The AI Transformation of Augmented Reality: Advances, Limits, and the Path to the Next-Generation AI Glasses Ecosystem

November 2026

Technical Report DK-ECE-2026-11-26

Hanting Ye[1], Jiasi Chen[2], Maria Gorlatova[1]

[1]Duke University, Durham, United States [2]University of Michigan, Ann Arbor, United States

ABSTRACT

Rapid advances in AI are transforming augmented reality (AR) across the stack, from how smart glasses perceive the physical world to how AR content and applications are created and deployed. At times, it feels as if science fiction has become reality: we can generate virtual content on demand and command robots with gestures and gaze through our lightweight, comfortable smart glasses. Yet the illusion quickly breaks when AI misinterprets what we can clearly see, responds to requests slowly or not at all, or the glasses run out of battery in under an hour. In this survey, we examine the emerging possibilities and practical limitations of AI-enabled AR across AR hardware, software, and applications, drawing on research in wearable computing, human–computer interaction, AI, and computer vision. First, we trace the evolution of smart glasses and examine how different hardware form factors constrain sensing, computation, interaction, and display capabilities. Second, we investigate the underlying systems, focusing on distributed intelligence, emerging sensing and interaction technologies, and the challenges of sustaining AI workloads under strict energy, latency, and thermal constraints. Third, we examine the transformation of AR software, from AI-based perception and generated spatial content to automated application development and emerging approaches to composing complete XR systems. Fourth, we present a functionality-based taxonomy of AI-enabled AR applications, explicitly identifying their required AI and system capabilities, fundamental limitations, and current deployment readiness. We conclude by presenting our vision for AR of the future: multi-device ecosystems, written and composed on demand for the specific set of applications at hand, operating reliably with unreliable AI.

Keywords: Augmented reality (AR), smart glasses, wearable AI, spatial computing, multimodal AI, egocentric perception, edge intelligence, human–AI interaction.

This work was supported in part by NSF grants CSR-2312760, CNS-2112562, and IIS-2231975, NSF CAREER Awards IIS-2046072 and NeTS-2537187, NSF NAIAD Award 2332744, a Cisco Research Award, a Meta Research Award, Defense Advanced Research Projects Agency Young Faculty Award HR0011-24-1-0001, and the Army Research Laboratory under Cooperative Agreement Number W911NF-23-2-0224. The views and conclusions contained in this document are those of the authors. They should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency, the Army Research Laboratory, or the U.S. Government. This paper has been approved for public release; distribution is unlimited. No official endorsement should be inferred. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes, notwithstanding any copyright notation herein.

I. Introduction

Augmented reality (AR) has captured the imagination for decades, as reflected in science fiction: Iron Man's J.A.R.V.I.S. provides Tony Stark with contextual information, Spider-Man's E.D.I.T.H. lets Peter Parker control remote devices through glasses, and the Z-Eyes in Black Mirror: White Christmas allow users to selectively obscure people in their field of view. Researchers have long been pursuing and refining this vision, from pioneering prototypes such as Sutherland's Sword of Damocles (1968), Mann's WearComp (1981), and Feiner et al.'s KARMA (1993) to commercial devices such as Google Glass, Microsoft HoloLens, Magic Leap, and Apple Vision Pro.

Currently, the field of AR is undergoing a radical transformation. Once seen primarily as a means of displaying information, AR is increasingly seen as a gateway to AI. As James Choi, the Executive VP and Head of XR R&D at Samsung, put it in July 2026, *"Intelligent eyewear introduces a more natural and intuitive way to experience mobile AI*" [204]. Different AI-based applications are driving the development of diverse form factors – a "Cambrian explosion" of eyewear with different sensing, computing, and display capabilities. For example, glasses designed to assist blind or low-vision users may rely on audio capabilities and lack a visual display. AR devices increasingly rely on companion wearables for sensing and interaction and on distributed architectures spanning on-device, edge, and cloud components for data processing and computing. Meanwhile, foundation models are lifting key constraints on perception and 3D content, reducing the need for task-specific ML models and manual content authoring. Advances in AI and mobile computing are bringing long-envisioned, sci-fi-esque applications into the hands of consumers: in 2026, asking questions about one's physical surroundings and receiving contextual answers is a default feature of many commercial devices. Less anticipated possibilities, including turning everyday human activities into training data for robots and generating AR tutorials from human demonstrations, are now on the cusp of realization. Fig. 1 shows three broad categories of AI-enabled AR applications and the AI transformations taking place within each. The process of creating AR applications is changing as well: the Tony Stark of tomorrow may rely on J.A.R.V.I.S. not only to carry out tasks, but also to create specialized AI assistants that extend its capabilities.

Yet in AI-integrated augmented reality, the gap between expectations and reality is enormous. During an on-stage demonstration of AI-assisted cooking at Meta Connect in September 2025, the Ray-Ban Meta glasses repeatedly provided instructions inconsistent with the chef's actual progress [186]. In day-to-day interactions, user experience is similar. Advanced, sci-fi-like capabilities rely on centralized server-based AI, resulting in response times that depend on cellular network conditions and limiting their usefulness in areas with less-than-stellar cellular coverage. High error rates of AI models are widely reported in both numerous first-person accounts and formal studies: e.g., a recent study of Gen 2 Ray-Ban Meta AI eyeglasses reports 50% accuracy on object counting and 2% accuracy on coin identification [160]. Moreover, commercial smart glasses typically last less than an hour with continuous visual sensing, with RayNeo X3 Pro supporting only 36 minutes of continuous video recording [200], and the battery life of Ray-Ban Meta glasses dropping to just 30 minutes when using Live AI [220]. Next-generation capabilities fare no better: e.g., auto-generation of 3D content is shown to succeed in only 1 of 3 attempts in one recent study [282] and in only 17% of attempts in another [134]. Even basic device control remains cumbersome, frustrating, and restrictive, with users struggling to communicate intent and verify the outcomes of their requests, and refraining from interacting with their devices in public settings [110], [143]. While AI models are likely to continue improving, these usability challenges span the entire stack and are unlikely to be resolved through advances in AI alone.

In this survey, we aim to separate exciting futuristic visions for AI-aided AR from the realities of what current technology can deliver. We ask: *How is AI transforming the capabilities, architectures, and development process of AR systems? How close are AI-enabled AR applications to practical realization? What advances in AR systems and software are needed to realize the next generation of AI-enabled AR applications?* To address these questions, we look across the stack, from individual device components to hardware platforms to software and applications, and bring together research from multiple fields, including embedded sensing and mobile and wearable computing, AR and human-computer interaction, user interface design, and AI, machine learning, and computer vision.

Organization of this article: First, in Section II, we trace the evolution of smart glasses and examine their modern form factors. Then, in Section III, we review trends and open problems at the hardware layer: distributed intelligence, sensing, interaction, and display technologies. Next, in Section IV, we examine AI software for XR, including automated coding, content generation, and foundation models for AR, as well as evaluation methods.

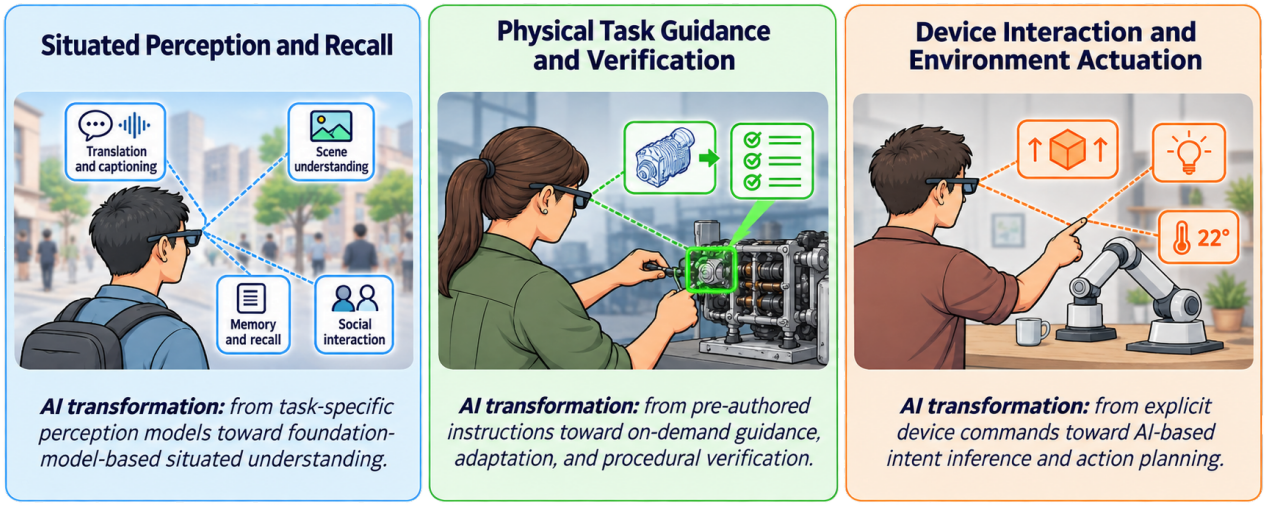

Fig. 1. AI transformations across three functional categories of AR applications.

In Section V, we present a functional taxonomy of AI-enabled AR applications and examine their requirements, limitations, and deployment readiness. Finally, in Section VI, we outline challenges and future research directions. We conclude our survey in Section VII.

Comparison to other survey papers: The intersection of AI and AR is widely recognized as a key technical and economic opportunity of our era. Currently, shipments of glasses combining AI and AR are rapidly growing [175]; Citi Research projects the AI smart glasses market to reach $40 billion by 2030 [33]. Unsurprisingly, given the extent of the opportunity, multiple surveys have examined the intersection of AI and AR from different angles. In early 2023, Hirzle et al. [72] provided an important snapshot of the field prior to the widespread adoption of generative AI. Subsequent surveys summarized applications of machine learning and deep learning in XR [4] and examined the interaction of XR and LLMs and generative AI [231], [232], but did not examine system constraints. Recently, [272], [274] provide a complementary perspective, specifically focusing on "AI smart glasses" as a category of devices and systematically examining their hardware and AI capabilities, interaction paradigms, and application scenarios across different domains. Our survey complements existing work by examining the broader AI transformation of AR beyond applications of egocentric computer vision. In contrast to [272], [274], we present a *functionality-based*, rather than domain-centered, taxonomy of AI-enabled AR applications, connecting user-facing application capabilities to computer system and AI requirements. We identify comprehensive AI-enhanced development, AI-enhanced orchestration, and multi-device system integration as the next frontiers for AI-enabled AR.

II. AI Glasses: Precursors and Modern Form Factors

Transforming glasses, one of our most familiar personal wearables, into smart devices has long been envisioned, from early head-mounted eyewear prototypes to the introduction of Google Glass nearly 15 years ago and the recent rapid emergence of various types of smart glasses (see Table I). This section introduces the evolution of smart glasses and their current categories, the system capabilities they offer, and the evolving requirements of hardware in the AI era.

A. Smart Glasses Precursors

We examine the foundational concepts in *perception*, *display*, *computing*, and *interaction* that have shaped the evolution of smart glasses. Fig. 2 provides an overview of early wearable glasses prototypes and near-future smart glasses.

TABLE I

A HARDWARE-ORIENTED TAXONOMY OF SMART GLASSES. ✓ INDICATES A COMPONENT OR SPATIAL-TRACKING CAPABILITY THAT IS TYPICALLY SUPPORTED, ○ INDICATES THAT ITS SUPPORT VARIES ACROSS PRODUCTS IN THE SUBBRANCH, AND – INDICATES THAT IT IS TYPICALLY UNSUPPORTED.

