AI-Assisted In-Context Writing on OHMD During Travels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1053, 24 pages. https://doi.org/10.1145/3613904.3642320

[20] Yanpeng Cao, Baobei Xu, Zhangyu Ye, Jiangxin Yang, Yanlong Cao, Christel-Loïc Tisse, and Xin Li. 2018. Depth and thermal sensor fusion to enhance 3D thermographic reconstruction. *Optics express* 26 7 (2018), 8179–8193. https://api.semanticscholar.org/CorpusID:25437618

[21] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[22] Chen Chen, Cuong Nguyen, Jane Hoffswell, Jennifer Healey, Trung Bui, and Nadir Weibel. 2023. PaperToPlace: Transforming Instruction Documents into Spatialized and Context-Aware Mixed Reality Experiences. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 118, 21 pages. https://doi.org/10.1145/3586183.3606832

[23] Liuqing Chen, Yu Cai, Ruyue Wang, Shixian Ding, Yilin Tang, Preben Hansen, and Lingyun Sun. 2024. Supporting Text Entry in Virtual Reality with Large Language Models. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE, 524–534.

[24] Tuochao Chen, Yaxuan Li, Songyun Tao, Hyunchul Lim, Mose Sakashita, Ruidong Zhang, François Guimbretière, and Cheng Zhang. 2021. NeckFace: Continuously Tracking Full Facial Expressions on Neck-mounted Wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5 (2021), 1 – 31. https://api.semanticscholar.org/CorpusID:235631104

[25] Xiaoming Chen, Zeke Zexi Hu, Guangxin Zhao, Haisheng Li, Vera Chung, and Aaron Quigley. 2024. Video2Haptics: Converting Video Motion to Dynamic Haptic Feedback with Bio-Inspired Event Processing. *IEEE Transactions on Visualization and Computer Graphics* (2024).

[26] Yifei Cheng, Yukang Yan, Xin Yi, Yuanchun Shi, and David Lindlbauer. 2021. SemanticAdapt: Optimization-based Adaptation of Mixed Reality Layouts Leveraging Virtual-Physical Semantic Connections. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 282–297. https://doi.org/10.1145/3472749.3474750

[27] Youngjun Cho, Nadia Bianchi-Berthouze, Nicolai Marquardt, and Simon J. Julier. 2018. Deep Thermal Imaging: Proximate Material Type Recognition in the Wild through Deep Learning of Spatial Surface Temperature Patterns. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173576

[28] Chieh Chou, Haifeng Li, and Dezhen Song. 2020. Encoder-Camera-Ground Penetrating Radar Sensor Fusion: Bimodal Calibration and Subsurface Mapping. *IEEE Transactions on Robotics* 37 (2020), 67–81. https://api.semanticscholar.org/CorpusID:225506468

[29] Michael Compton, Cory Henson, Laurent Lefort, Holger Neuhaus, and Amit Sheth. 2009. A survey of the semantic specification of sensors. In *Proceedings of the 2nd International Conference on Semantic Sensor Networks - Volume 522* (Washington DC) *(SSN'09)*. CEUR-WS.org, Aachen, DEU, 17–32.

[30] Fernanda De La Torre, Cathy Mengying Fang, Han Huang, Andrzej Banburski-Fahey, Judith Amores Fernandez, and Jaron Lanier. 2024. Llmr: Real-time prompting of interactive worlds using large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–22.

[31] Aidin Delnavaz and Jérémie Voix. 2013. Piezo-earpiece for micro-power generation from ear canal dynamic motion. *Journal of Micromechanics and microengineering* 23, 11 (2013), 114001.

[32] Pengchao Deng, Chenyang Ge, Hao Wei, Yuan Sun, and Xin Qiao. 2023. Attention-Aware Dual-Stream Network for Multimodal Face Anti-Spoofing. *IEEE Transactions on Information Forensics and Security* 18 (2023), 4258–4271. https://api.semanticscholar.org/CorpusID:259612172

[33] Yuanzhi Deng, Cheng Chi, Huajie Wen, Yang Zhou, Gang Xu, and Jianhao Shen. 2023. Context-Aware Fusion for 3D Object Detection in LiDAR-Camera Systems. In *2023 4th International Conference on Computer Vision, Image and Deep Learning (CVIDL)*. IEEE, 601–608. https://doi.org/10.1109/CVIDL58838.2023.10166260

[34] Anind K Dey, Gregory D Abowd, and Daniel Salber. 2001. A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human–Computer Interaction* 16, 2-4 (2001), 97–166.

[35] Anind K Dey and Jennifer Mankoff. 2005. Designing mediation for context-aware applications. *ACM Transactions on Computer-Human Interaction (TOCHI)* 12, 1 (2005), 53–80.

[36] Alan Dix. 2004. *Human-computer interaction*. Vol. 1. Pearson Education.

[37] Mustafa Doga Dogan, Steven Vidal Acevedo Colon, Varnika Sinha, Kaan Akşit, and Stefanie Mueller. 2021. SensiCut: Material-Aware Laser Cutting Using Speckle Sensing and Deep Learning. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 24–38. https://doi.org/10.1145/3472749.3474733

[38] Messaoud Doudou, Abdelmadjid Bouabdallah, and Véronique Berge-Cherfaoui. 2019. Driver Drowsiness Measurement Technologies: Current Research, Market Solutions, and Challenges. *International Journal of Intelligent Transportation Systems Research* 18 (2019), 297 – 319. https://api.semanticscholar.org/CorpusID:203081957

[39] Paul Dourish. 2004. What we talk about when we talk about context. *Personal and ubiquitous computing* 8 (2004), 19–30.

[40] Bruno Dumas, Denis Lalanne, and Sharon Oviatt. 2009. *Multimodal Interfaces: A Survey of Principles, Models and Frameworks*. Vol. 5440. Springer, 3–26. https://doi.org/10.1007/978-3-642-00437-7_1

[41] Don Samitha Elvitigala, Yunfan Wang, Yongquan Hu, and Aaron J Quigley. 2023. RadarFoot: Fine-grain Ground Surface Context Awareness for Smart Shoes. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 87, 13 pages. https://doi.org/10.1145/3586183.3606738

[42] Zachary Englhardt, Chengqian Ma, Margaret E. Morris, Chun-Cheng Chang, Xuhai Orson Xu, Lianhui Qin, Daniel McDuff, Xin Liu, Shwetak N. Patel, and Vikram Iyer. 2023. From Classification to Clinical Insights. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8 (2023), 1 – 25. https://api.semanticscholar.org/CorpusID:265351685

[43] Eva Eriksson, Thomas Riisgaard Hansen, and Andreas Lykke-Olesen. 2007. Movement-based interaction in camera spaces: a conceptual framework. *Personal and Ubiquitous Computing* 11 (2007), 621–632.

[44] Xianzhe Fan, Zihan Wu, Chun Yu, Fenggui Rao, Weinan Shi, and Teng Tu. 2024. ContextCam: Bridging Context Awareness with Creative Human-AI Image Co-Creation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 157, 17 pages. https://doi.org/10.1145/3613904.3642129

[45] Andreas Fender and Jörg Müller. 2018. Velt: A Framework for Multi RGB-D Camera Systems. In *Proceedings of the 2018 ACM International Conference on Interactive Surfaces and Spaces* (Tokyo, Japan) *(ISS '18)*. Association for Computing Machinery, New York, NY, USA, 73–83. https://doi.org/10.1145/3279778.3279794

[46] Hasan Shahid Ferdous, Thuong Hoang, Zaher Joukhadar, Martin N Reinoso, Frank Vetere, David Kelly, and Louisa Remedios. 2019. " What's Happening at that Hip?" Evaluating an On-body Projection based Augmented Reality System for Physiotherapy Classroom. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[47] David Fleer and Christian Leichsenring. 2012. MISO: a context-sensitive multimodal interface for smart objects based on hand gestures and finger snaps. In *Adjunct Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology* (Cambridge, Massachusetts, USA) *(UIST Adjunct Proceedings '12)*. Association for Computing Machinery, New York, NY, USA, 93–94. https://doi.org/10.1145/2380296.2380338

[48] Fang Fu and Yan Luximon. 2020. A systematic review on ear anthropometry and its industrial design applications. *Human Factors and Ergonomics in Manufacturing & Service Industries* 30, 3 (2020), 176–194.

[49] Mana Fukasawa and Yu Nakayama. 2022. Spatial Augmented Reality Assistance System with Accelerometer and Projection Mapping at Cleaning Activities. In *ACM SIGGRAPH 2022 Posters*. 1–2.

[50] Raghuraman Gopalan and Behzad Dariush. 2009. Toward a vision based hand gesture interface for robotic grasping. In *Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'09)*. IEEE Press, St. Louis, MO, USA, 1452–1459.

[51] Jens Grubert, Tobias Langlotz, Stefanie Zollmann, and Holger Regenbrecht. 2017. Towards Pervasive Augmented Reality: Context-Awareness in Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics* 23 (2017), 1706–1724. https://api.semanticscholar.org/CorpusID:2560516

[52] Renan Guarese, João Becker, Henrique Fensterseifer, Marcelo Walter, Carla Freitas, Luciana Nedel, and Anderson Maciel. 2020. Augmented Situated Visualization for Spatial and Context-Aware Decision-Making. In *Proceedings of the 2020 International Conference on Advanced Visual Interfaces* (Salerno, Italy) *(AVI '20)*. Association for Computing Machinery, New York, NY, USA, Article 48, 5 pages. https://doi.org/10.1145/3399715.3399838

[53] Kunal Gupta, Yuewei Zhang, Tamil Selvan Gunasekaran, Prasanth Sasikumar, Nanditha Krishna, Philip Pits, Conor Russomanno, and Mark Billinghurst. 2023. SensoryScape: Context-Aware Empathic VR Photography. In *SIGGRAPH Asia 2023 XR* (Sydney, NSW, Australia) *(SA '23)*. Association for Computing Machinery, New York, NY, USA, Article 26, 2 pages. https://doi.org/10.1145/3610549.3614596

[54] R. Spencer Hallyburton, Yupei Liu, Yulong Cao, Z. Morley Mao, and Miroslav Pajic. 2022. Security Analysis of Camera-LiDAR Fusion Against Black-Box Attacks on Autonomous Vehicles. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association, Boston, MA, 1903–1920. https://www.usenix.org/conference/usenixsecurity22/presentation/hallyburton

[55] Albert Haque, Arnold Milstein, and Li Fei-Fei. 2020. Illuminating the dark spaces of healthcare with ambient intelligence. *Nature* 585, 7824 (2020), 193–202.

[56] Chris Harrison and Scott E. Hudson. 2008. Lightweight material detection for placement-aware mobile computing. In *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology* (Monterey, CA, USA) *(UIST '08)*. Association for Computing Machinery, New York, NY, USA, 279–282. https://doi.org/10.1145/1449715.1449761

[57] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (2006), 904–908. https://doi.org/10.1177/154193120605000909 arXiv:https://doi.org/10.1177/154193120605000909

[58] Andy Harter, Andy Hopper, Pete Steggles, Andy Ward, and Paul Webster. 1999. The anatomy of a context-aware application. In *Proceedings of the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking* (Seattle, Washington, USA) *(MobiCom '99)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/313451.313476

[59] Haitham Sabah Hasan and S. Abdul Kareem. 2012. Human Computer Interaction for Vision Based Hand Gesture Recognition: A Survey. In *2012 International Conference on Advanced Computer Science Applications and Technologies (ACSAT)*. IEEE, 55–60. https://doi.org/10.1109/ACSAT.2012.37

[60] Ari Hautasaari, Minami Aramaki, Rintaro Chujo, and Takeshi Naemura. 2024. EmoScribe Camera: A Virtual Camera System to Enliven Online Conferencing with Automatically Generated Emotional Text Captions. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 121, 7 pages. https://doi.org/10.1145/3613905.3650987

[61] Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* 21, 9 (2009), 1263–1284.

[62] Liwen He, Yifan Li, Mingming Fan, Liang He, and Yuhang Zhao. 2023. A Multimodal Toolkit to Support DIY Assistive Technology Creation for Blind and Low Vision People. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23 Adjunct)*. Association for Computing Machinery, New York, NY, USA, Article 3, 3 pages. https://doi.org/10.1145/3586182.3616646

[63] Tom Heath, Christian Bizer, and J Hendler. 2011. Synthesis lectures on the Semantic Web: theory and technology. *Linked data: Evolving the Web into a global data space* 1 (2011), 1–136.

[64] Trong-Vu Hoang, Quang-Binh Nguyen, Duy-Nam Ly, Khanh-Duy Le, Tam Nguyen, Minh-Triet Tran, and Trung-Nghia Le. 2024. ARtVista: Gateway To Empower Anyone Into Artist. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[65] Yongquan Hu, Wen Hu, and Aaron J. Quigley. 2024. Towards Enhanced Context Awareness with Vision-based Multimodal Interfaces. In *Adjunct Proceedings of the 26th International Conference on Mobile Human-Computer Interaction* (Melbourne, VIC, Australia) *(MobileHCI '24 Adjunct)*. Association for Computing Machinery, New York, NY, USA, Article 41, 3 pages. https://doi.org/10.1145/3640471.3686646

[66] Yongquan Hu, Black Sun, Pengcheng An, Zhuying Li, Wen Hu, and Aaron J Quigley. 2024. MultiSurf-GPT: Facilitating Context-Aware Reasoning with Large-Scale Language Models for Multimodal Surface Sensing. *arXiv preprint arXiv:2408.07311* (2024).

[67] Yongquan Hu, Hui-Shyong Yeo, Mingyue Yuan, Haoran Fan, Don Samitha Elvitigala, Wen Hu, and Aaron Quigley. 2023. Microcam: Leveraging smartphone microscope camera for context-aware contact surface sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–28.

[68] Yongquan Hu, Mingyue Yuan, Kaiqi Xian, Don Samitha Elvitigala, and Aaron Quigley. 2023. Exploring the design space of employing ai-generated content for augmented reality display. *arXiv preprint arXiv:2303.16593* (2023).

[69] Yongquan Hu, Shuning Zhang, Ting Dang, Hong Jia, Flora D. Salim, Wen Hu, and Aaron J. Quigley. 2024. Exploring Large-Scale Language Models to Evaluate EEG-Based Multimodal Data for Mental Health. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Melbourne VIC, Australia) *(UbiComp '24)*. Association for Computing Machinery, New York, NY, USA, 412–417. https://doi.org/10.1145/3675094.3678494

[70] Gang Hua and Matthew Turk. 2022. *Vision-based interaction*. Springer Nature.

[71] Ming Huang, Toshiyo Tamura, Takumi Yoshimura, Tadahiro Tsuchikawa, and Shigehiko Kanaya. 2016. Wearable deep body thermometers and their uses in continuous monitoring for daily healthcare. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 177–180. https://doi.org/10.1109/EMBC.2016.7590669

[72] Shaoshuai Huang, Xuandong Zhao, Dapeng Wei, Xinheng Song, and Yuanbo Sun. 2024. Chatbot and Fatigued Driver: Exploring the Use of LLM-Based Voice Assistants for Driving Fatigue. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 74, 8 pages. https://doi.org/10.1145/3613905.3651031

[73] Dong-Hyun Hwang, Kohei Aso, Ye Yuan, Kris Kitani, and Hideki Koike. 2020. MonoEye: Multimodal Human Motion Capture System Using A Single Ultra-Wide Fisheye Camera. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 98–111. https://doi.org/10.1145/3379337.3415856

[74] Ajune Wanis Ismail, Mark Billinghurst, and Mohd Shahrizal Sunar. 2015. Vision-Based Technique and Issues for Multimodal Interaction in Augmented Reality. In *Proceedings of the 8th International Symposium on Visual Information Communication and Interaction* (Tokyo, AA, Japan) *(VINCI '15)*. Association for Computing Machinery, New York, NY, USA, 75–82. https://doi.org/10.1145/2801040.2801058

[75] Suphanut Jamonnak, Ye Zhao, Xinyi Huang, and Md. Amiruzzaman. 2021. Geo-Context Aware Study of Vision-Based Autonomous Driving Models and Spatial Video Data. *IEEE Transactions on Visualization and Computer Graphics* PP (2021), 1–1. https://api.semanticscholar.org/CorpusID:237592808

[76] Mingyu Jin, Qinkai Yu, Dong Shu, Chong Zhang, Lizhou Fan, Wenyue Hua, Suiyuan Zhu, Yanda Meng, Zhenting Wang, Mengnan Du, and Yongfeng Zhang. 2024. Health-LLM: Personalized Retrieval-Augmented Disease Prediction System. arXiv:2402.00746 [cs.CL] https://arxiv.org/abs/2402.00746

[77] Qiao Jin, Yu Liu, Ruixuan Sun, Chen Chen, Puqi Zhou, Bo Han, Feng Qian, and Svetlana Yarosh. 2023. Collaborative Online Learning with VR Video: Roles of Collaborative Tools and Shared Video Control. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 713, 18 pages. https://doi.org/10.1145/3544548.3581395

[78] Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology* 7, 4

(2015), 396–403.

[79] R. Kernchen, P.P. Boda, K. Moessner, B. Mrohs, M. Boussard, and G. Giuliani. 2005. Multimodal user interfaces for context-aware mobile applications. In *2005 IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications*, Vol. 4. IEEE, 2268–2273 Vol. 4. https://doi.org/10.1109/PIMRC.2005.1651849

[80] Mina Khan, Glenn Fernandes, and Pattie Maes. 2021. PAL: Wearable and Personalized Habit-support Interventions in Egocentric Visual and Physiological Contexts. In *Proceedings of the Augmented Humans International Conference 2021* (Rovaniemi, Finland) *(AHs '21)*. Association for Computing Machinery, New York, NY, USA, 265–267. https://doi.org/10.1145/3458709.3458963

[81] Mohammad Kianpisheh, Alex Mariakakis, and Khai-Nghi Truong. 2024. exHAR: An Interface for Helping Non-Experts Develop and Debug Knowledge-based Human Activity Recognition Systems. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (2024), 1 – 30. https://api.semanticscholar.org/CorpusID:268286108

[82] Taejun Kim, Youngbo Aram Shim, Youngin Kim, Sunbum Kim, Jaeyeon Lee, and Geehyuk Lee. 2024. QuadStretcher: A Forearm-Worn Skin Stretch Display for Bare-Hand Interaction in AR/VR. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 409, 15 pages. https://doi.org/10.1145/3613904.3642067

[83] Kevin Koch, Martin Maritsch, Eva Van Weenen, Stefan Feuerriegel, Matthias Pfäffli, Elgar Fleisch, Wolfgang Weinmann, and Felix Wortmann. 2023. Leveraging driver vehicle and environment interaction: Machine learning using driver monitoring cameras to detect drunk driving. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 322, 32 pages. https://doi.org/10.1145/3544548.3580975

[84] M. Kolsch, M. Turk, and T. Hollerer. 2004. Vision-based interfaces for mobility. In *The First Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services, 2004. MOBIQUITOUS 2004*. IEEE, 86–94. https://doi.org/10.1109/MOBIQ.2004.1331713

[85] Andy Kong, Karan Ahuja, Mayank Goel, and Chris Harrison. 2021. EyeMU Interactions: Gaze + IMU Gestures on Mobile Devices. In *Proceedings of the 2021 International Conference on Multimodal Interaction* (Montréal, QC, Canada) *(ICMI '21)*. Association for Computing Machinery, New York, NY, USA, 577–585. https://doi.org/10.1145/3462244.3479938

[86] Yuki Kubo, Ryosuke Takada, Buntarou Shizuki, and Shin Takahashi. 2017. SynCro: context-aware user interface system for smartphone-smartwatch cross-device interaction. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 1794–1801.