Category	Branch	Representative Smart Glasses	Cam.	Mic.	IMU	Eye	Spatial		Disp.	Spk.	Compute chip
							3 DoF	6 DoF			
Google Glass	Precursor	Glass Enterprise Edition 2	✓	✓	✓	–	–	–	✓	✓	**Snapdragon XR1**
All-day wearable	Audio-only	Bose Frames; Ray-Ban Meta Audio; Dymesty Cook Edge	–	✓	○	–	–	–	–	✓	**Qualcomm QCC**: Bose Frames, Dymesty Cook Edge
	Camera-enabled	Ray-Ban Meta; Rokid AI Glasses Style; Xiaomi AI Glasses; Samsung Intelligent Eyewear; Solos AirGo V2	✓	✓	✓	–	–	–	–	✓	**Snapdragon AR1 Gen 1**: Ray-Ban Meta/Rokid/ Xiaomi/Samsung; *2nd low-power chip*: **BES2800**: Xiaomi; **NXP RT600**: Rokid
	AR display	Even Realities G1/G2; Meta Ray-Ban Display; Quark AI Glasses S1; Rokid Glasses; INMO GO3; MemoMind One; RayNeo X3 Pro	○	✓	✓	–	○	○	✓	○	**Snapdragon AR1 Gen 1**: Meta Display/Quark/ Rokid/RayNeo; **Ambiq Apollo510**: Even Realities; *2nd low-power chip*: **BES2800**: Quark; **NXP RT600**: Rokid
	Spatial computing	Meta Orion; Snap Spectacles (2024); Snap SPECS (2026)	✓	✓	✓	○	✓	✓	✓	✓	**Snapdragon SoC**: Snap Spectacles (2024); Snap SPECS (2026)
Near-eye display	Birdbath display	XREAL AURA, Air, and One series; VITURE Pro and Luma	○	○	✓	–	✓	○	✓	✓	**Snapdragon Reality Elite**: XREAL AURA; **XREAL X1**: XREAL One
	Micro-OLED display	Meta VR Glasses, URXR One	○	○	✓	○	✓	○	✓	○	**Snapdragon Reality Elite**: Meta VR Glasses; **GravityXR**: URXR One

Perception: First-person personal imaging. The egocentric view of the people, objects, and activities that constitute the user's immediate experience empowers smart glasses to serve as a perceptual interface to the physical world. Steve Mann introduced this concept in 1997 by using a wearable camera to enable continuous visual capture and processing [156]. Meanwhile, egocentric perception can also support *context-aware augmented reality* by providing information about the user's surroundings and activities [62], thereby better augmenting the physical world by overlaying adaptive digital information based on changing requirements and constraints. In this context, egocentric captures could not only be used for lifelogging and retrospective memory [73], but also be used to learn and recognize handled objects and to jointly infer daily activities, actions, hands, and object interactions [46].

Display: Unobtrusive near-eye viewing. Unobtrusive near-eye viewing keeps digital information in AR continuously accessible while preserving situational awareness of the physical environment. The idea that "the wearer could effectively forget its presence" was proposed for head-mounted displays in 2000 [103]. Two display approaches have evolved to achieve such viewing [212]: *video-see-through* and *optical-see-through*. *Video-see-through* uses the aforementioned egocentric cameras to continuously capture the physical world and reproject it onto high-resolution displays, allowing users to view the physical world even when their view is fully enclosed by the headset, as demonstrated by prototypes in 1997 [156] and recently by Apple Vision Pro, Samsung Galaxy XR, and Meta Quest. However, this approach also hinders eye contact, which is important for natural social interactions [195]. *Optical-see-through*, in contrast, uses optical components to create a semi-transparent display that preserves a direct view of the physical world while overlaying virtual content, as explored in Boeing's 1992 optical-see-through display for aircraft manufacturing [21], and has now become the mainstream display approach in smart glasses [3].

Computing: Eyeglass form factor and distributed computing. The miniaturization of smart glasses toward an eyeglass-scale form factor that closely resembles ordinary eyewear is another important line of development.

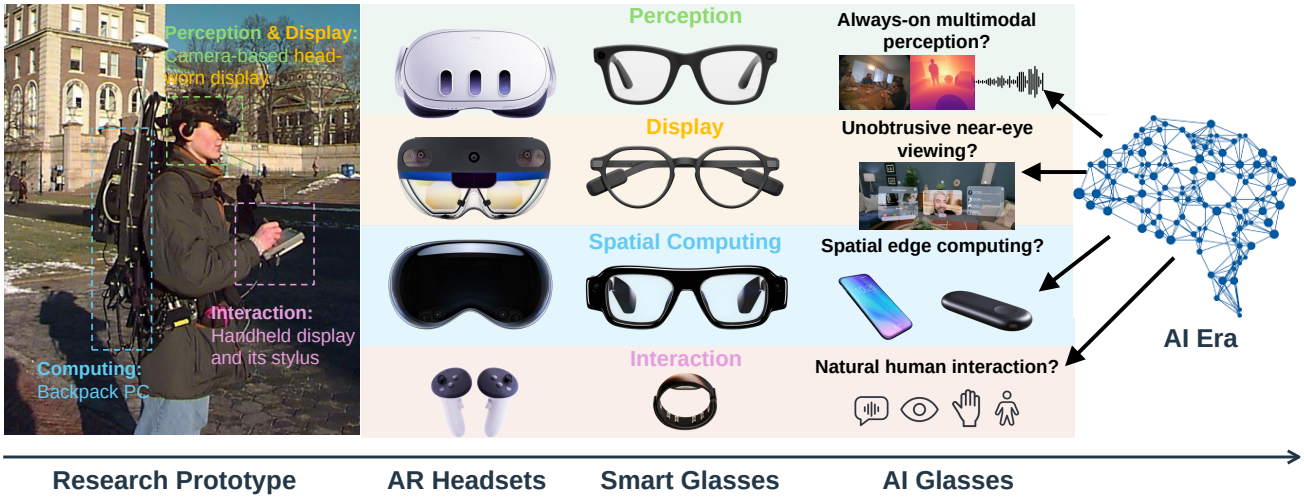

Fig. 2. Representative evolution of smart-glasses hardware from 1997 to 2026.

To realize this, a distributed computing architecture was demonstrated by the Touring Machine in 1997 [47], in which cameras and displays were incorporated into an eyeglass form factor to meet the wearable-system design requirements of ordinary eyeglasses in terms of weight, thermal management, and aesthetics [56], while a backpack was used for computation, communication, storage, and power supply, and a touchpad for interaction. Meanwhile, as smartphones became more compact, various electronic components, e.g., IMUs and high-resolution displays, also became smaller, contributing to increasingly lightweight AR headset designs. This concept of distributed computing and the miniaturization of key electronic components influenced the development of subsequent eyeglass-form-factor prototypes in 2009 [48], Google Glass in 2012, and modern smart glasses.

Interaction: Natural interaction modalities. The development of interaction techniques has also gradually shifted toward more natural forms. In 1996, in prototypes such as Studierstube, interaction still relied on a handheld pen-and-panel interface [229]. These interfaces later evolved into lighter handheld controllers, which have now been widely adopted in commercial AR headsets, such as Meta Quest and Samsung Galaxy XR. Meanwhile, interaction has evolved beyond these handheld devices toward more natural and direct interaction methods. Hand tracking was introduced as early as 2014 in headsets such as the Oculus Rift and Meta 1, enabling interactions with digital content, such as clicking, grasping, and moving virtual objects. As we enter the AI era, natural-language voice interaction has also become one of the most natural ways for users to interact with AI models through smart glasses. Beyond these interaction modalities, the integration of eye trackers into smart glasses has also attracted considerable attention, as it offers another natural form of interaction while providing additional cues about users' attention.

B. Google Glass

Google Glass, introduced in 2012, became an important turning point. It was the first commercial eyeglass-scale device, rather than a bulky prototype, to integrate the aforementioned concepts.

Hardware and software ecosystem. Google Glass integrated a single first-person camera and a monocular near-eye display based on a liquid crystal on silicon prism design. The Explorer Edition also featured an OMAP4430 processor with image signal processing capabilities and supported computation offloading to edge servers via Wi-Fi or Bluetooth [137]. Meanwhile, Google Glass also integrated a microphone, bone-conduction audio, and a temple-mounted touch panel to support audio and touch-based interactions. On the software side, its Android-based platform supported various applications. The *Mirror API* allowed web services to deliver notifications and information cards to the glasses' display, while the *Glass Development Kit* enabled native applications with hardware access [60]. Together, these capabilities made Google Glass a relatively comprehensive system. These capabilities also led to its selection as one of TIME's Inventions of the Year upon its introduction in 2012, with the device described as

"*the device that will make augmented reality part of our daily lives*," despite its high price of $1,500 at the time (approximately $2,200 in 2026 dollars).

Barriers to consumer adoption. Although Google Glass successfully integrated various functionalities into a comprehensive system, it was eventually withdrawn from the consumer market in 2015 and shifted toward the enterprise market until its complete discontinuation in 2023. In 2014, many developers had already abandoned their projects due to a lack of users or limitations of the device itself [176]. The reasons were multifaceted. The *camera*, a core component of Google Glass, raised bystander privacy concerns, as bystanders could not reliably determine whether Google Glass was recording them [38]. This triggered strong public reactions at the time, to the extent that Google had to issue an etiquette guide warning wearers against being "creepy or rude" or a "Glasshole" [6]. In terms of the *glasses form factor*, its lightweight design made it difficult to meet the substantial energy demands of camera use, video communication, wireless networking, and on-device computation, limiting Google Glass's battery life to only 45 minutes [137], which was far from sufficient for everyday use. At the same time, this lightweight design also posed an overheating risk near the wearer's temples, with Google Glass reaching 52 °C during use [137]. Yet this lightweight form factor still failed to gain user acceptance. Users complained about ergonomic design issues, including difficulty viewing AR content and interacting with the device [14], [270], and multiple reports described awkwardness during face-to-face interactions involving wearers due to Google Glass's conspicuous appearance [42], [109]. Together, these experiences show that, despite the substantial progress achieved by Google Glass, the widespread adoption of smart glasses still requires attention to appearance, privacy, and public acceptance, as well as comprehensive improvements in system efficiency, ergonomics, and interaction for wearable devices intended to be worn close to users throughout the day.

C. Modern Smart Glasses Development Branch: All-Day Wearable Smart Glasses

Guided by the lessons from Google Glass, the development of smart glasses has shifted toward designs that keep them lightweight and suitable for everyday wear, and manufacturers have increasingly adopted a "subtractive" design approach by reconsidering which components are truly necessary for users' intended applications. Based on this design philosophy, we further classify smart glasses into four categories: *Audio-only*, *Camera-enabled*, *AR display-enabled*, and *Spatial* smart glasses. Fig. 3 illustrates the differences in component integration across these categories of smart glasses.

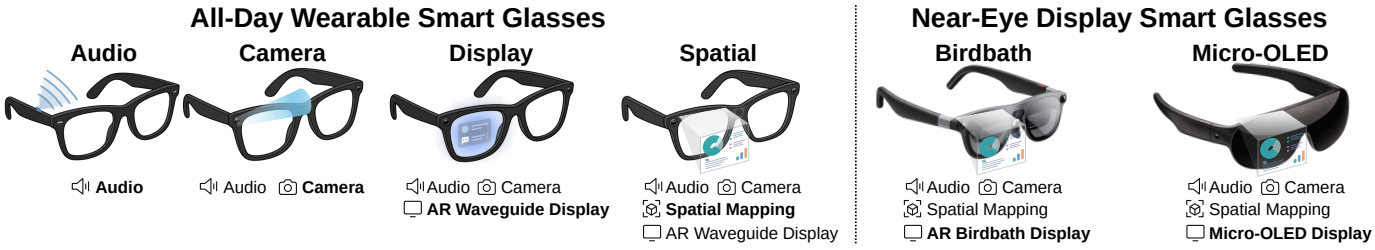

Fig. 3. Different hardware compositions of modern smart glasses.

Audio-only smart glasses represent the simplest form of smart glasses, discarding visual elements and relying solely on audio input from microphones and audio output from speakers. As a result, they can maintain a lightweight form factor resembling ordinary eyeglasses, making them suitable for all-day wear. Early audio-only glasses primarily served as accessories to other mobile devices to provide conventional audio functions, such as music playback and phone calls. With the rapid advancement of AI, audio-only smart glasses have also increasingly served as an all-day wearable interaction layer between users and AI assistants, with examples including Amazon Echo Frames, Solos AirGo, and the recently introduced Samsung glasses and Ray-Ban Meta Audio, which provide access to AI assistants and language-model-based services.

Camera-enabled smart glasses extend audio-based interaction by incorporating an important visual input channel envisioned for smart glasses: *first-person imaging*, also referred to as *egocentric vision* [156]. A representative example of this category is the Ray-Ban Meta series of smart glasses, which has seen substantial market growth: EssilorLuxottica reported that its AI-glasses portfolio developed with Meta sold more than seven million units in

2025 [44]. A comparative study of six head-mounted wearable devices found that Ray-Ban Meta smart glasses received the highest social-acceptability ratings [92], mainly because of their compact size and resemblance to traditional eyewear [15]. Camera-enabled smart glasses have been developed to provide AI models with an important visual input from the wearer's egocentric perspective, helping them better understand the physical world.

AR waveguide display-enabled smart glasses. Advances in AR waveguide display technology have enabled smart glasses to provide a visual information interface. The waveguide leverages total internal reflection to guide light from a light engine located in the glasses' temples through the lens and direct it toward the user's eye. Such an interface, overlaid on the real world, provides AI systems with a direct user-facing channel for conveying guidance and decision-support information. The number of display-enabled smart-glasses models has grown rapidly in recent years, with examples including Meta Ray-Ban Display with a monocular full-color display, Even Realities glasses and Quark AI Glasses S1 with binocular monochrome displays, and RayNeo X3 Pro with a binocular full-color display.

Spatial smart glasses. Advanced MR headsets, such as Apple Vision Pro, have already demonstrated the promise of immersive spatial computing: *spatial anchoring in the physical environment*, i.e., understanding the surrounding environment, and *tracking the wearer's pose*, i.e., determining their position and changes in pose. Can such powerful spatial computing capabilities eventually be brought to smart glasses? To date, Snap SPECS and Meta Orion prototypes have offered glimpses of this future. Both face a common challenge: the full spatial-tracking stack that previously resided in bulky video-see-through MR headsets has to fit within a limited form factor and operate under a tight power budget. To address this, SPECS use a standalone dual-processor architecture, assigning computer vision and AR applications to separate Snapdragon processors [215]. Nevertheless, SPECS still have a relatively bulky eyeglass form factor and weigh 132 g, compared with many camera-enabled and display-enabled smart glasses weighing around 50 g. This weight may make them challenging to wear throughout the day. Meta Orion adopts a distributed computing approach, using a wireless compute puck to handle application logic, rendering, and other daily user workloads, and retains hand tracking, eye tracking, simultaneous localization and mapping (SLAM), and other spatial computing components on the glasses.

D. Modern Smart Glasses Development Branch: Near-Eye Display Smart Glasses

These near-eye display glasses can be divided into two categories: *birdbath display glasses* and *Micro-OLED display glasses*. In the first category, "birdbath" mainly refers to a distinctive optical design in which the display content is reflected toward the user's eyes through a birdbath-shaped optical structure above the glasses frame, representing an indirect display approach. A representative example of this category is the recently introduced XREAL AURA, which was developed with Google for Android XR. A drawback of this display approach is that its transmission of real-world light is generally lower than that of waveguide-based AR displays. However, its advantages in optical display performance are also significant. XREAL AURA provides a 70° FoV, substantially wider than the approximately 30° FoV of many waveguide-based displays.

Micro-OLED display glasses take a different approach, prioritizing bringing the immersive visual experiences of headsets such as Apple Vision Pro into the constrained form factor of eyewear. To this end, Meta VR Glasses were recently announced with Qualcomm's Snapdragon Reality Elite chipset to support 5K Micro-OLED displays [162], with high-resolution displays placed close to the user's eyes. Both types of near-eye-display smart glasses can provide portable workspaces for knowledge work or leisure activities, such as document editing, web browsing, email, and video viewing [31]. However, a limitation is that they still function as extended monitors for work and entertainment, lacking novel XR interface designs that help users adopt new spatial workflows while preserving familiar interactions [31]. Micro-OLED displays can also serve as light sources in birdbath optical systems, e.g., in the XREAL AURA.

III. Hardware Architectures, Limitations, and Frontiers

A. Computing: Distributed Intelligence Across Glasses, Companion Edge, and Cloud

The demand for intelligence has led most current smart glasses to invoke cloud-based AI services to perform various complex tasks [256], [17], [83], [143], [262], while this cloud computing architecture still faces challenges



(a) Glasses-Cloud Architecture (b) Glasses-Companion Edge-Cloud Architecture

Fig. 4. Emerging distributed computing architectures for smart glasses with LLMs of varying parameter sizes.

related to data privacy and the need for reliable, real-time connectivity. This has led to intelligence being increasingly distributed across the smart glasses themselves and their companion mobile edge devices, as shown in Fig. 4.

On-glasses computing. The mainstream chip architecture currently adopted by smart glasses is the System-on-Chip (SoC), which integrates different functional modules onto a single chip as shown in Fig. 5. One of the most widely used platforms is Qualcomm's Snapdragon AR1 Gen 1, which integrates an image signal processor (ISP) for camera image processing and a digital signal processor (DSP) for audio signal processing. For display output, it supports 2D and 3D graphics rendering pipelines with a display resolution of up to 1280 × 1280 pixels at 60 frames per second (FPS)[1]. For communication, it supports Wi-Fi and Bluetooth connectivity. Most smart glasses, such as Snap Spectacles (2024), INMO Air3, and RayNeo X3 Pro, access cloud-based AI services through Wi-Fi connections. Though direct 5G cellular connectivity could allow smart glasses to connect almost anywhere, currently only a few smart glasses support it, e.g., the recently launched Snap SPECS. Lastly, its computing components include a CPU for general-purpose processing and a dedicated neural processing unit (NPU), i.e., Qualcomm Hexagon, for accelerating AI models. For example, with NPU acceleration, Llama 3.2 1B-Instruct could run directly on glasses at approximately 6 tokens per second, with a time to first token of 185 ms [226]. Meanwhile, Snapdragon AR2 Gen 1 further enhances perception capabilities for next-generation spatial smart glasses. It supports up to nine concurrent cameras to enable a spatial-computing platform, while also supporting on-glasses eye tracking and iris authentication. However, the high performance of these SoC platforms comes at a cost. Despite the integration of low-power designs, their overall power consumption remains significant for smart glasses. Therefore, a multi-chip architecture that adds an additional low-power microcontroller (MCU) alongside the main SoC has become another option, as shown in Fig. 5. In this architecture, the auxiliary MCU, e.g., the NXP RT600 in Rokid Glasses, is dedicated to low-power always-on sensing and wake-word detection for routine monitoring. Once a relevant event or task is detected, it wakes up the AR1 Gen 1 to perform more power-intensive tasks, such as ISP processing for photo capture and video recording, thereby avoiding unnecessary activity.

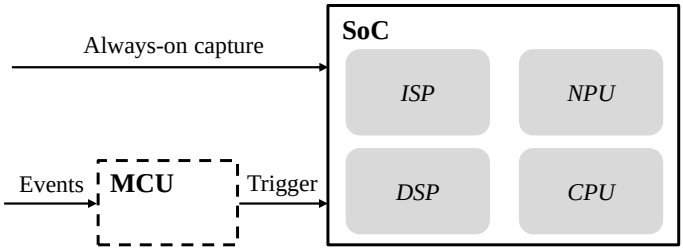

Fig. 5. On-glasses computing chip architectures and components.

Open problems: A current bottleneck in deploying AI models on-device is limited *on-chip memory*. YOLOv11n, requiring only 632 kB of internal activation memory, was selected for deployment on the STM32N6 with 4 MB of memory [57]. Similarly, TinyissimoYOLO [168] uses techniques such as quantization and tiling to reduce the memory requirements for deploying object detectors on GAP9-based glasses. Commercial platforms, such as Spectacles, also encourage the use of compact models smaller than 10 MB, such as YOLOv7-tiny and MobileNet [216]. Due to these memory limitations, current smart glasses can only deploy lightweight object detectors or classifiers and language models with up to 1B parameters. Emerging neuromorphic computing [116], with its low-memory and low-power advantages, will open up more possibilities for on-glasses computing.

Companion edge computing. Edge computing refers to offloading computation to nearby computing devices. Early edge-computing deployments for smart-glasses-based wearable cognitive assistance primarily relied on desktop

[1]Since most current smart glasses primarily display 2D content, such as text and images, 3D rendering capabilities are less frequently used.

servers [66] and showed that using edge computing resources could significantly reduce the overall response time across all smart glasses applications compared to cloud computing [30]. More recently, with the increasing computing capabilities of mobile chips, workload offloading has been moving toward portable companion devices, such as *compute pucks* and *smartphones*, giving rise to a new paradigm of *companion edge computing*. Following this trend, Google's Android XR provides APIs that enable companion smartphone-hosted applications to deliver audio or visual experiences on smart glasses [59].

In *puck-based edge computing*, dedicated portable computing pucks have been developed for more advanced smart glasses with higher spatial computing demands, such as the Meta Orion, and near-eye display glasses, such as the XREAL AURA, to support the spatial computational requirements of the glasses through wireless connections or wired connections that also provide power, respectively. Meanwhile, communication modules can also be integrated into these computing pucks, enabling data connections with smartphones via Bluetooth or Wi-Fi, as well as direct cloud connectivity over Wi-Fi or 5G cellular networks, as exemplified by XREAL Beam Pro 5G. This dedicated portable computing offloading architecture allows smart glasses to meet the wearability requirements in terms of weight and heat distribution while providing the substantial computing resources required for high-accuracy tracking and spatial localization applications, e.g., 6DoF spatial anchoring and hand tracking provided by the AURA. **Open problems:** Orion's report shows that wireless camera-data transmission can consume more energy than local processing for some workloads, and can introduce more severe thermal issues [269]. Therefore, sustained operation requires finding the best trade-off between distributed communication costs and on-glasses computational workloads, while some studies have also explored optimizing the spatial distribution of heat across the glasses' temples, especially in regions near the skin, through component placement and material choices [159].

In *smartphone-based edge computing*, increasingly powerful smartphones, which are ubiquitous among users, have provided the computing capabilities to deploy AI models, such as object detection models [213] and multimodal large language models [146], offering benefits including the potential to mitigate privacy concerns associated with continuous sensing, cost advantages, personalization, and offline availability. **Open problems:** Many studies have investigated this direction to enable edge intelligence on smartphones [146], but three key limitations have also emerged: smartphones can currently deploy only compact models (¡4B parameters, compared with 1B-parameter models deployed on glasses), model performance degrades substantially due to scaling laws, and even with mobile NPU acceleration, inference latency still reaches 3 s for visual question answering (VQA) and 10 s for video-streaming inference (whereas calling lightweight cloud models can also achieve latency below 10 s). As a result, smart-glasses applications relying solely on smartphone computing remain limited to tasks that do not require advanced intelligence or real-time responsiveness [238].