[87] Y. Kuno, M. Sakamoto, K. Sakata, and Y. Shirai. 1994. Vision-based human interface with user-centered frame. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'94)*, Vol. 3. IEEE, 2023–2029 vol.3. https://doi.org/10.1109/IROS.1994.407586

[88] Alexander Kunze, Steve J. Summerskill, Russell Marshall, and Ashleigh J. Filtness. 2019. Automation transparency: implications of uncertainty communication for human-automation interaction and interfaces. *Ergonomics* 62 (2019), 345 – 360. https://api.semanticscholar.org/CorpusID:54507824

[89] Gierad Laput, Xiang 'Anthony' Chen, and Chris Harrison. 2016. SweepSense: Ad Hoc Configuration Sensing Using Reflected Swept-Frequency Ultrasonics. In *Proceedings of the 21st International Conference on Intelligent User Interfaces* (Sonoma, California, USA) *(IUI '16)*. Association for Computing Machinery, New York, NY, USA, 332–335. https://doi.org/10.1145/2856767.2856812

[90] Danh Le-Phuoc and Manfred Hauswirth. 2009. Linked open data in sensor data mashups. In *Proceedings of the 2nd International Conference on Semantic Sensor Networks - Volume 522* (Washington DC) *(SSN'09)*. CEUR-WS.org, Aachen, DEU, 1–16.

[91] Jeungchan Lee, Ishtiaq Mawla, Jieun Kim, Marco L Loggia, Ana Ortiz, Changjin Jung, Suk-Tak Chan, Jessica Gerber, Vincent J Schmithorst, Robert R Edwards, et al. 2019. Machine learning–based prediction of clinical pain using multimodal neuroimaging and autonomic metrics. *pain* 160, 3 (2019), 550–560.

[92] Jaewook Lee, Jun Wang, Elizabeth Brown, Liam Chu, Sebastian S. Rodriguez, and Jon E. Froehlich. 2024. GazePointAR: A Context-Aware Multimodal Voice Assistant for Pronoun Disambiguation in Wearable Augmented Reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 408, 20 pages. https://doi.org/10.1145/3613904.3642230

[93] John D Lee, Dary Fiorentino, Michelle L Reyes, Timothy L Brown, Omar Ahmad, James Fell, Nic Ward, and Robert Dufour. 2010. Assessing the feasibility of vehicle-based sensors to detect alcohol impairment. *Washington, DC: National Highway Traffic Safety Administration* 1, 2 (2010), 7.

[94] Wonsup Lee, Xiaopeng Yang, Hayoung Jung, Ilgeun Bok, Chulwoo Kim, Ochae Kwon, and Heecheon You. 2018. Anthropometric analysis of 3D ear scans of Koreans and Caucasians for ear product design. *Ergonomics* 61, 11 (2018), 1480–1495.

[95] Shengyu Li, Xingxing Li, Shuolong Chen, Yuxuan Zhou, and Shiwen Wang. 2023. Two-Step LiDAR/Camera/IMU Spatial and Temporal Calibration Based on Continuous-Time Trajectory Estimation. *IEEE Transactions on Industrial Electronics* 71 (2023), 3182–3191. https://api.semanticscholar.org/CorpusID:258452547

[96] Yifang Li, Nishant Vishwamitra, Hongxin Hu, and Kelly Caine. 2020. Towards a taxonomy of content sensitivity and sharing preferences for photos. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[97] Yaxuan Li, Yongjae Yoo, Antoine Weill-Duflos, and Jeremy Cooperstock. 2021. Towards Context-aware Automatic Haptic Effect Generation for Home Theatre Environments. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology* (Osaka, Japan) *(VRST '21)*. Association for Computing Machinery, New York, NY, USA, Article 13, 11 pages. https://doi.org/10.1145/3489849.3489887

[98] Zhipeng Li, Yi Fei Cheng, Yukang Yan, and David Lindlbauer. 2024. Predicting the Noticeability of Dynamic Virtual Elements in Virtual Reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.

[99] Yuanfeng Lian, Xu Shi, ShaoChen Shen, and Jing Hua. 2024. Multitask learning for image translation and salient object detection from multimodal remote sensing images. *The Visual Computer* 40, 3 (2024), 1395–1414.

[100] Chen Liang, Chun Yu, Xiaoying Wei, Xuhai Xu, Yongquan Hu, Yuntao Wang, and Yuanchun Shi. 2021. Auth+Track: Enabling Authentication Free Interaction on Smartphone by Continuous User Tracking. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 2, 16 pages. https://doi.org/10.1145/3411764.3445624

[101] Haicheng Liao, Huanming Shen, Zhenning Li, Chengyue Wang, Guofa Li, Yiming Bie, and Chengzhong Xu. 2024. GPT-4 enhanced multimodal grounding for autonomous driving: Leveraging cross-modal attention with large language models. *Communications in Transportation Research* 4 (2024), 100116. https://doi.org/10.1016/j.commtr.2023.100116

[102] Jian Liao, Kevin Van, Zhijie Xia, and Ryo Suzuki. 2024. RealityEffects: Augmenting 3D Volumetric Videos with Object-Centric Annotation and Dynamic Visual Effects. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 1248–1261.

[103] Jieun Lim, Youngji Koh, Auk Kim, and Uichin Lee. 2024. Exploring Context-Aware Mental Health Self-Tracking Using Multimodal Smart Speakers in Home Environments. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 699, 18 pages. https://doi.org/10.1145/3613904.3642846

[104] Qing Lin and Youngjoon Han. 2014. A Context-Aware-Based Audio Guidance System for Blind People Using a Multimodal Profile Model. *Sensors (Basel, Switzerland)* 14 (2014), 18670 – 18700. https://api.semanticscholar.org/CorpusID:1897887

[105] Zihan Lin, Francisco Cruz, and Eduardo Benitez Sandoval. 2024. Self context-aware emotion perception on human-robot interaction. arXiv:2401.10946 [cs.HC] https://arxiv.org/abs/2401.10946

[106] Xubo Liu, Qiushi Huang, Xinhao Mei, Haohe Liu, Qiuqiang Kong, Jianyuan Sun, Shengchen Li, Tom Ko, Yu Zhang, Lilian H. Tang, Mark D. Plumbley, Volkan Kılıç, and Wenwu Wang. 2023. Visually-Aware Audio Captioning With Adaptive Audio-Visual Attention. arXiv:2210.16428 [eess.AS] https://arxiv.org/abs/2210.16428

[107] Yiting Liu, Liang Li, Beichen Zhang, Shan Huang, Zheng-Jun Zha, and Qingming Huang. 2023. MaTCR: Modality-Aligned Thought Chain Reasoning for Multimodal Task-Oriented Dialogue Generation. In *Proceedings of the 31st ACM International Conference on Multimedia* (Ottawa ON, Canada) *(MM '23)*. Association for Computing Machinery, New York, NY, USA, 5776–5785. https://doi.org/10.1145/3581783.3612268

[108] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).

[109] Jarmo Makkonen, Ivan Avdouevski, Riitta Kerminen, and Ari Visa. 2009. *Context Awareness in Human-Computer Interaction*. IntechOpen, Rijeka, Chapter 1. https://doi.org/10.5772/7743

[110] Arnav Vaibhav Malawade, Trier Mortlock, and Mohammad Abdullah Al Faruque. 2022. HydraFusion: Context-Aware Selective Sensor Fusion for Robust and Efficient Autonomous Vehicle Perception. In *2022 ACM/IEEE 13th International Conference on Cyber-Physical Systems (ICCPS)*. IEEE, 68–79. https://doi.org/10.1109/ICCPS54341.2022.00013

[111] Nicolai Marquardt, Nathalie Henry Riche, Christian Holz, Hugo Romat, Michel Pahud, Frederik Brudy, David Ledo, Chunjong Park, Molly Jane Nicholas, Teddy Seyed, Eyal Ofek, Bongshin Lee, William A.S. Buxton, and Ken Hinckley. 2021. AirConstellations: In-Air Device Formations for Cross-Device Interaction via Multiple Spatially-Aware Armatures. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST*

'21). Association for Computing Machinery, New York, NY, USA, 1252–1268. https://doi.org/10.1145/3472749.3474820

[112] Yuki Matsuda, Dmitrii Fedotov, Yuta Takahashi, Yutaka Arakawa, Keiichi Yasumoto, and Wolfgang Minker. 2018. EmoTour: Multimodal Emotion Recognition using Physiological and Audio-Visual Features. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (Singapore, Singapore) *(UbiComp '18)*. Association for Computing Machinery, New York, NY, USA, 946–951. https://doi.org/10.1145/3267305.3267687

[113] Daniel McDuff, Kael Rowan, Piali Choudhury, Jessica Wolk, ThuVan Pham, and Mary Czerwinski. 2019. A Multimodal Emotion Sensing Platform for Building Emotion-Aware Applications. arXiv:1903.12133 [cs.HC] https://arxiv.org/abs/1903.12133

[114] Liyu Meng, Yuchen Liu, Xiaolong Liu, Zhaopei Huang, Wenqiang Jiang, Tenggan Zhang, Chuanhe Liu, and Qin Jin. 2022. Valence and Arousal Estimation based on Multimodal Temporal-Aware Features for Videos in the Wild. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2344–2351. https://doi.org/10.1109/CVPRW56347.2022.00261

[115] Johannes Meyer, Adrian Frank, Thomas Schlebusch, and Enkelejda Kasneci. 2021. A CNN-based Human Activity Recognition System Combining a Laser Feedback Interferometry Eye Movement Sensor and an IMU for Context-aware Smart Glasses. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1 – 24. https://api.semanticscholar.org/CorpusID:245568506

[116] Johannes Meyer, Adrian Frank, Thomas Schlebusch, and Enkelejda Kasneci. 2022. U-har: A convolutional approach to human activity recognition combining head and eye movements for context-aware smart glasses. *Proceedings of the ACM on Human-Computer Interaction* 6, ETRA (2022), 1–19. https://api.semanticscholar.org/CorpusID:248753659

[117] Chulhong Min, Akhil Mathur, Alessandro Montanari, and Fahim Kawsar. 2019. An early characterisation of wearing variability on motion signals for wearables. In *Proceedings of the 2019 ACM International Symposium on Wearable Computers* (London, United Kingdom) *(ISWC '19)*. Association for Computing Machinery, New York, NY, USA, 166–168. https://doi.org/10.1145/3341163.3347716

[118] Trisha Mittal, Aniket Bera, and Dinesh Manocha. 2021. Multimodal and Context-Aware Emotion Perception Model With Multiplicative Fusion. *IEEE MultiMedia* 28 (2021), 67–75. https://api.semanticscholar.org/CorpusID:234228590

[119] Gloria Anahi Molina-Barron, Rebeca Elizabeth Alvarado-Ramirez, and Angeles Aguirre-Acosta. 2023. Storytelling: Digital Narration Enhanced by Artificial Intelligence in the Metaverse. In *Proceedings of the 2023 7th International Conference on Education and E-Learning*. 33–40.

[120] Colver Ken Howe Ne, Jameel Muzaffar, Aakash Amlani, and Manohar Bance. 2021. Hearables, in-ear sensing devices for bio-signal acquisition: a narrative review. *Expert Review of Medical Devices* 18, sup1 (2021), 95–128.

[121] Ilpo Niskanen, Guoyong Duan, Erik Vartiainen, Matti Immonen, Lauri W. Hallman, Juha Kostamovaara, and Rauno Heikkilä. 2024. Enhancing point cloud data fusion through 2D thermal infrared camera and 2D lidar scanning. *Infrared Physics & Technology* (2024). https://api.semanticscholar.org/CorpusID:271042263

[122] Mina Nouredanesh, Alan Godfrey, Dylan Powell, and James Tung. 2022. Egocentric vision-based detection of surfaces: towards context-aware free-living digital biomarkers for gait and fall risk assessment. *Journal of neuroengineering and rehabilitation* 19, 1 (2022), 79.

[123] Sharon Oviatt. 2002. *Multimodal interfaces*. L. Erlbaum Associates Inc., USA, 286–304.

[124] José Ramón Padilla-López, Alexandros Andre Chaaraoui, and Francisco Flórez-Revuelta. 2015. Visual privacy protection methods: A survey. *Expert Systems with Applications* 42, 9 (2015), 4177–4195.

[125] Frederik Pahde, Mihai Puscas, Tassilo Klein, and Moin Nabi. 2020. Multimodal Prototypical Networks for Few-shot Learning. arXiv:2011.08899 [cs.CV] https://arxiv.org/abs/2011.08899

[126] Matthias Peissner, Dagmar Häbe, Doris Janssen, and Thomas Sellner. 2012. MyUI: generating accessible user interfaces from multimodal design patterns. In *Proceedings of the 4th ACM SIGCHI Symposium on Engineering Interactive Computing Systems* (Copenhagen, Denmark) *(EICS '12)*. Association for Computing Machinery, New York, NY, USA, 81–90. https://doi.org/10.1145/2305484.2305500

[127] Charith Perera, Prem Jayaraman, Arkady Zaslavsky, Peter Christen, and Dimitrios Georgakopoulos. 2013. Dynamic configuration of sensors using mobile sensor hub in internet of things paradigm. In *2013 IEEE Eighth International Conference on Intelligent Sensors, Sensor Networks and Information Processing*. IEEE, 473–478. https://doi.org/10.1109/ISSNIP.2013.6529836

[128] Pavan Kartheek Rachabatuni, Filippo Principi, Paolo Mazzanti, and Marco Bertini. 2024. Context-aware chatbot using MLLMs for Cultural Heritage. In *Proceedings of the 15th ACM Multimedia Systems Conference* (Bari, Italy) *(MMSys '24)*. Association for Computing Machinery, New York, NY, USA, 459–463. https://doi.org/10.1145/3625468.3652193

[129] Yogesh Singh Rawat and M. Kankanhalli. 2017. ClickSmart: A Context-Aware Viewpoint Recommendation System for Mobile Photography. *IEEE Transactions on Circuits and Systems for Video Technology* 27 (2017), 149–158. https://api.semanticscholar.org/CorpusID:9415762

[130] Natalie Ruiz, Fang Chen, and Sharon Oviatt. 2010. Chapter 12 - Multimodal Input. In *Multimodal Signal Processing*, Jean-Philippe Thiran, Ferran Marqués, and Hervé Bourlard (Eds.). Academic Press, Oxford, 231–255. https://doi.org/10.1016/B978-0-12-374825-6.00010-1

[131] Vítor Sá, Cornelius Malerczyk, and Michael Schnaider. 2001. Vision-Based Interaction within a Multimodal Framework. In *Proceedings of the 10th Conference of the Eurographics Portuguese Chapter*. The Eurographics Association. https://doi.org/10.2312/pt.20011318

[132] Alia Saad, Kian Izadi, Anam Ahmad Khan, Pascal Knierim, Stefan Schneegass, Florian Alt, and Yomna Abdelrahman. 2023. HotFoot: Foot-Based User Identification Using Thermal Imaging. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 262, 13 pages. https://doi.org/10.1145/3544548.3580924

[133] Daniel Salber. 2000. Context-awareness and multimodality. In *Colloque sur la multimodalité*.

[134] Javad Sameri, Sam Van Damme, Susanna Schwarzmann, Qing Wei, Riccardo Trivisonno, Filip De Turck, and Maria Torres Vega. 2024. Collaborative Cooking in VR: Effects of Network Distortion in Multi-User Virtual Environments. In *Proceedings of the 15th ACM Multimedia Systems Conference* (Bari, Italy) *(MMSys '24)*. Association for Computing Machinery, New York, NY, USA, 509–515. https://doi.org/10.1145/3625468.3652201

[135] B. Schilit, N. Adams, and R. Want. 1994. Context-Aware Computing Applications. In *1994 First Workshop on Mobile Computing Systems and Applications*. IEEE, 85–90. https://doi.org/10.1109/WMCSA.1994.16

[136] Albrecht Schmidt. 2000. Implicit human computer interaction through context. *Personal technologies* 4 (2000), 191–199.

[137] Maximilian Schrapel, Philipp Etgeton, and Michael Rohs. 2021. SpectroPhone: Enabling Material Surface Sensing with Rear Camera and Flashlight LEDs. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 336, 5 pages. https://doi.org/10.1145/3411763.3451753

[138] Björn W. Schuller, Tuomas Virtanen, Maria Riveiro, Georgios Rizos, Jing Han, Annamaria Mesaros, and Konstantinos Drossos. 2021. Towards Sonification in Multimodal and User-friendlyExplainable Artificial Intelligence. In *Proceedings of the 2021 International Conference on Multimodal Interaction* (Montréal, QC, Canada) *(ICMI '21)*. Association for Computing Machinery, New York, NY, USA, 788–792. https://doi.org/10.1145/3462244.3479879

[139] Tristan Schwörer, Jonathan Eichild Schmidt, and Dimitrios Chrysostomou. 2023. Nav2CAN: Achieving Context Aware Navigation in ROS2 Using Nav2 and RGB-D sensing. In *2023 IEEE International Conference on Imaging Systems and Techniques (IST)*. IEEE Signal Processing Society, United States, 1–6. https://doi.org/10.1109/IST59124.2023.10355731

[140] Omer Berat Sezer, Erdogan Dogdu, and Ahmet Murat Ozbayoglu. 2017. Context-aware computing, learning, and big data in internet of things: a survey. *IEEE Internet of Things Journal* 5, 1 (2017), 1–27.

[141] Rajeev Sharma, Vladimir I Pavlovic, and Thomas S Huang. 1998. Toward multimodal human-computer interface. *Proc. IEEE* 86, 5 (1998), 853–869.

[142] Mali Shen, Yun Gu, Ning Liu, and Guang-Zhong Yang. 2019. Context-Aware Depth and Pose Estimation for Bronchoscopic Navigation. *IEEE Robotics and Automation Letters* 4 (2019), 732–739. https://api.semanticscholar.org/CorpusID:59619567

[143] Xiyuan Shen, Chun Yu, Xutong Wang, Chen Liang, Haozhan Chen, and Yuanchun Shi. 2024. MouseRing: Always-available Touchpad Interaction with IMU Rings. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 412, 19 pages. https://doi.org/10.1145/3613904.3642225

[144] Jingyu Shi, Rahul Jain, Hyungjun Doh, Ryo Suzuki, and Karthik Ramani. 2024. An HCI-Centric Survey and Taxonomy of Human-Generative-AI Interactions. arXiv:2310.07127 [cs.HC] https://arxiv.org/abs/2310.07127

[145] Gihan Shin and Junchul Chun. 2007. Vision-Based Multimodal Human Computer Interface Based on Parallel Tracking of Eye and Hand Motion. In *2007 International Conference on Convergence Information Technology (ICCIT 2007)*. IEEE, 2443–2448. https://doi.org/10.1109/ICCIT.2007.142

[146] Alon Shoa, Ramon Oliva, Mel Slater, and Doron Friedman. 2023. Sushi with Einstein: Enhancing Hybrid Live Events with LLM-Based Virtual Humans. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*. 1–6.