Given the current limitations of mobile devices in computing resources, a new *glass-phone-cloud* computing architecture is emerging. This architecture uses smartphones as users' trusted edge devices and as relays to the cloud, allowing data to be compressed or anonymized, e.g., users' privacy-sensitive egocentric data, before being uploaded over either 5G cellular networks or Wi-Fi [105]. Recent commercial smart glasses, including Ray-Ban Meta and Even Realities, have also adopted this architecture. **Open problems:** A measurement study [104] of this architecture has revealed three key issues and opportunities in communication and computing introduced by such multi-hop pipelines: (1) latency challenges, where 1 s of end-to-end latency cannot meet the 100 ms responsiveness commonly associated with interactive communication; (2) asymmetry in uplink and downlink data processing, where data sent from smartphones to smart glasses undergoes more encoding and image preprocessing steps than uplink data to accommodate the hardware display constraints of different smart glasses; and (3) opportunities for system-wide optimization, where the lack of unified interfaces and standards across vendors in the overall pipeline leads to redundant transcoding operations when smartphones relay video to the cloud, resulting in 2–3 times the latency compared with phone-native live streaming.

B. Perception: Expanded Viewpoints and Multimodal Sensing

The egocentric perception provided by smart glasses has become key to AI's perception of the physical world and to real-world tasks, e.g., progress verification [83], but transient actions and incomplete camera views pose barriers to these tasks, which require preserving both the context of the wearer's activity and the visual details needed to assist with it. To this end, Meta has also introduced the Project Aria research glasses platform to explore how

AI systems can better understand the world from a human perspective, and the latest Aria Gen 2 uses a camera array with a larger field of view (FoV) and more wide-angle cameras (including a 12-megapixel RGB camera with a 133° horizontal FoV and four wide-angle computer-vision cameras) to enhance the glasses' perception capabilities. Existing approaches are also exploring supplementary devices to complement egocentric perception, such as using handheld [224] or finger-worn [136] cameras to provide more *flexible viewpoints*, while environmental viewpoints, such as IoT cameras, can provide third-person views to complement egocentric vision in *global spatial awareness* [206]. In addition, *mutual egocentric views* from multiple wearers in shared spaces can complement one another by providing tracking information about individual wearers, such as 3D body motions [94], as well as shared spatial information, such as long-term spatial memory [108].

Multimodal augmented egocentric vision. To overcome the limitations of traditional RGB camera-based perception, various multimodal sensing approaches have been proposed, including depth, acoustic, thermal, and millimeter-wave sensing. Depth sensing can provide AI models with distance cues about the physical world and also provide blind users with accurate distance information for obstacle avoidance [53], [268]. To overcome the effects of changes in environmental lighting conditions, acoustic sensing approaches have also been proposed, mainly using microphone arrays to analyze the reflections of inaudible acoustic signals to support various sensing tasks, such as body pose tracking [130], facial-expression recognition [132], and dietary activity recognition [152]. In addition, thermal sensing is also unaffected by adverse lighting conditions and offers important cues at night and highlights key biological information [2]. To address the occlusion limitations of cameras, non-line-of-sight sensing using wireless signals has also been proposed [12], [149] to detect targets behind occlusions.

Open problems: As the sensing capabilities of smart glasses continue to be enhanced to meet AI demands, incorporating additional wide-angle camera arrays or new multimodal sensing components will pose great challenges to the already energy-constrained glasses while maintaining their effectiveness in real-world scenarios. *Perception energy consumption:* Active image capture was previously measured to account for a major portion of Google Glass's energy consumption [137]. Although the energy consumption of this sensing component has been optimized in recent years, experiments show that active perception is still a major contributor to the total energy consumption of smart glasses [50], [57]. For example, in GAPses' continuous eye-movement classification configuration, sensing electronics account for 50.3% of the 12.3 mW system power, while classification computation contributes only 0.97% [50]. How to balance the design of the overall sensing system is therefore a key consideration for the future. *Perception robustness:* Although numerous laboratory prototypes of multimodal sensing have been proposed, these sensing approaches still face challenges in accurately tracking diverse body poses across large user populations and in diverse environments in practical deployments [130], [149]. Addressing these challenges calls for future efforts to increase the scale of user data collection to improve perception performance and optimize the overall perception system design.

C. Display: AR Waveguide Architectures and Optical Challenges

Current waveguide-based display glasses mainly adopt two optical display approaches: *diffractive waveguides* and *geometric waveguides*. Diffractive waveguides mainly use diffractive gratings to control light paths and can employ different microscopic structures, including *surface-relief gratings*, *volume holographic gratings*, and *metasurfaces*. Geometric waveguides, in contrast, mainly use a sequence of reflective mirrors or refractive prisms to control light propagation. Diffractive waveguides offer a brightness efficiency of 400–1000 nits/lm and a field of view (FoV) of only 30°, whereas geometric waveguides offer higher brightness efficiency of 650–1000 nits/lm and a FoV of 50° [39]. Due to constraints such as manufacturing costs, most commercial display-enabled smart glasses, such as Even Realities G1 and G2, Rokid Glasses, INMO GO3, and RayNeo X3 Pro, currently adopt lower-cost waveguide solutions based on surface-relief gratings. More expensive geometric waveguides are used in a few display glasses such as Meta Ray-Ban Display and INMO Air3.

Open problems: Current waveguide displays still face challenges related to *optical display performance, optical leakage, and optical efficiency*. In terms of *optical display performance*, some current waveguide-based displays have relatively low resolutions, i.e., 480p, and limited FoVs, i.e., 30°, which constrain the angular region in which content is visible to the wearer's eyes [39]. Software can rearrange content within the optical viewport, but cannot extend that viewport across the wearer's entire visual field due to the hardware constraints of waveguide displays [3].

In terms of *optical leakage*, not all light emitted from the light engine in the glasses' temples reaches the wearer's eyes [39], and some of the light leaks outward and can be observed from other viewing angles [264]. Such optical leakage from AR displays also raises concerns about social acceptability and privacy [112]. Lastly, the *low optical efficiency* of waveguide displays makes display power consumption a system-level bottleneck for smart glasses. In a representative 30° diffractive-waveguide scenario, a waveguide with an efficiency of approximately 1300 nits/lm requires around 3 lm at the in-coupler to deliver about 4000 nits at the eye, resulting in additional power consumption of 200–800 mW [196]. At the hardware level, additional dimmers can reduce the brightness of the real-world scene and thus lower the required display brightness, but reduced transparency to the real world when no content is displayed remains a concern [3]. At the software level, existing approaches mainly adjust pixel luminance according to the display-power budget to reduce power consumption without compromising users' perceived visual quality [139].

D. Interaction: On-Glasses and Wearable-Assisted Input

Interaction methods for smart glasses currently rely either solely on sensors integrated into the glasses, such as microphones for voice input in Ray-Ban Meta glasses and touchpads in RayNeo X3 Pro, or on *additional wearable devices*, such as smart rings [17] for Star V Air 2 and Even Realities G2 and wristbands [100], [249] for Meta Ray-Ban Display glasses. In addition, other promising wearable devices include smartwatches [107], wearable straps [181], and gloves [77].

Open problems: Despite the unprecedented progress of the last few years, several interaction requirements from users remain unmet. *Silent input.* Voice control is susceptible to noisy environments and can be awkward to use in public, and silent input has been envisioned to enable private human–AI interaction [110]. Therefore, several speech-related cues, including nasal-pad vibrations [158], facial movements [277], head movements [222], and jaw movements [223], are leveraged to infer speech content. However, these non-audible cues are mainly captured using on-device IMUs, and interference caused by motion noise in the surrounding environment remains a challenge to be addressed. *Fine-grained control.* Current touchpads mounted on the glasses' temples lack fine-grained interaction capabilities, such as keyboard input. As an additional wearable device, the Meta Neural Band has already enabled neural handwriting on hard surfaces in the environment when paired with Meta Ray-Ban Display glasses through the recognition of surface electromyography signals. However, current neural handwriting capabilities still face challenges related to low recognition accuracy. Recent work [249] has also highlighted that such surface-based handwriting still exhibits inconsistent recognition accuracy across a broad user population in new environments.

E. Hardware Frontiers

AI needs smart glasses to serve as a multimodal sensing interface that enables the sensing and interpretation of what the user sees, hears, and potentially even smells or feels, paving the way for a new generation of intelligent, always-on agents, each serving as an assistive embodiment of AI, a digital co-pilot that continuously senses both the user and their environment, adapts in real time, and can even guide behavior. Advancing toward this vision presents a complex roadmap necessitating tight co-design across hardware, algorithms, and power management. We review emerging sensors for smart glasses and discuss the open problems associated with each.

Event-based cameras. Frame-based cameras are currently among the most energy-consuming components in smart glasses, reaching 1.2 W during video recording, while low-power event-based cameras consuming only 20 mW are emerging as promising alternatives. Unlike frame-based cameras, which capture redundant full-scene images at fixed intervals, event cameras respond asynchronously to luminance changes, emitting sparse, microsecond-resolution data only when the scene changes [22]. This yields dramatic power savings during the static conditions that dominate daily wear, while preserving high temporal resolution when motion occurs. The native sparsity of the event stream also aligns naturally with efficient sensing of scene and object changes during everyday interactions. These properties have also recently led to the proposal of a new class of event-based visual perception approaches for smart glasses [8], [167]. **Open problems:** One key limitation of event-based cameras is their low spatial resolution, typically only QVGA (240 × 320), whereas the RGB camera resolution of Meta Ray-Ban Display glasses can reach 1080p (1920 × 1080) or higher. Inspired by previous triggered-perception architectures, e.g., approaches that activate RGB cameras using grayscale cameras [18] or audio-based interaction cues [182], recent studies have

also explored hybrid event- and frame-based camera architectures [11], [275]. These architectures leverage the low-power characteristics of event cameras for always-on perception of the physical world and wake up the RGB camera when an event is detected, such as hand–object interaction [275]. This approach also has the potential to be integrated into a single image sensor in the future, forming a *hybrid vision sensor* [263], which would eliminate the need to integrate multiple types of cameras and instead enable such low-power visual perception using a single camera, making always-on perception for future AI assistants possible.

Eye trackers. Eye trackers, as an important means of attention-based user input, have attracted considerable attention for their potential to address the fine-grained control challenges of smart glasses. For example, in Apple Vision Pro, eye tracking is used to "locate" letters on a keyboard, while hand gestures are used to "click" them. However, few smart glasses currently integrate eye trackers since existing camera-based eye-tracking hardware faces limitations in cost, power consumption, accuracy, and form factor. These challenges have led to the development of new eye trackers using alternative sensing modalities: acoustic [131], millimeter-wave [151], infrared light [187], bioelectrical (i.e., electrooculography (EOG)) [280], and capacitive [188]. **Open problems:** Among these, acoustic sensing offers advantages in weight and cost but faces challenges in form-factor integration. Millimeter-wave sensing offers robustness to ambient noise and advantages in form-factor integration and performance, but faces calibration challenges. Optical sensing using infrared light offers advantages in accuracy, power consumption, weight, and cost, but remains susceptible to ambient-light noise. Therefore, event cameras are also being explored to improve robustness against ambient-light interference while reducing energy consumption [69]. Bioelectrical and capacitive sensing are also attracting attention, offering high accuracy, low power consumption, low weight, and low cost, but requiring the integration of contact electrodes into smart glasses. However, fully integrating eye trackers into smart glasses still presents challenges in maintaining tracking accuracy across a broad user population and in optimizing the cumbersome calibration process.

Additional sensors. Beyond the most common visual and auditory sensing modalities, other human senses, such as olfaction, can also provide additional information to enhance AI capabilities. For example, electronic nose sensors can be integrated into the nose pads of smart glasses for odor sensing in industrial maintenance [43] and remote odor sharing [259]. In addition, smart glasses can also help AI understand users' own physiological states, such as heart rate and blood oxygen saturation, and mental states, such as cognitive load. The proximity of smart glasses to highly vascularized facial areas has motivated the integration of various biosensors into smart glasses for monitoring key physiological and behavioral biomarkers [234], [174]. Common sensing modalities include photoplethysmography (PPG), electroencephalography (EEG), electromyography (EMG), and electrodermal activity (EDA), as well as biochemical sensing for monitoring biomarkers such as sweat electrolytes and metabolites. PPG sensors detect changes in reflected infrared light to obtain PPG signals and can be integrated into smart glasses near the nose [74], forehead [234], and auricular region [1] to monitor various physiological parameters. Another category involves electrophysiological (ExG) sensing, including EEG, which records brain electrical activity using electrodes placed on the scalp near the glasses' temples [115]; and EMG, which uses bilateral electrodes placed around the ears to measure the electrical activity of skeletal muscles and detect chewing and eating activities, supporting applications in dietary monitoring [276]. In addition, two electrodes placed at the temples and behind the ears can measure skin conductance to assess EDA and infer emotional arousal and stress levels [203]. Wireless sensing techniques, such as radar sensors placed on the glasses' temples near major arteries in the ear region, can also enable heart rate monitoring [80]. Other chemical sensors, such as an amperometric lactate biosensor and a potentiometric potassium ion-selective electrode [210], can be placed on the two nose bridge pads of the glasses to enable monitoring of sweat lactate and potassium levels. **Open problems:** Data collected from these sensors across large user populations exhibit significant variations due to anatomical differences in head size and shape, skin tone and skin conductivity, and hair distribution, as well as subtle differences in how individuals wear the glasses. In addition, sensing accuracy is also affected by other ergonomic factors, e.g., how well the glasses' temples fit the user's head. These challenges call for future efforts to increase the scale of user data collection to improve recognition performance and optimize the overall glasses design.

From 2D to 3D holographic displays. Current mainstream waveguide display solutions are primarily 2D displays with a fixed optical depth, i.e., presenting content at a fixed focal plane at a certain distance from the user's eyes, without providing 3D depth cues for virtual objects. Such fixed-focal-plane displays also introduce a fundamental

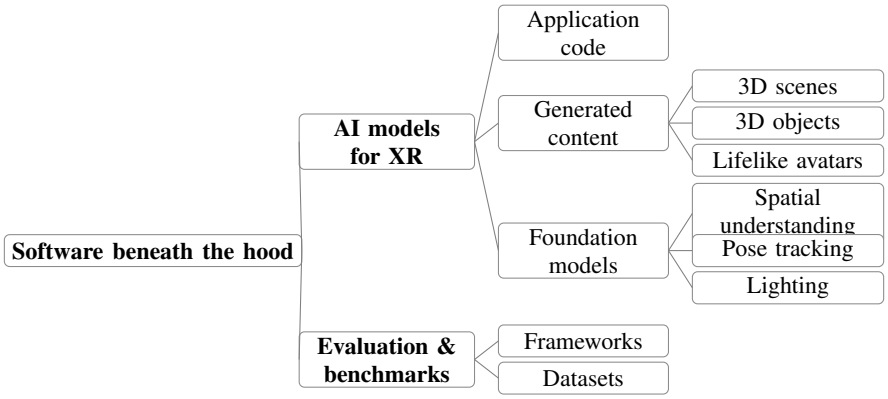

Fig. 6. Taxonomy of software beneath the hood for XR, including AI models used for XR, and evaluation and benchmarks.

perceptual issue in AR related to depth perception, known as the vergence-accommodation conflict [123], in which the eyes converge at the perceived depth of virtual objects while accommodating to the display's fixed focal plane. Therefore, recent research has also attempted to combine the advantages of waveguide displays and holographic displays to develop true 3D holographic AR glasses [91], [61]. This type of 3D holographic AR display can not only reduce the size of the light source, thereby alleviating the form-factor constraints of smart glasses [61], but also provide a larger depth range, significantly mitigating the effects of the vergence-accommodation conflict on the human visual system [123]. Meanwhile, the introduction of such 3D display capabilities also brings greater immersion to how AI conveys information in the physical world, allowing AI to move beyond merely existing on displays and potentially step out of the display into the real world perceived by the user's eyes. **Open problems:** Current 3D holographic displays are still limited by a narrow FoV (e.g., only 10° compared with at least 30° for current 2D waveguide display glasses), the need for precise optical calibration (e.g., even minor mechanical perturbations can affect display calibration), and time-consuming computational image generation (e.g., generating each display image can require up to several minutes of computation). More compact and efficient display light sources and further optimization of display algorithms have the potential to address these challenges in the future.

IV. Software Beneath the Hood

Using the hardware and sensing substrate, AI is widely being explored to provide new capabilities to AR glasses, ranging from basic system functions all the way to automatic UI design. In this section, AI usage in three broad categories is described, as shown in Fig. 6: generated content at the application level (including code, 3D scenes, and 3D objects), foundation models to support XR functions (such as spatial understanding and human pose tracking), and evaluation methods (including benchmarking frameworks and datasets). Selected works from this section are summarized in Table II. Although this work focuses on AR, there are some references to VR because their methods are also applicable to AR.

A. Application code

An exciting new use of LLMs is the "vibe coding" of XR applications through text prompts. This enables people with less coding experience to create AR apps from scratch, potentially leading to the next killer application to help drive the XR ecosystem. Google's XR Blocks [127] and Unity AI [240] are industry products that take text inputs to generate application code, 3D objects, and their associated scripts, alongside academic approaches [285], [147]. Unity AI is arguably the most usable product. It allows developers to pick their favorite LLMs to help program Unity applications (e.g., Claude, Codex, Grok), through a choice of different Unity-LLM integration mechanisms: an AI assistant inside the Unity IDE, the Model Context Protocol (MCP) with a command-line interface, or as a plugin loaded directly into existing LLM clients. The latter integration allows the LLM to control the Unity Editor and also utilize pre-defined skills [239] for common tasks, such as 3D physics, performance analysis, and setting up multiplayer capabilities. Unity AI also includes content generation tools for creating audio, animations, 3D models, etc. In this way, the Unity game engine can directly integrate with existing LLMs and development workflows.

XR Blocks, developed by another major industry player (Google), targets WebXR applications on Android devices, an alternative to the Unity ecosystem. It seeks to enable a "flywheel" of positive XR development, similar to AI development (programming frameworks like PyTorch, plus model repositories like Hugging Face, plus benchmarks like Arena AI). XR Blocks positions itself as a member of this ecosystem, namely the high-level programming abstraction. It includes a high-level scripting language to access low-level components like depth maps or camera feeds, and an agent module that can interpret the user's commands. The agent can orchestrate other modules (e.g., design a simple interaction app via text prompt), or can provide low-level functionality (e.g., an object detector in XR-Objects [40]). XR Blocks relies on Gemini but teaches it specifics about the XR Blocks abstraction. Different from Unity AI, XR Blocks can run directly in an XR headset through a web browser, rather than in a desktop PC environment. AgentAR [285] also calls tools for low-level functionality, such as object detection or optical character recognition, mainly sourced from the HCI community, rather than foundational computer vision models. Finally, XARP Tools [16] is a very interesting work that also extends a hand to AI development by making it easy to load ML modules that are typically coded in Python. XARP acts as an MCP server that exposes access to XR tools (aka ML modules) that run on the device, and the MCP client is a locally run LLM that the developer interacts with. Unity is the underlying game engine.

Along with end-to-end application development, several works study specific components to enable interactive applications. Meta has a collection of agentic tools [161], [165] to help with automation for application publishing, deployment, and testing after the code has been generated. GROMIT [93] focuses specifically on user interactions, enabling new interactions to be automatically generated during runtime. Several works [143], [253], [257] focus on developing the agent itself for specific user-facing applications such as proactive assistance while grocery shopping, described later in Section V-B.

B. Generated Content

XR applications include 3D content such as scenes, objects, and avatars. The input to content generation methods can be text prompts, images, or other multimodal inputs (e.g., an RGB-D snapshot of a real-world object to create a digital twin). The output can be a complete 3D scene or an object to be used within a scene, represented traditionally as a mesh or a point cloud, or more recently as a neural radiance field (NeRF) or set of 3D Gaussian splats (3DGS). Some content generation methods are specific to XR [37], [95], while others are drawn from computer vision [192], [173], [99], [233].

3D scenes. From text input, LLMR [37] creates Unity scenes following an agentic workflow, with an LLM acting as a Planner, code Builder, and Inspector, among other roles. The XR-specific aspects include a Scene Analyzer LLM that analyzes object hierarchy in the current Unity scene, as well as a Skill Library that includes skills such as retrieving 3D models from Sketchfab or applying interaction scripts from the MRTK library. LLMR can assemble complex scenes from primitives, such as a kitchen scene made out of basic cubes, spheres, etc. The vibe coding tools described earlier (Section IV-A) could also be used for similar purposes as LLMR, to generate scenes from text input.

Methods that take images as input to produce 3D scenes include the classical KinectFusion [171], which uses RGB-D images from a mobile Kinect camera to produce 3D reconstructions of the real world, and the more recent "world models", which use a single image to produce a virtual 3D scene, including NVIDIA's Project Lyra [211] and Atlas by World Labs / AMD [250]. Lyra uses vision transformer-based diffusion models to produce multiple video trajectories from an input image, then fuses these trajectories to produce a 3D scene. The key challenge is ensuring temporal and spatial consistency in these videos so that the final scene will be self-consistent, so Lyra calls the model's attention to relevant history and spatial correspondences. Once the video frames are generated, an off-the-shelf depth model is used to lift the video to 3D, with final outputs in the form of 3D Gaussian splats or meshes.

3D objects. Google's DreamFusion [192] is a well-known text-to-3D model that uses a frozen text-to-2D diffusion model to render multiple views, then uses these view to update the weights of a multilayer perceptron (NeRF model) that generates the 3D volumetric representation. Similar to Lyra [211], which seeks to overcome spatial consistency issues in large scenes, DreamFusion needs to address spatial consistency between multiple views of a 3D object. To do this, DreamFusion relies on a well-trained 2D generation model conditioned on different views, thus also

avoiding the need for large datasets of 3D models, which are not commonly available. LGM [230] generates a 3D Gaussian splats object from text or image inputs, also using frozen 2D models to produce multiple views of the 3D object, followed by U-Net diffusion to generate the final 3D output, a similar pipeline to DreamFusion. Rather than outputting NeRF or 3D Gaussian splat models, which are non-standard 3D representations and hence can be slow to render on resource-constrained devices like AR glasses, OpenAI's Point-E [173] and Shap-E [99] generate point clouds and textured meshes respectively, which are faster to render.