[147] Muhammad Hameed Siddiqi, Nabil Almashfi, Amjad Ali, Madallah Alruwaili, Yousef Alhwaiti, Saad Awadh Alanazi, and M. M. Kamruzzaman. 2021. A Unified Approach for Patient Activity Recognition in Healthcare Using Depth Camera. *IEEE Access* 9 (2021), 92300–92317. https://api.semanticscholar.org/CorpusID:235780004

[148] Amit Kumar Sikder, Leonardo Babun, Z Berkay Celik, Hidayet Aksu, Patrick McDaniel, Engin Kirda, and A Selcuk Uluagac. 2022. Who's controlling my device? Multi-user multi-device-aware access control system for shared smart home environment. *ACM Transactions on Internet of Things* 3, 4 (2022), 1–39.

[149] Rukshani Somarathna, Don Samitha Elvitigala, Yijun Yan, Aaron J Quigley, and Gelareh Mohammadi. 2023. Exploring User Engagement in Immersive Virtual Reality Games through Multimodal Body Movements. In *Proceedings of the 29th ACM Symposium on Virtual Reality Software and Technology* (Christchurch, New Zealand) *(VRST '23)*. Association for Computing Machinery, New York, NY, USA, Article 3, 8 pages. https://doi.org/10.1145/3611659.3615687

[150] Dimitrios-Emmanuel Spanos, Periklis Stavrou, Nikolas Mitrou, and Nikolas Konstantinou. 2012. SensorStream: A semantic real–time stream management system. *International Journal of Ad Hoc and Ubiquitous Computing* 11, 2-3 (2012), 178–193.

[151] Thad Starner. 1995. *Visual recognition of american sign language using hidden markov models.* Ph. D. Dissertation. Massachusetts Institute of Technology.

[152] Xia Su, Eunyee Koh, and Chang Xiao. 2024. SonifyAR: Context-Aware Sound Effect Generation in Augmented Reality. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 297, 7 pages. https://doi.org/10.1145/3613905.3650927

[153] Xia Su, Han Zhang, Kaiming Cheng, Jaewook Lee, Qiaochu Liu, Wyatt Olson, and Jon E. Froehlich. 2024. RASSAR: Room Accessibility and Safety Scanning in Augmented Reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 968, 17 pages. https://doi.org/10.1145/3613904.3642140

[154] Zixiong Su, Xinlei Zhang, Naoki Kimura, and Jun Rekimoto. 2021. Gaze+ Lip: rapid, precise and expressive interactions combining gaze input and silent speech commands for hands-free smart TV control. In *ACM symposium on eye tracking research and applications*. 1–6.

[155] Fengyuan Sun, Sezer Karaoglu, and Theo Gevers. 2023. Temporally Consistent Semantic Segmentation using Spatially Aware Multi-view Semantic Fusion for Indoor RGB-D videos. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. IEEE, 4250–4259. https://doi.org/10.1109/ICCVW60793.2023.00459

[156] Xin Suo, Minye Wu, Yanshun Zhang, Yingliang Zhang, Lan Xu, Qiang Hu, and Jingyi Yu. 2020. Neural3D: Light-weight Neural Portrait Scanning via Context-aware Correspondence Learning. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 3651–3660. https://doi.org/10.1145/3394171.3413734

[157] Hemant Bhaskar Surale, Aakar Gupta, Mark Hancock, and Daniel Vogel. 2019. TabletInVR: Exploring the Design Space for Using a Multi-Touch Tablet in Virtual Reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300243

[158] Ryo Suzuki, Adnan Karim, Tian Xia, Hooman Hedayati, and Nicolai Marquardt. 2022. Augmented Reality and Robotics: A Survey and Taxonomy for AR-enhanced Human-Robot Interaction and Robotic Interfaces. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 553, 33 pages. https://doi.org/10.1145/3491102.3517719

[159] Ryosuke Suzuki, Tadachika Ozono, and Toramatsu Shintani. 2019. An Offline Mahjong Support System Based on Augmented Reality with Context-aware Image Recognition. In *2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI)*. IEEE, 127–132. https://doi.org/10.1109/IIAI-AAI.2019.00035

[160] Anil Kumar Thandlam, T. Vignesh, A Saravanan, Prakash Subramani, Ramaganesh Marimuthu, and Sandeep Gupta. 2024. "Next-Gen Vehicle Safety: The Futuristic Approach to Auto-Stop in Modern Vehicles". In *2024 10th International Conference on Communication and Signal Processing (ICCSP)*. IEEE, 265–270. https://doi.org/10.1109/ICCSP60870.2024.10543666

[161] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. arXiv:2406.16860 [cs.CV] https://arxiv.org/abs/2406.16860

[162] Hsin-Ruey Tsai, Shih-Kang Chiu, and Bryan Wang. 2024. GazeNoter: Co-Piloted AR Note-Taking via Gaze Selection of LLM Suggestions to Match Users' Intentions. arXiv:2407.01161 [cs.HC] https://arxiv.org/abs/2407.01161

[163] Mikael Uimonen, Paul Kemppi, and Taru Hakanen. 2023. A Gesture-based Multimodal Interface for Human-Robot Interaction. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, IEEE, 165–170. https://doi.org/10.1109/RO-MAN57019.2023.10309404

[164] Chongyang Wang, Yuan Feng, Lin Xiao Zhong, Siyi Zhu, Chi Zhang, Siqi Zheng, Chen Liang, Yuntao Wang, Chen-Jun He, Chun Yu, and Yuanchun Shi. 2023. UbiPhysio: Support Daily Functioning, Fitness, and Rehabilitation with Action Understanding and Feedback in Natural Language. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8 (2023), 1 – 27. https://api.semanticscholar.org/CorpusID:261049773

[165] Chongyang Wang, Siqi Zheng, Lingxiao Zhong, Chun Yu, Chen Liang, Yuntao Wang, Yuan Gao, Tin Lun Lam, and Yuanchun Shi. 2024. PepperPose: Full-Body Pose Estimation with a Companion Robot. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 586, 16 pages. https://doi.org/10.1145/3613904.3642231

[166] Ru Wang, Zach Potter, Yun Ho, Daniel Killough, Linxiu Zeng, Sanbrita Mondal, and Yuhang Zhao. 2024. GazePrompt: Enhancing Low Vision People's Reading Experience with Gaze-Aware Augmentations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 894, 17 pages. https://doi.org/10.1145/3613904.3642878

[167] Tianyi Wang, Xun Qian, Fengming He, Xiyun Hu, Ke Huo, Yuanzhi Cao, and Karthik Ramani. 2020. CAPturAR: An Augmented Reality Tool for Authoring Human-Involved Context-Aware Applications. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 328–341. https://doi.org/10.1145/3379337.3415815

[168] Xue Wang, Zixiong Su, Jun Rekimoto, and Yang Zhang. 2024. Watch Your Mouth: Silent Speech Recognition with Depth Sensing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 323, 15 pages. https://doi.org/10.1145/3613904.3642092

[169] X.H. Wang, D.Q. Zhang, T. Gu, and H.K. Pung. 2004. Ontology based context modeling and reasoning using OWL. In *IEEE Annual Conference on Pervasive Computing and Communications Workshops, 2004. Proceedings of the Second*. IEEE, 18–22. https://doi.org/10.1109/PERCOMW.2004.1276898

[170] Yikai Wang, Wenbing Huang, Bin Fang, Fuchun Sun, and Chang Li. 2021. Elastic tactile simulation towards tactile-visual perception. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2690–2698.

[171] Zeyu Wang, Yuanchun Shi, Yuntao Wang, Yuchen Yao, Kun Yan, Yuhan Wang, Lei Ji, Xuhai Xu, and Chun Yu. 2024. G-VOILA: Gaze-Facilitated Information Querying in Daily Scenarios. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8 (2024), 1 – 33. https://api.semanticscholar.org/CorpusID:269758074

[172] Zhan Wang, Lin-Ping Yuan, Liangwei Wang, Bingchuan Jiang, and Wei Zeng. 2024. Virtuwander: Enhancing multi-modal interaction for virtual tour guidance through large language models. In *Proceedings of the CHI conference on human factors in computing systems*. 1–20.

[173] Mark Weiser. 1999. The computer for the 21st century. *ACM SIGMOBILE mobile computing and communications review* 3, 3 (1999), 3–11.

[174] Mark Weiser and John Seely Brown. 1996. Designing calm technology. *PowerGrid Journal* 1, 1 (1996), 75–85.

[175] Linda Yilin Wen, Cecily Morrison, Martin Grayson, Rita Faia Marques, Daniela Massiceti, Camilla Longden, and Edward Cutrell. 2024. Find My Things: Personalized Accessibility through Teachable AI for People who are Blind or Low Vision. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (2024). https://api.semanticscholar.org/CorpusID:269748025

[176] Shaoyue Wen, Songming Ping, Jialin Wang, Hai-Ning Liang, Xuhai Xu, and Yukang Yan. 2024. AdaptiveVoice: Cognitively Adaptive Voice Interface for Driving Assistance. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 253, 18 pages. https://doi.org/10.1145/3613904.3642876

[177] Johann Wentzel, Fraser Anderson, George Fitzmaurice, Tovi Grossman, and Daniel Vogel. 2024. SwitchSpace: Understanding Context-Aware Peeking Between VR and Desktop Interfaces. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 801, 16 pages. https://doi.org/10.1145/3613904.3642358

[178] Erwin Wu, Chen-Chieh Liao, Ruofan Liu, and Hideki Koike. 2022. Context-aware Risk Degree Prediction for Smartphone Zombies. In *ACM SIGGRAPH 2022 Posters* (Vancouver, BC, Canada) *(SIGGRAPH '22)*. Association for Computing Machinery, New York, NY, USA, Article 48, 2 pages. https://doi.org/10.1145/3532719.3543197

[179] Xinyu Xie, Xiaozhi Zhang, Dongping Xiong, and Lijun Ouyang. 2023. MFA-DAF: Unsupervised Multimodal Medical Image Fusion via Multiscale Fourier Attention and Detail-Aware Fusion Strategy. *2023 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)* (2023), 208–214. https://api.semanticscholar.org/CorpusID:267659859

[180] Anran Xu, Shitao Fang, Huan Yang, Simo Hosio, and Koji Yatani. 2024. Examining Human Perception of Generative Content Replacement in Image Privacy Protection. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.