From image inputs, Tencent's Hunyuan 3D 2.0 [233] generates a 3D model consisting of a mesh and its texture. It does this in two steps: first, generating the mesh using a transformer-based diffusion model, and second, generating the texture using multi-task attention, where the tasks are ensuring consistency with the input image, 3D mesh from the previous step, and views from different camera viewpoints. Such models have been incorporated into XR authoring tools; for example, Hunyuan 3D 3.0 is used by Unity AI [240], and Thing2Reality [78] uses LGM internally to produce 3D Gaussian models for their multi-user meeting platform.

Lifelike avatars. Many academic works study the generation of photorealistic avatars for use in XR [286], [118], [26], [198], alongside industry approaches like Meta's Codec Avatars [145]. There are two main steps in generating real-time animated avatars. (1) *Generate an avatar*, consisting of an underlying model of human joints (e.g., SMPL-X) plus the physical appearance or "skin" of the avatar. LHM++ [198] fuses 3D geometry from a SMPL-X template with 2D images of the subject by using a transformer model to produce tokens, which are then decoded into 3D Gaussian splats that represent a human avatar. The target pose of the avatar is achieved using a standard computer graphics technique called Linear Blend Skinning, plus a custom neural network, to deform the splats into the target pose. (2) *Animate the avatar* in real time. The overhead of rendering in real time on resource-constrained commercial XR devices can be high. VRGaussianAvatar [219] uses an earlier version of LHM++ as its base model, and achieves real-time pose changes by tracking the user's head and hand motions on the Meta Quest 3, rendering the pre-generated LHM avatar model with the current pose on a nearby server, then streaming the rendering results as JPEG images to the headset. This consumes approximately 5 Mbps over a network link. Meta's Codec Avatars focuses on human faces and performs both steps using a variational autoencoder: first, images of the face and a mesh representation are fed into an encoder to produce a latent representation; second, during runtime, the latent representation is decoded into a mesh and texture using additional partial images of the user's face, captured by the headset's inward-facing RGB and infrared cameras. The renderings must meet a target frame rate of 90 FPS for VR. ESCA [286] further optimizes the frame rate by designing a custom hardware accelerator to achieve 100 FPS.

C. Foundation Models

Beneath 3D content generation, many works have investigated foundation models to support basic support functions of XR systems. Key functions include spatial understanding, object detection, human pose tracking, and lighting estimation, to understand the world around the AR device and the human user's relationship to it.

Spatial understanding. Computer visiontechniques are core to spatial understanding like depth estimation [138], [144], [243], scene segmentation [273], [265], and object detection [98], [142], and they are being explored for SLAM [25], [169]. For depth estimation, ByteDance's Depth Anything [138] is a popular model for estimating scene depth primarily for monocular cameras, although it also works with multiple input views. It uses DINOv2 image features plus a custom head to predict depth maps. Meta's Argos [243] is valuable for its overview of lightweight depth estimation on resource-constrained devices such as a mobile phone (Samsung Galaxy S8). It uses another Meta model as its internal depth estimation network and extends it to stereo cameras. MultiHeadDepth [144] further optimizes Argos for mobile devices, by replacing its expensive pre-processing with ML models, evaluating on dataset collected from Project Aria smart glasses and running on a Jetson. For scene segmentation, the Segment Anything Model (SAM) and its lightweight version, MobileSAM [273], are foundation models used in several other works previously mentioned (SemanticXR [214] and Thing2Reality [78]). SAM is built on a vision transformer encoder to produce an image embedding, an optional prompt input, and a lightweight decoder to produce the masks. MobileSAM is a distilled version of the encoder (100$\times$ fewer parameters) for speedup on mobile devices. Gaussian Grouping [265] also segments scenes, but specifically scenes comprised of 3D Gaussian splats. It assigns an ID number to every splat, so that splats with the same ID belong to the same segment/object, and learns the ID by

TABLE II
SELECTED WORKS USING AI FOR AR, INCLUDING CODE AVAILABILITY AND DEPENDENCIES ON OTHER AI METHODS.

Category	Name	Inputs	Outputs	AI methods	Open source
Application code	Unity AI [240]	Text	AR application	Claude, Codex, Grok	No
	XR Blocks [127]	Text	AR application	Gemini	Yes
	XARP Tools [16]	Text	AR application	LM Studio	Yes
	AgentAR [285]	Text	AR application	GPT-4o	No
	GROMIT [93]	Player action	Interaction code	GPT-4	Yes
3D content	LLMR [37]	Text	3D scene	GPT-4	Yes
	Project Lyra [211]	Image, (text)	3D scene (3DGS, mesh)	Diffusion (DiT), Depth Anything	Yes
	SemanticXR [214]	Image, depth, pose	3D map	RAM, Grounding DINO, MobileCLIP, MobileSAM [273]	No
	DreamFusion [192]	Text	3D objects (NeRF)	Diffusion (U-Net)	No
	LGM [230]	Image, (text)	3D objects (3DGS)	Diffusion (U-Net)	Yes
	Point-E [173]	Text	3D objects (point cloud)	Diffusion (transformer)	Yes
	Shap-E [99]	Text	3D objects (mesh)	Diffusion (transformer)	Yes
	Hunyuan 3D 2.0 [233]	Image, (text)	3D objects (mesh, texture)	Diffusion (DiT)	Yes
	PartInteractor [95]	Speech/image/sketch	3D objects	GPT-4o, GPT-Image-1, OmniPart	No
	Thing2Reality [78]	Text, image, sketch	3D objects (3DGS)	LGM [230], MobileSAM [273]	No
	Codec Avatars [286]	Images, mesh	3D avatar (mesh, texture)	Variational autoencoder	Yes
	LHM++ [198]	Image	3D avatar (3DGS)	Transformer, DINOv2	Yes
	VR Gaussian Avatars [219]	Image	3D avatar (3DGS)	LHM++ [198]	Yes
Foundation	Depth Anything [138]	Image	Depth map	DINOv2, attention	Yes
	Argos [243]	Stereo images	Depth map	CNN (Tiefenrausch)	No
	MultiHeadDepth [144]	Stereo images	Depth map	CNN	Yes
	MobileSAM [273]	Image	Segmentation masks (2D)	ViT	Yes
	Gaussian Grouping [265]	3D scene (3DGS)	Segmentation masks (3DGS)	Custom loss function	Yes
	MASt3R-SLAM [169]	Monocular images	3D scene, pose	Mast3r	Yes
	Active Neural SLAM [25]	Images	2D scene, pose	CNN	Yes
	RealityReplay [32]	Images	AR labels of motion	Segmentation (Detic), saliency detection (TASED-Net)	Yes
	RL-LABEL [29]	Images	AR label positions	Deep reinforcement learning	No
	Neural Band [100]	EMG (wrist)	Hand gestures, handwriting, cursor position	1D conv, conformer, LSTM	No
	Gardner et al. [54]	Image (LDR)	Image (HDR)	CNN	No
	DeepLight [125]	Image (LDR)	Image (HDR)	Encoder-decoder	No
	EnvMapNet [217]	Image (LDR)	Image (HDR)	Encoder-decoder	No

training with a custom grouping loss function. Object detection models such as YOLO [98] are commonly used for fast inference, while models such as Grounding DINO [142] support natural-language input. In practice, Meta's AI Building Blocks [161] provides Unity developers with access to object detection, speech-to-text, and text-to-speech models, either cloud-hosted (e.g., OpenAI's Whisper model or any Hugging Face model like ResNet), on the same Wi-Fi network, or locally on the device.

Simultaneous localization and mapping (SLAM) is key in XR devices. SLAM constructs a map of an unknown environment and also localizes the device in that environment. Klein and Murray [111] wrote the seminal work on real-time SLAM on a mobile camera for AR, paving the way for modern AR devices. Popular open-source codebases include ORB-SLAM3 [19] and VINS-Mono [197], which rely on classic SLAM techniques like structure from motion, pose graph construction, and nonlinear optimization. Attempts to incorporate AI into SLAM include MAST3R-SLAM [169], which replaces classical image matching through feature correspondences with a learned 3D prior, and Active Neural SLAM [25], which uses CNNs to predict a 2D map and the device pose, as well as path planning for a robot navigation. Beyond general scene understanding, there are proposals unique to XR, like highlighting motion changes in the real world [32] through a combination of scene segmentation filtered by predicted saliency, or placing AR labels on moving objects to avoid occlusion [29] using deep reinforcement learning.

Pose tracking. Tracking of human pose can be done from egocentric videos [75], [242], of human hands specifically [55], [68], or of objects being manipulated by hands [266]. Even the lower body can be estimated (somewhat inaccurately) from the pose of the head and hands by Meta's Generative Legs [163]. Beyond popular camera and IMU sensors, exotic sensors like EMG have recently shown promise in Meta's Neural Band [100], which analyzes data from more than 11,000 participants to train a model to detect hand gestures, writing, and cursor movements. The models used a combination of frequency/power features, 1D convolutions, LSTMs, and conformers. XRBench [117] provides a wealth of further references on the underlying ML methods for human sensing and spatial understanding for XR systems.

Lighting estimation. There is a large body of work on lighting estimation, where the problem is to estimate the lighting in the real world, so that the virtual objects can blend in by being rendered with similar lighting. This is typically done using light probes placed in the virtual scene that contain empirical data about how nearby objects should be lit. These light probes are created from high dynamic range (HDR) images, which store data in 32 bits per pixel in order to better capture the full range of real-world lighting conditions, compared to the 8 bits in standard low dynamic range (LDR) images. The challenge is that mobile devices like XR headsets only capture low dynamic range images, so it is difficult to create light probes representing the real world and use these to re-light virtual objects. Seminal work by Gardner et al. [54] uses a CNN to learn the mapping between LDR and HDR images, using a training dataset of 2,100 HDR panoramas. Further work by Google was incorporated into their ARCore Lighting Estimation API [125]. Their key insight was to overcome the dearth of HDR images by collecting additional training data using a custom rig, with three spheres of different reflectivity visible in images captured by a mobile phone. Related approaches were also explored by Apple using generative adversarial networks [217] and by academics optimizing for mobile devices [193].

D. Evaluation and Benchmarks

Benchmarks and datasets are an important component of evaluating AI-enabled AR applications. They are often borrowed from other domains and repurposed for AR.

Datasets. There are many egocentric video datasets used for evaluating AR, sometimes with other multimodal data included, such as the ego category in OVO-Bench [135], and others [128], [191], [36]. These datasets are often used for visual question-answering tasks. Some egocentric datasets specifically focus on procedural tasks [5], [258], [46], [124]. Open4D [221] is an academic effort to store 3D and 4D models and provide tools to manipulate them. There are a few datasets recorded specifically from XR devices, such as SuperGlasses [96] (recorded from the Ray-Ban Meta, Xiaomi, and RayNeo smart glasses) and HoloSet [23] (recorded from the HoloLens 2), which contains additional information from sensors beyond cameras, such as IMU. Project Aria is a hardware effort from Meta designed for data collection from a variety of sensors on the glasses, such as RGB cameras, eye tracker, and PPG sensors. A very nice collection of datasets has resulted from this effort [194].