[181] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2024. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 1, Article 31 (March 2024), 32 pages. https://doi.org/10.1145/3643540

[182] Yating Xu, Conghui Hu, and Gim Hee Lee. 2022. Motion and Context-Aware Audio-Visual Conditioned Video Prediction. In *British Machine Vision Conference*. https://api.semanticscholar.org/CorpusID:254536032

[183] Zhenyu Xu, Hailin Xu, Zhouyang Lu, Yingying Zhao, Rui Zhu, Yujiang Wang, Mingzhi Dong, Yuhu Chang, Qin Lv, Robert P Dick, et al. 2024. Can Large Language Models Be Good Companions? An LLM-Based Eyewear System with Conversational Common Ground. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2, Article 87 (May 2024), 41 pages. https://doi.org/10.1145/3659600

[184] Songlin Yang, Wei Wang, Jun Ling, Bo Peng, Xu Tan, and Jing Dong. 2023. Context-Aware Talking-Head Video Editing. In *Proceedings of the 31st ACM International Conference on Multimedia* (Ottawa ON, Canada) *(MM '23)*. Association for Computing Machinery, New York, NY, USA, 7718–7727. https://doi.org/10.1145/3581783.3611765

[185] Xing-Dong Yang, Tovi Grossman, Daniel Wigdor, and George Fitzmaurice. 2012. Magic finger: always-available input through finger instrumentation. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology* (Cambridge, Massachusetts, USA) *(UIST '12)*. Association for Computing Machinery, New York, NY, USA, 147–156. https://doi.org/10.1145/2380116.2380137

[186] Mengmei Ye, Zhongze Tang, Huy Phan, Yi Xie, Bo Yuan, and Sheng Wei. 2022. Visual privacy protection in mobile image recognition using protective perturbation. In *Proceedings of the 13th ACM Multimedia Systems Conference*. 164–176.

[187] Hui-Shyong Yeo, Juyoung Lee, Andrea Bianchi, David Harris-Birtill, and Aaron Quigley. 2017. SpeCam: sensing surface color and material with the front-facing camera of a mobile device. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Vienna, Austria) *(MobileHCI '17)*. Association for Computing Machinery, New York, NY, USA, Article 25, 9 pages. https://doi.org/10.1145/3098279.3098541

[188] Nur Yildirim, Hannah Richardson, Maria Teodora Wetscherek, Junaid Bajwa, Joseph Jacob, Mark Ames Pinnock, Stephen Harris, Daniel Coelho De Castro, Shruthi Bannur, Stephanie Hyland, Pratik Ghosh, Mercy Ranjit, Kenza Bouzid, Anton Schwaighofer, Fernando Pérez-García, Harshita Sharma, Ozan Oktay, Matthew Lungren, Javier Alvarez-Valle, Aditya Nori, and Anja Thieme. 2024. Multimodal Healthcare AI: Identifying and Designing Clinically Relevant Vision-Language Applications for Radiology. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 444, 22 pages. https://doi.org/10.1145/3613904.3642013

[189] Zhizhuo Yin, Yuyang Wang, Theodoros Papatheodorou, and Pan Hui. 2024. Text2VRScene: Exploring the Framework of Automated Text-driven Generation System for VR Experience. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE, 701–711.

[190] Watcharaphong Yookwan, Krisana Chinnasarn, C. So-In, and Paramate Horkaew. 2022. Multimodal Fusion of Deeply Inferred Point Clouds for 3D Scene Reconstruction Using Cross-Entropy ICP. *IEEE Access* 10 (2022), 77123–77136. https://api.semanticscholar.org/CorpusID:250976713

[191] Xenophon Zabulis, Haris Baltzakis, and Antonis A Argyros. 2009. Vision-Based Hand Gesture Recognition for Human-Computer Interaction. *The universal access handbook* 34 (2009), 30.

[192] Nima Zargham, Mohamed Lamine Fetni, Laura Spillner, Thomas Muender, and Rainer Malaka. 2024. " 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*. 1–18.

[193] Chao Zhang, Xuechen Liu, Katherine Ziska, Soobin Jeon, Chi-Lin Yu, and Ying Xu. 2024. Mathemyths: Leveraging Large Language Models to Teach Mathematical Language through Child-AI Co-Creative Storytelling. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 274, 23 pages. https://doi.org/10.1145/3613904.3642647

[194] Xiyuxing Zhang, Yuntao Wang, Yuxuan Han, Chen Liang, Ishan Chatterjee, Jiankai Tang, Xin Yi, Shwetak Patel, and Yuanchun Shi. 2024. The EarSAVAS Dataset: Enabling Subject-Aware Vocal Activity Sensing on Earables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–26.

[195] Fei Zhao, Chengcui Zhang, and Baocheng Geng. 2024. Deep Multimodal Data Fusion. *ACM Comput. Surv.* 56, 9, Article 216 (April 2024), 36 pages. https://doi.org/10.1145/3649447

[196] Nan Zhao, Elena Kodama, and Joseph A. Paradiso. 2022. Mediated Atmosphere Table (MAT): Adaptive Multimodal Media System for Stress Restoration. *IEEE Internet of Things Journal* 9 (2022), 23614–23625. https://api.semanticscholar.org/CorpusID:250564158

[197] Y. Zheng, Yuejie Zhang, Rui Feng, Tao Zhang, and Weiguo Fan. 2021. Stacked Multimodal Attention Network for Context-Aware Video Captioning. *IEEE Transactions on Circuits and Systems for Video Technology* 32 (2021), 31–42. https://api.semanticscholar.org/CorpusID:236657677

[198] Dingfu Zhou, Xibin Song, Jin Fang, Yuchao Dai, Hongdong Li, and Liangjun Zhang. 2022. Context-Aware 3D Object Detection From a Single Image in Autonomous Driving. *IEEE Transactions on Intelligent Transportation Systems* 23 (2022), 18568–18580. https://api.semanticscholar.org/CorpusID:247339223

[199] Zhongyi Zhou, Jing Jin, Vrushank Phadnis, Xiuxiu Yuan, Jun Jiang, Xun Qian, Jingtao Zhou, Yiyi Huang, Zheng Xu, Yinda Zhang, et al. 2023. InstructPipe: Building Visual Programming Pipelines with Human Instructions. *arXiv preprint arXiv:2312.09672* (2023).

[200] Zhongyi Zhou and Koji Yatani. 2022. Gesture-aware Interactive Machine Teaching with In-situ Object Annotations. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 27, 14 pages. https://doi.org/10.1145/3526113.3545648

[201] Rongrong Zhu, Liang Shi, Yunpeng Song, and Zhongmin Cai. 2023. Integrating Gaze and Mouse Via Joint Cross-Attention Fusion Net for Students' Activity Recognition in E-learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7 (2023), 1 – 35. https://api.semanticscholar.org/CorpusID:263153323

[202] Chris Zimmerer, Martin Fischbach, and Marc Erich Latoschik. 2022. A Case Study on the Rapid Development of Natural and Synergistic Multimodal Interfaces for XR Use-Cases. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 19, 8 pages. https://doi.org/10.1145/3491101.3503552

A Appendix A: PRISMA-style Literature Selection

Following the PRISMA framework [1], we conducted a four-stage systematic review, summarized in Figure 8.

Initial Phase: A keyword search was conducted in digital libraries, including ACM and IEEE databases, using `("vision-based") AND ("multimodal") AND ("context aware")` and related synonyms (e.g., `"vision"`, `"camera"`, `"visual"`). This search yielded 4,494 works. A temporal filter was applied, excluding 1,723 items published in or before 2018.

Screening Phase: After removing duplicates and non-computational works, 1,842 exclusions were made. The remaining works were reviewed independently by two pairs of coders to ensure consistent application of criteria and inter-rater reliability [1]. Coders assessed abstracts and full texts against predefined inclusion and exclusion criteria, and conflicts were resolved through the following steps:

- Flagging conflicting cases for discussion in bi-weekly meetings.
- Applying a predefined resolution protocol, prioritizing alignment with the study scope.
- Revisiting discrepancies after each meeting to ensure a unified approach.

Eligibility Phase: Specific exclusion criteria were applied, resulting in the removal of 831 works and leaving 98 eligible papers. The excluded works fell into the following categories:

- Incomplete research processes, such as workshops, symposia, tutorials, or technical briefs (99).
- Mentioning multimodality without utilizing it in the final system, or focusing solely on a single modality (306).
- Conceptual or theoretical works lacking demonstrable implementation, including case studies without prototypes or demos (64).
- Outside the HCI domain, unpublished in HCI-related venues (e.g., CHI, UIST), or lacking relevant keywords (297).
- Misaligned with the definition of VMIs, as described in Section 2.1.3 (65).

Final Phase: Expert discussions added 11 relevant works to the dataset, resulting in a curated collection of 109 papers, primarily sourced from ACM (65%), IEEE (23%), and other libraries (12%).

To ensure consistency and reduce subjectivity in coding, the following procedures were implemented across the phases:

- Initial coding was performed on a small subset of the dataset to identify recurring themes and patterns.
- Discrepancies in coding were flagged and discussed during iterative group meetings, with consensus reached through majority agreement. In critical cases, an independent expert was consulted.
- Categories were refined through successive iterations, merging similar ones and expanding underrepresented dimensions.
- A final validation step was conducted on a randomly sampled subset (10%) of the dataset to verify inter-coder reliability and alignment.

Through these steps, the final categorization of dimensions converged, ensuring robustness and consistency in the analysis process.

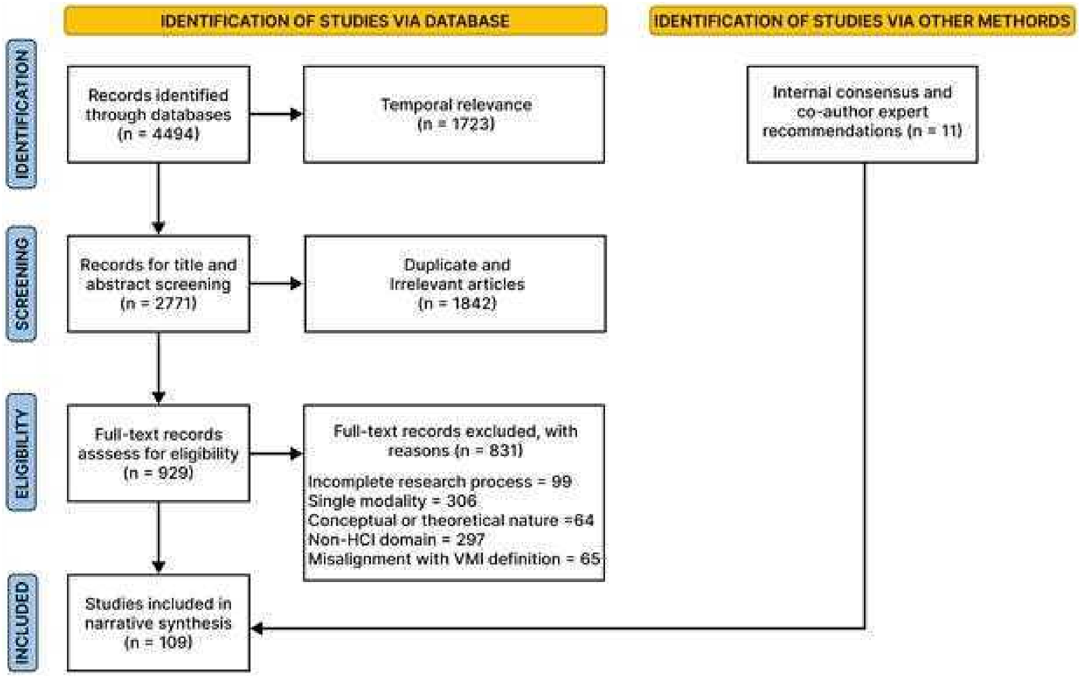

IDENTIFICATION OF STUDIES VIA DATABASE
IDENTIFICATION OF STUDIES VIA OTHER METHORDS
IDENTIFICATION
Records identified through databases (n = 4494)
Temporal relevance (n = 1723)
Internal consensus and co-author expert recommendations (n = 11)
SCREENING
Records for title and abstract screening (n = 2771)
Duplicate and Irrelevant articles (n = 1842)
ELIGIBILITY
Full-text records asssess for eligibility (n = 929)
Full-text records excluded, with reasons (n = 831)
Incomplete research process = 99
Single modality = 306
Conceptual or theoretical nature =64
Non-HCI domain = 297
Misalignment with VMI definition = 65
INCLUDED
Studies included in narrative synthesis (n = 109)

Figure 8: A PRISMA-style flowchart of the selection of studies for the systematic review and meta-analysis.

B Appendix B: Details of the Survey Literature Statistics

Category	Count	Citations
Section 3: Context of VMIs		
Context Source Factors		
→ *Human Factors*	87	[4, 8, 11–13, 19, 22, 24–26, 30, 33, 37, 44, 45, 47, 49, 52–54, 56, 60, 64, 66–69, 75, 80, 81, 83, 85, 86, 92, 95, 97–100, 102, 105–107, 110–112, 115, 116, 118, 119, 121, 122, 128, 129, 132, 137, 139, 143, 146, 149, 152, 153, 159, 162–165, 167, 168, 170–172, 175–178, 182–184, 187, 189, 190, 194, 196, 197, 200, 201]
→ *Environment Factors*	92	[4, 5, 8, 11, 12, 19, 20, 23, 25–28, 30, 32, 33, 37, 44–47, 49, 52, 54, 56, 64, 66, 67, 69, 73, 75, 81, 83, 85, 86, 92, 95, 97–101, 103–107, 110, 111, 113–116, 118, 119, 121, 122, 128, 132, 137, 139, 142, 143, 147, 149, 152–155, 162–165, 167, 168, 170–172, 176–179, 182, 184, 187, 189, 190, 192, 196–198, 200, 202]
→ *System Factors*	106	[4, 5, 8, 11–13, 19, 20, 22–28, 30, 32, 33, 37, 44–47, 49, 52–54, 56, 60, 64, 66–69, 73, 75, 80, 81, 83, 85, 86, 92, 95, 97–107, 110–116, 118, 119, 121, 122, 128, 129, 132, 137, 139, 142, 143, 146, 147, 149, 152–155, 159, 162–165, 167, 168, 170–172, 175, 177–179, 182–185, 189, 190, 192, 194, 196, 198, 200–202]
Context Categories		
→ *Activity*	78	[4, 5, 13, 19, 22, 23, 25, 30, 45–47, 49, 56, 60, 64, 66–69, 73, 75, 80, 81, 83, 85, 86, 97, 98, 100–107, 110–112, 115, 116, 118, 119, 122, 128, 129, 139, 143, 146, 147, 149, 152, 154, 159, 162–165, 167, 168, 170–172, 175–178, 182–185, 192, 194, 196, 197, 200–202]
→ *Location*	56	[8, 11, 19, 20, 25–28, 30, 44–46, 49, 52, 54, 56, 66–68, 75, 80, 81, 85, 86, 92, 95, 103, 111, 112, 118, 119, 121, 129, 132, 137, 139, 142, 152, 153, 155, 159, 167, 168, 170, 172, 176–179, 183, 185, 187, 190, 192, 198, 202]
→ *Identity*	42	[8, 11–13, 19, 24, 27, 30, 32, 33, 37, 44, 46, 53, 56, 60, 66, 67, 80, 99–101, 104, 105, 112–114, 118, 122, 137, 146, 152, 153, 175–177, 185, 187, 192, 194, 196, 202]
→ *Time*	46	[13, 19, 23, 25–27, 30, 46, 52, 64, 66, 68, 69, 75, 81, 85, 86, 92, 98, 100, 102, 103, 111, 112, 114, 119, 129, 143, 146, 149, 152–155, 167, 170, 172, 175, 177, 178, 182, 189, 192, 194, 200, 202]

Category	Count	Citations
Section 4: Input Data Modality		
Visual Modality and Visual Dimensions		
→ *Standard-Vision*	108	[4, 5, 8, 11–13, 19, 20, 22–28, 30, 32, 33, 37, 44–47, 49, 52–54, 56, 60, 64, 66–69, 73, 75, 80, 81, 83, 85, 86, 92, 95, 97–107, 110–116, 118, 119, 121, 122, 128, 129, 132, 137, 139, 142, 143, 146, 147, 149, 152–155, 159, 162–164, 167, 168, 170–172, 175–179, 182–185, 187, 189, 190, 192, 194, 196–198, 200–202]
→ *Scale*	25	[19, 33, 37, 44, 56, 64, 66, 67, 73, 98, 100, 101, 110, 122, 129, 132, 137, 162, 163, 179, 185, 187, 189, 196, 198]
→ *Spatial*	65	[4, 5, 8, 12, 13, 19, 20, 22, 23, 26, 28, 30, 32, 33, 37, 44–47, 49, 52–54, 73, 75, 80, 92, 95, 97, 98, 101, 102, 104, 105, 112, 116, 118, 122, 129, 132, 139, 142, 146, 147, 152, 155, 159, 163, 165, 167, 168, 170–172, 175–177, 179, 182, 189, 190, 196–198, 202]
→ *Temporal*	47	[11, 19, 22, 24, 25, 30, 47, 53, 60, 73, 75, 80, 81, 83, 85, 95, 97, 98, 100, 102, 103, 105, 106, 110, 112, 114, 118, 119, 122, 132, 142, 147, 152, 155, 162, 163, 168, 171, 175, 176, 182, 184, 194, 196, 197, 201, 202]
→ *Beyond-Human-Vision*	22	[12, 20, 24, 25, 27, 30, 33, 49, 54, 56, 66, 95, 99, 104, 110, 115, 121, 132, 137, 143, 153, 187]
Other Sensing Modalities		
→ *Audio*	36	[11, 19, 47, 60, 64, 80, 85, 92, 97, 102–106, 110, 112–114, 116, 118, 152, 154, 162–164, 171, 172, 175, 176, 182–184, 192, 194, 196, 202]
→ *Text*	24	[11, 19, 30, 44, 64, 66, 68, 69, 81, 101, 103, 107, 114, 118, 128, 129, 146, 152, 171, 172, 183, 189, 197, 200]
→ *Motion*	34	[4, 11, 19, 30, 46, 47, 49, 67, 81, 83, 85, 86, 95, 100, 103, 105, 110, 112, 113, 115, 116, 118, 122, 143, 149, 152, 163–165, 167, 178, 183, 194, 201]
→ *Haptic*	3	[4, 5, 170]
→ *Positional*	9	[8, 11, 28, 80, 110, 111, 121, 175, 178]
→ *Physiological*	8	[11, 53, 69, 80, 112, 149, 176, 196]