Benchmarking frameworks. Datasets inherently have the limitation that they cannot be used for closed-loop experimentation, i.e., the dataset remains the same no matter the behavior of the XR system or the guidance provided. Open-source AR systems allow for dynamic experimentation and potentially user studies. While some of the previously covered works have released open-source code, as noted in Table II, several works have the explicit goal of benchmarking AR systems. SIGMA [10] from Microsoft Research focuses on procedural guidance and is built on top of the more general-purpose \psi framework to combine AI and sensors when working with multimodal streaming data. XRBench [117] from Meta provides examples of pipelined AI models for AR-specific purposes like eye tracking and speech recognition, along tabulating useful datasets and specific performance requirements. Their provided code is focused more on reproducing the evaluation results from the paper, rather than actively running the AI models in the pipeline. Although both the SIGMA and XRBench repositories no longer seem to be maintained, they are good examples of efforts to encourage reproducible systems building in AR and AI. ILLIXR [84] is actively maintained and support for foundation models is being added, according to the authors' personal communications. There are also simulation environments for AR-adjacent scenarios, like vehicles [227], speech-mediated human-robot interactions [71], as well as general simulators for training autonomous agents, like Habitat [205] or Gymnasium [235] that could be adapted to AR agents, albeit without humans in the loop.

V. AI-Enabled AR Applications

AI-enabled AR applications range from providing real-time speech captioning to guiding robotic swarms via gestures and high-level voice commands. In this section, we present a *functional taxonomy* of AI-enabled AR applications, centered on their core functionalities. For each functionality, we examine the underlying AI and computer system requirements, current bottlenecks, and deployment readiness. An overview of our taxonomy is presented in Fig. 7.

A. Situated Perception and Recall

The use cases, requirements, limitations, and readiness of different application capabilities in the *situated perception and recall* application category are summarized in Tables III and IV. *AI transformation:* Long-proposed egocentric perception-based applications have been restricted, in practice, by the need to train dedicated models for each targeted application setup. The emergence of foundation models lifted this restriction, moving long-envisioned applications closer to practical deployment.

Speech translation and captioning. Translators that convert foreign speech into speech in a language that a person understands are omnipresent in sci-fi: Universal Translator (Star Trek) and Babel fish (Hitchhiker's Guide to the Galaxy) are among numerous examples of this vision. Translation and live captioning are now default features of many consumer smart glasses, including Meta Ray-Ban Display, Even Realities G2, and Rokid Glasses. Several smart glasses specifically designed for *deaf and hard-of-hearing users*, such as Captify Pro and Hearview Lite, are also on the market. A 2025 *New Yorker* article titled "Subtitling your life" referred to eyeglasses that display real-time speech transcription on their lenses as a "game-changing breakthrough" [177]. *Limitations*: Recent studies have compared the performance of different commercial offerings and characterized their limitations [201], which include reduced performance in challenging acoustic conditions and difficulty establishing sound direction and speaker attribution. Due to strict conversational latency requirements, current systems are largely limited to captioning or translating speech verbatim, without conveying tone, emphasis, and other prosodic and paralinguistic information that humans process when communicating. Additionally, an inherent challenge for Babel fish-like live translation applications is that translations of partial utterances may need to be revised as more speech becomes available. While users may expect captions to resemble those provided with prerecorded media, live translated captions are inherently less stable and more error-prone [180], reflecting fundamental tradeoffs between latency, stability, and translation accuracy.

Visual and multimodal understanding: on-demand. This category relies on visual understanding invoked on demand: the wearer directs the system to something in view and asks about it. This capability is now a default feature of many consumer smart glasses, including Ray-Ban Meta's Meta AI and RayNeo X3 Pro's AI Assistant. Studies of smart glasses use demonstrate users exploiting these capabilities to read and translate text; identify unfamiliar objects and landmarks; ask referential questions; obtain context-sensitive recommendations; and support everyday problem

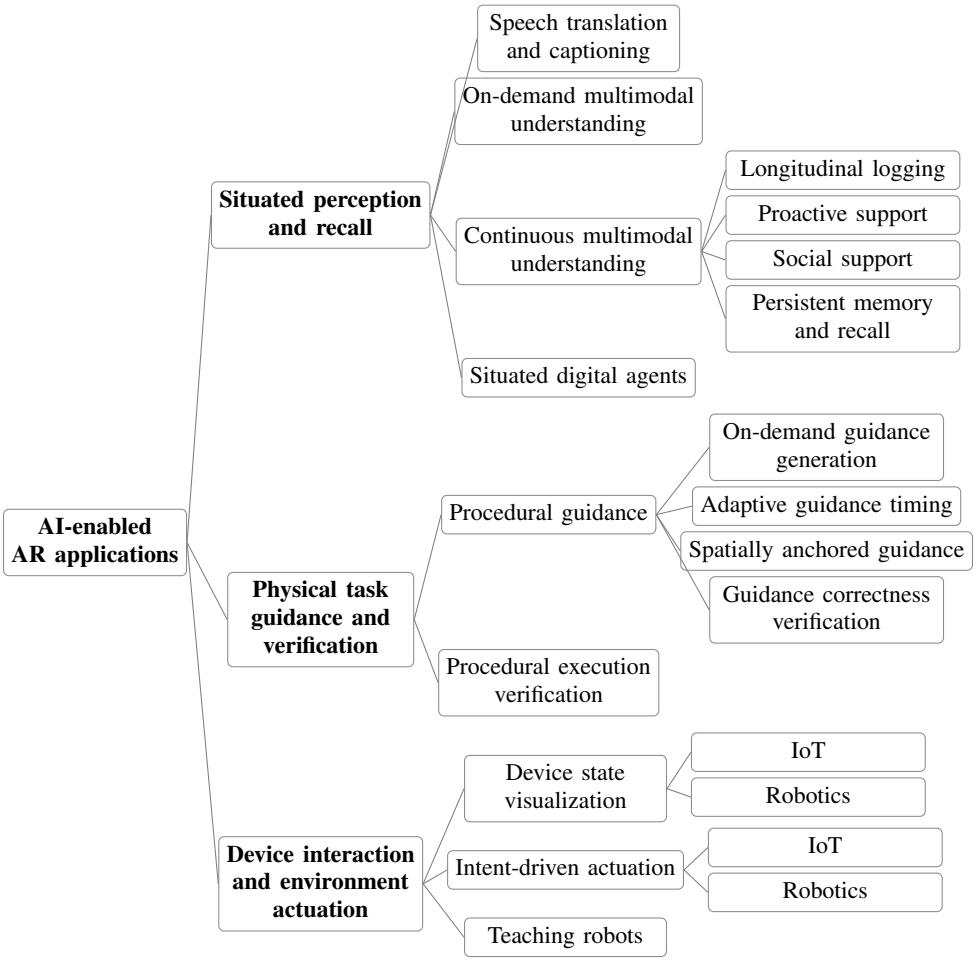

Fig. 7. A functional taxonomy of AI smart-glasses applications.

solving and decision making [122], [247], [287]. A particularly important line of applications is egocentric AI for *blind and low-vision (BLV) users*, including reading text, identifying products, describing visual attributes [58], and recognizing faces. Meta explicitly promotes its glasses for this population. Documented real-world misuses of these capabilities have also emerged, including smart glasses-assisted cheating on exams and competitions, and social exploits that turn face recognition into a real-time doxxing tool [141], [172]. *Limitations*: Despite rapid commercialization, **reliable question answering from unconstrained egocentric views remains far from solved**. Egocentric cameras produce blurred, occluded, poorly framed, and low-light observations that challenge even advanced models, which can also confidently give incorrect answers. Evaluations find large performance losses under these conditions [24] and a substantial gap between multimodal models and humans on BLV-relevant queries [251]; a wide range of first-person accounts detail notable instances of misidentification, miscounting, and confidently incorrect answers [7], [287]. Multiple safety-critical use cases for smart glasses have been proposed, including identifying food allergens, interpreting medical packaging, supporting BLV users in safety-critical settings, and distinguishing edible from toxic wild mushrooms. Realizing those applications would require significant improvements in model accuracy and reliable estimates of the basis and level of confidence of their judgments.

Visual and multimodal understanding: continuous. This broad category of applications becomes possible when an egocentric AI system monitors the wearer and surrounding environment *persistently*, rather than only when explicitly invoked. *Limitations:* These applications are constrained by the limited battery, compute, memory, and thermal budgets of wearable devices, motivating extensive work on low-power sensing, efficient on-device inference, and selective activation of resource-intensive sensors. Continuous sensing also raises particularly acute privacy

TABLE III
SITUATED PERCEPTION AND RECALL: USE CASES AND REQUIREMENTS

Application Capabilities	Illustrative Use Cases	Required System Capabilities	Required AI Capabilities	Latency Requirements
Speech translation and captioning	Foreign-language conversations; accessibility for deaf and hard-of-hearing users	Microphone + audio output (spoken translation) or near-eye display (captions)	Speech recognition; translation; speech synthesis for audio output	Conversational
Visual & multim. understanding: On demand (O)	BLV assistance; object identification; everyday scene questions	Egocentric camera; user query input; audio or display output	Visual scene understanding; visual question answering	Interactive; task-dependent
Visual & multim. understanding: Continuous (C)	Ongoing monitoring of activities and surroundings	O + continuous or event-sampled sensing	O + ongoing scene/activity understanding	Application-dependent
Logging	Activity and dietary monitoring; lifelogging; personal or research analytics	C + time-stamped logging/storage	C + activity and event extraction	Relaxed
Proactive support	Hazard alerts; just-in-time nudges; everyday cognitive assistance	C + event detection and trigger processing	C + intervention-need and timing decisions	Task-dependent; near-real-time
Social support	Communication accessibility; people needing social or attentional support	C + multimodal human sensing; audio/visual feedback	C + speech/turn-taking and social-cue understanding	Conversational
Persistent memory and recall	People with memory challenges; everyday recall and contextual reminders	C + persistent storage, indexing and retrieval	C + entity continuity, temporal memory and retrieval	Retrieval: interactive. Reminders: task-dependent
Situated digital agents	Hands-free notes, scheduling, shopping, and digital follow-through	O + authorized external tools and services	O + planning and tool-use agents	Task-dependent; interactive for confirmations

concerns [202], [106].

Longitudinal logging. Continuous multimodal sensing can turn glasses into a behavioral measurement device for monitoring a user's eating behavior [97], facial expressions and emotions [255], and daily activities as a whole [260], among many others. These applications build on a long tradition of lifelogging and egocentric activity recognition, which used wearable cameras and sensors to capture everyday experience and increasingly to infer activities and behavioral patterns from it [189], [65]. *Limitations*: These applications' requirements are more forgiving than those that depend on real-time inference: logging is generally delay-tolerant, and its outputs can usually be reviewed and corrected before they are used for subsequent analysis or decision making (inference accuracy, however, remains important for usability). However, for many applications, prolonged always-on sensing is critical; this motivates research in low-power solutions across the stack.

Proactive support. A system that proactively monitors the environment can provide assistance based on the context captured by the glasses, rather than an explicit user request. Such assistance can surface contextually relevant facts, describe environmental changes, or warn about hazards [184]. It can also help users monitor or change their behavior, such as by prompting healthier choices or interrupting unwanted habits, as long envisioned in the literature on *just-in-time adaptive interventions* (JITAIs). Relevant work has explored monitoring smoking behavior [88], delivering real-time prompts during eating [178], and generating personalized contextual nudges [45]. Wearable cognitive assistance, which uses first-person sensing to provide timely prompts or guidance for everyday functioning [66], [67], falls in this category as well. *Once sufficiently reliable, these applications could be societally transformative: proactive wearable AI could improve safety by detecting hazards that escape human attention and substantially increase the independence of people with sensory impairments*. *Limitations:* Providing proactive

TABLE IV
SITUATED PERCEPTION AND RECALL: LIMITATIONS AND DEPLOYMENT READINESS.