Category	Count	Citations
Section 5: System Design Foundations		
Data Integration Stages		
→ *Sensor-Level Integration*	27	[4, 12, 13, 28, 45, 47, 49, 52, 54, 56, 67, 73, 81, 86, 95, 100, 103, 104, 111, 115, 121, 149, 163, 165, 168, 187, 201]
→ *Feature-Level Integration*	16	[27, 32, 37, 60, 98, 99, 107, 116, 122, 132, 137, 143, 159, 170, 178, 190]
→ *Information-Level Integration*	12	[11, 26, 30, 75, 102, 119, 129, 147, 172, 176, 177, 192]
→ *Hybrid Integration*	54	[5, 8, 19, 20, 22–25, 33, 44, 46, 53, 64, 66, 68, 69, 80, 83, 85, 92, 97, 101, 105, 106, 110, 112–114, 118, 128, 139, 142, 146, 152–155, 162, 164, 167, 171, 175, 179, 182–185, 189, 194, 196–198, 200, 202]
Multimodal Data Processing		
→ *Rapid Solution Prototyping via Foundational Model APIs*	38	[5, 19, 22, 25–27, 30, 44, 53, 60, 64, 66, 69, 81, 83, 85, 92, 100, 102, 103, 113, 119, 128, 139, 146, 154, 162, 164, 167, 168, 171, 172, 175, 177, 183, 189, 196, 202]
→ *Developing A Dedicated ML Models*	91	[4, 5, 8, 11–13, 20, 22–28, 30, 32, 33, 37, 46, 49, 52, 53, 60, 64, 66–69, 73, 75, 80, 81, 83, 85, 86, 92, 95, 97–99, 101–107, 110–112, 114–116, 118, 121, 122, 132, 137, 142, 143, 147, 149, 152, 153, 155, 159, 163–165, 167, 170, 171, 175–179, 182–185, 187, 189, 190, 192, 194, 197, 198, 200–202]
→ *Heuristic Methods*	21	[23, 28, 45, 47, 49, 54, 56, 67, 104, 112, 114, 118, 129, 132, 147, 149, 159, 163, 185, 194, 196]
Evaluation Strategies		
→ *Prototyping and Demonstration*	79	[4, 5, 8, 12, 13, 19, 20, 22–24, 26–28, 30, 37, 44, 46, 47, 49, 52–54, 56, 60, 64, 66–69, 73, 75, 80, 81, 83, 85, 86, 92, 97, 100, 102–104, 110, 111, 113, 114, 119, 122, 128, 129, 132, 139, 142, 143, 146, 147, 152–154, 159, 162–165, 167, 170–172, 175–178, 183, 185, 189, 192, 196, 200, 202]
→ *Technical Evaluation*	92	[4, 5, 8, 11–13, 20, 22–25, 27, 28, 30, 32, 33, 37, 45, 47, 49, 52, 54, 56, 66, 67, 69, 73, 75, 81, 83, 85, 86, 92, 95, 97–101, 104–107, 110–116, 118, 121, 122, 128, 129, 132, 137, 139, 142, 143, 146, 147, 149, 152–155, 159, 162–165, 167, 168, 170, 171, 175–177, 179, 182–185, 187, 189, 190, 194, 197, 198, 200, 201]
→ *User Evaluation*	76	[4, 11, 19, 22–26, 30, 33, 37, 44, 46, 47, 49, 52–54, 56, 60, 64, 67, 68, 73, 75, 80, 81, 83, 85, 92, 97, 98, 100–103, 105, 107, 110–112, 115, 116, 119, 129, 143, 146, 149, 152–155, 159, 162–165, 167, 168, 170–172, 175–178, 183, 184, 187, 189, 192, 194, 196, 200–202]

Category	Count	Citations
Section 6: Application Domains		
→ *Location and Identity Recognition*	19	[8, 28, 32, 37, 45, 56, 66, 67, 95, 100, 103, 132, 137, 139, 171, 183, 185, 187, 198]
→ *Activity Detection and Understanding*	46	[4, 5, 11, 23–25, 47, 49, 52, 53, 56, 60, 67, 68, 73, 80, 81, 85, 86, 97, 104, 105, 112, 113, 115, 116, 122, 139, 143, 146, 147, 149, 154, 162–165, 167, 178, 182, 187, 192, 194, 197, 200, 201]
→ *Autonomous and Assistive Driving*	11	[11, 33, 54, 75, 83, 101, 110, 139, 176, 178, 198]
→ *Content Retrieval, Editing and Creation*	29	[19, 23, 30, 44, 64, 68, 92, 97, 101, 102, 106, 107, 114, 118, 119, 128, 129, 154, 162, 168, 171, 182–184, 189, 197, 200–202]
→ *Spatial Computing and Perception*	48	[4, 5, 8, 12, 13, 20, 22, 23, 25, 26, 28, 30, 33, 37, 45, 46, 49, 52, 53, 66, 68, 73, 86, 92, 95, 98, 99, 102, 104, 111, 116, 119, 121, 122, 142, 143, 146, 149, 155, 159, 162, 167, 170–172, 177, 190, 202]
→ *Well-being and Health Care*	28	[20, 53, 60, 69, 80, 81, 103, 105, 112–114, 116, 118, 121, 122, 132, 142, 147, 149, 164, 165, 168, 170, 179, 182, 183, 194, 196]
→ *Education*	16	[22, 30, 46, 52, 81, 119, 128, 146, 152, 162, 164, 165, 167, 197, 200, 201]
→ *Accessibility*	27	[5, 22, 23, 27, 37, 45, 47, 52, 64, 66, 73, 85, 86, 92, 99, 104, 106, 116, 122, 143, 153, 154, 163, 170, 175, 178, 185]
→ *Game*	18	[4, 25, 30, 68, 73, 97, 143, 146, 149, 152, 154, 159, 165, 167, 189, 192, 200, 202]

Category	Count	Citations
Section 7: Design Considerations and Key Challenges		
→ *Privacy and Security-aware Systems*	46	[8, 11, 19, 23, 27, 30, 37, 44, 54, 56, 60, 66, 67, 69, 73, 75, 80, 83, 92, 98, 100, 101, 106, 110, 113, 116, 128, 132, 142, 143, 147, 153, 162, 163, 165, 168, 171, 177, 183, 187, 189, 192, 194, 196, 200, 201]
→ *User Variability*	51	[4, 11, 24, 26, 44, 46, 47, 60, 64, 67, 68, 75, 80, 81, 85, 86, 92, 100, 103, 111, 113, 115, 116, 119, 121, 128, 139, 143, 146, 147, 149, 152, 153, 159, 162–165, 167, 168, 171, 175–177, 183–185, 192, 194, 196, 201]
→ *Ethics*	27	[11, 19, 27, 30, 32, 37, 52, 54, 60, 67, 75, 80, 101, 110, 113, 128, 146, 147, 163, 165, 168, 171, 183, 189, 194, 201, 202]
→ *Cognitive Load and User Engagement*	58	[5, 13, 19, 22, 25, 37, 46, 47, 49, 56, 60, 64, 66, 67, 69, 80, 83, 92, 97, 98, 101–103, 105–107, 110, 113, 115, 118, 119, 129, 147, 149, 152–154, 162–165, 167, 170–172, 175–178, 183–185, 192, 196, 197, 200–202]
→ *Automated Sensor Configuration*	32	[4, 12, 19, 28, 30, 45, 54, 56, 66, 67, 85, 95, 101–103, 110, 113, 116, 121, 143, 147, 162, 163, 165, 168, 171, 189, 194, 196, 200–202]
→ *Context Discovery*	48	[11, 12, 19, 22, 26, 30, 44, 49, 52, 54, 56, 66, 67, 75, 80, 86, 92, 95, 98, 101–103, 110, 113, 116, 119, 129, 137, 139, 147, 152, 153, 159, 162, 163, 165, 167, 171, 176–178, 182, 184, 192, 194, 196, 201, 202]
→ *Semantic Multimodal Data Integration*	54	[5, 12, 19, 20, 22, 26, 28, 30, 33, 37, 44, 45, 49, 52, 54, 56, 66, 67, 69, 75, 92, 95, 97, 99, 101, 103, 104, 106, 107, 110, 113, 115, 116, 118, 119, 121, 128, 147, 152, 153, 162–165, 170–172, 176, 182–184, 194, 196, 201]
→ *Multimodal Contextual Reasoning*	73	[5, 11, 19, 20, 22, 26, 27, 30, 32, 37, 44, 53, 54, 66, 67, 69, 75, 80, 81, 83, 86, 92, 97–99, 101–107, 110, 112–116, 118, 121, 122, 128, 129, 132, 139, 142, 147, 149, 152–155, 159, 162–165, 167, 168, 170, 171, 175, 176, 178, 182–184, 187, 192, 194, 196, 198, 201]
→ *Imbalanced Data*	27	[8, 27, 30, 32, 67, 75, 81, 83, 99, 105, 114, 116, 137, 147, 149, 153, 155, 165, 167, 172, 175, 187, 189, 194, 198, 200, 201]
→ *Assessment Heterogeneity*	11	[8, 75, 112, 114, 118, 122, 153, 155, 165, 172, 189]
→ *Scalable Architecture*	63	[4, 11–13, 20, 24, 25, 27, 28, 30, 33, 37, 44, 45, 47, 53, 54, 60, 64, 66–68, 81, 83, 85, 92, 95, 100–103, 110, 111, 113, 115, 116, 121, 132, 137, 142, 143, 146, 147, 152, 153, 162–165, 168, 171, 176, 182, 183, 185, 187, 189, 190, 194, 196, 198, 201, 202]
→ *Power Consumption*	28	[4, 5, 24, 25, 49, 66, 73, 85, 101, 110, 113, 115, 116, 143, 147, 154, 162–165, 168, 178, 179, 185, 187, 190, 194, 196]