Application Capabilities	Representative Papers	Limitations	Readiness
Speech translation and captioning	[201], [27]	Noise and overlapping speech; speaker attribution; accuracy–latency tradeoffs	Commercial: live captions and translation (selected devices/languages)
Visual and multim. understanding: On demand (O)	MobiVQA [20], Vinci [81], [160], [52]	Robustness of egocentric perception	Mass-market for informational use; not safety-critical
Visual and multim. understanding: Continuous (C)	*See subcategories below*	O + limited continuous runtime; bystander privacy	*Application-dependent*
Logging	EgoLife [260], DietGlance [97]	C	Recording: commercial; semantic logging: field-tested research
Proactive support	AttentionAR [184], Mirai [45]	C + uncertain need for intervention; false or disruptive alerts	User-tested prototypes
Social support	Understood [278], Sign-to-911 [64], SocialMind [256]	C + ambiguous social cues; assistance disrupting interaction	Communication aids: commercial; social coaching: clinical research
Persistent memory and recall	MemPal [155], LightMem-Ego [28]	C + accumulating errors; identity continuity	Dialogue recall: commercial; visual episodic memory: field-tested research
Situated digital agents	VisionClaw [143], Egoc. Co-pilot [262]	O + unsafe or unauthorized actions; physical prompt injection	Bounded actions: commercial; scene-grounded agentic action: research

assistance requires both understanding the context captured by the glasses and deciding whether and when an intervention is desired by the user. The latter is particularly challenging, as user preferences are difficult to infer and collecting relevant training data is non-trivial. Unwanted interventions can be disruptive, while delayed interventions may arrive too late to be useful.

Social support. Smart glasses can support wearers in real time during social interactions, providing feedback on speaking style [13], recognizing conversational partners' affect [248], associating captions with individual speakers [185], bidirectional sign-to-speech mediation [64], real-time summarization, word suggestions, and topic-shift detection [278], and proactive suggestions based on verbal and nonverbal social cues [256]. Such assistance can require interpreting conversational context, tracking participants and turn-taking, understanding nonverbal behavior, translating between communication modalities, and generating responses at sufficiently low latency to avoid disrupting interactions. *Limitations*: While modern foundation models have improved recognition of visual and auditory content, many forms of social support require inferences about intent, affect, relationships, and conversational norms – states that are difficult to capture and label at scale, and that vary across individuals, cultures, and communication styles. They are likely to remain challenging for AI models, especially at the accuracy and low latency required by real-time interaction.

Continuous understanding with persistent memory. When egocentric audio or video are captured over long periods of time, AI-based reasoning can be coupled with persistent memory to support reasoning about past experience. This enables the long-envisioned application of smart glasses as a "memory prosthesis," including for people living with dementia and others with significant memory impairments [73], [120]. Recent AI and ML techniques extend visual question answering to queries over past experience [260] and construct temporal and spatial memories to track objects, events, locations, and their evolution over long egocentric recordings [150]. Representative user-facing capabilities include recalling past events [129]; verifying whether actions were completed or objects left in particular states, such as whether medication was taken or a stove turned off; and providing memory-driven, context-aware reminders [155]. These systems address an important need: human memory is notoriously imperfect [208], especially under sleep deprivation, interruptions, and disruptions in routines, and memory lapses can be safety-critical. *Limitations:* Systems with persistent memory are substantially more complex than systems that reason only over current observations. A key challenge is maintaining persistent identity for objects and people

across encounters, viewpoints, and environmental conditions. Additionally, unavoidable perception and reasoning errors accumulate and propagate over time. Current work recognizes the importance of allowing users to review and correct AI-inferred memories [129], [28]; this, however, creates a tension with open-ended question answering over a user's experience, as it is not feasible to require users to curate all remembered information that might later become relevant to unanticipated queries.

Situated digital agents. Smart glasses can serve as interfaces to software agents that take digital actions on the user's behalf, grounding those actions in observed physical context. This holds the promise of reducing the gap between physical and virtual worlds, as users can encounter something in the physical world and act on it digitally. Recent work couples egocentric perception with digital action: Ego2Web connects egocentric video understanding with downstream web tasks [271], while Egocentric Co-Pilot and VisionClaw use first-person context to ground tool use for external services [262], [143]. Wearable agents can also act proactively, inferring from visual, audio, digital, and historical context when assistance is needed rather than waiting for an explicit request [241], [258]. *Limitations:* Bringing agentic AI to smart glasses compounds the inherent safety and security risks of delegating actions to AI systems [166], since in these settings agents' actions are grounded in inherently error-prone glasses-based environmental context and user intent understanding. In addition, using environmental context as input to agentic behavior creates multiple additional opportunities for adversarial input manipulations, such as physical prompt injection attacks.

B. Physical Task Guidance and Verification

A key, long-envisioned opportunity for AR is to support people as they perform complex, multi-step physical procedures. Assembly [246], repair [70], and cooking [183] have all been extensively considered in this context. *AI transformation:* Previously, these long-envisioned applications of AR relied on procedural guidance that was carefully pre-authored, which restricted their applicability to standardized procedures. Advances in generative AI make it increasingly possible to generate guidance on demand, opening the door to using smart glasses to provide *guidance for custom protocols*, authored or adapted for a single instance rather than reused at scale. One example of this category is guidance and oversight for bench scientists following a bespoke experimental protocol, as targeted by the LabOS [34] system. Such bespoke guidance and customized guidance oversight also open exciting opportunities in healthcare, where patients and caregivers are often faced with performing complex, unfamiliar procedures.

Procedural guidance can be seen as having several components. First, the multi-step guidance itself needs to be conjured. For each individual step, the system then needs to establish *when* to present step-appropriate guidance, place the guidance where needed in the context of a scene, and check its correctness. Following the completion of a step, egocentric co-pilots are envisioned to verify the correctness of procedural execution and adjust the guidance if necessary. Tables V and VI summarize the requirements, limitations, and deployment readiness of these distinct procedural guidance-related application capabilities.

Guidance generation, placement, and oversight. Recent advances in generative AI made the very first step in this process, conjuring multi-step guidance, increasingly straightforward. Guidance for common procedures can be directly generated by LMMs, producing step-by-step AR instructions from a single user query without prior authoring or demonstration [282]. Additionally, a significant recent body of work has been exploring creation of guidance from instructional videos, mining large-scale collections of online tutorials to learn and generate step-by-step visual instructions [79]. Establishing *when to present guidance* is a harder problem: it requires recognizing the step of the procedure, itself the target of a significant body of work on egocentric action segmentation and procedural step tracking, at a latency that is low enough for the user to not be slowed down or inconvenienced. A wide range of recent papers, including Satori, ProAssist, and Pro^{2}Assist, explicitly target when to intervene, reasoning from task progress, user state, and streaming egocentric observations [126], [281], [253]. Placing guidance requires both strong low-latency semantic scene understanding and strong on-device 3D localization and mapping to anchor guidance at the correct location and scale, and keep it stable as the user moves; recent systems that aim to provide it include TAGGAR and Guided Reality [225], [282]. While the correctness of generated guidance has received little explicit study as a standalone problem, recent work shows that LMMs can jointly reason over physical scenes and virtual overlays, enabling automated verification [41], [252]. *Limitations:* Procedural guidance is significantly more complex than situated perception alone, as it requires maintaining an up-to-date, correct understanding of

TABLE V
PHYSICAL TASK GUIDANCE AND VERIFICATION: USE CASES AND REQUIREMENTS.

Application Capabilities	Illustrative Use Cases	Required System Capabilities	Required AI Capabilities	Latency Requirements
Custom-protocol guidance	Wet labs; healthcare; bespoke maintenance and repair	Task input (text/voice); edge/cloud compute; camera if scene-grounded	Scene understanding; guidance generation	Generation before execution can be slower; mid-task changes are task-dependent
Proactive guidance	Novices performing complex multi-step tasks; cooking; repair	Ongoing egocentric sensing; event-triggered or low-power processing	Scene understanding; step/progress recognition; decision when and how to intervene	Task-dependent, near-real-time; useful before next relevant action
Spatially anchored guidance	Assembly; complex repair; assistance for novices and apprentices	AR display, 6DoF tracking, scene mapping/depth and calibration	Scene understanding; semantic object grounding; 3D anchor and orientation/scale selection	Tracking and registration near-real-time; content generation may occur ahead of the step
Guidance correctness verification	Checking generated instructions and spatial overlays	Access to guidance content; egocentric camera for scene-grounded verification	Joint reasoning over guidance and physical scene; procedural knowledge; error detection	Task-dependent; before user action
Procedural execution verification	Manufacturing quality checks; lab protocols; healthcare; safety-critical procedures	Egocentric camera; optional other sensors; no spatial display required for monitoring.	Scene understanding; fine-grained step/action and object-state tracking; mistake detection.	Logging: relaxed. Corrective interventions: task-dependent, before errors propagate.

the evolving task state, and a reliable estimate of whether, when, and how specifically the user would want to be guided. State-of-the-art research platforms demonstrate substantial error rates in guidance generation, spatial placement, and intervention timing; while user studies conducted with these platforms suggest that some errors are tolerated and participants find the provided guidance useful regardless [126], [282], [253], high error rates have a significant potential to stand in the way of intended users' willingness to rely on these systems in unconstrained real-world scenarios.

Procedural execution verification. The possibility of real-time procedural verification is exciting; it represents another potential *AI transformation*: going from simply displaying the guidance to verifying that it is correctly followed is potentially societally transformative as procedural mistake detection can improve safety and efficiency of procedures across a wide range of tasks and industries. *Limitations*: Procedural verification is a complex AI task that requires a combination of procedural step detection, fine-grained user action and hand-object interaction recognition, and robustness to open-set procedural and execution mistakes, among others. A substantial body of work in the CV community has been targeting egocentric mistake detection [49]; however, the problem remains difficult, with recent work pointing out the contrast between modern models' excellent abilities to provide instructions and deficiencies in identifying errors and suggesting corrections [140]. The need to provide oversight at a latency that does not impede the user's progress complicates the problem further.

C. Device Interaction and Environment Actuation

A key functionality of smart glasses is their ability to mediate interaction with other devices, in particular robots and IoT-enabled environments. These ideas have a long history in robotics, where AR displays have been explored for human–robot communication since the 1990s [35], [228], and in ubiquitous computing and IoT, where wearable displays have long been envisioned as interfaces to "intelligent" environments [114], [283], [209]. In these systems, smart glasses and robots or IoT devices can function relatively independently or be more tightly integrated, combining their observations for enhanced spatial and semantic understanding. Tables VII and VIII summarize the system and AI requirements, key limitations, and current deployment readiness of these application capabilities.

TABLE VI
PHYSICAL TASK GUIDANCE AND VERIFICATION: LIMITATIONS AND DEPLOYMENT READINESS.

Application Capabilities	Representative Papers	Limitations	Readiness
Custom-protocol guidance	LabOS [34]	Correctness and completeness of generated procedures	Research prototypes; expert review needed for consequential tasks.
Proactive guidance	Satori [126], Pro^{2}Assist [253]	Ambiguous task progress; inappropriate or disruptive interventions	Research prototypes and controlled user studies.
Spatially anchored guidance	Guided Reality [282], TAGGAR [225]	Action-level spatial grounding; placement and registration errors	Authored spatial AR: commercial. AI-generated spatial placement: research prototypes.
Guidance correctness verification	VIDDAR [252], [41]	Unreliable procedural and spatial reasoning; missed guidance errors	Research
Procedural execution verification	PREGO [190], HoloAssist [246]	Unobservable task states; open-set mistakes; missed errors	Narrow visual pass/fail checks: commercial. Open-ended procedural verification: research.

TABLE VII
DEVICE INTERACTION AND ENVIRONMENT ACTUATION: USE CASES AND REQUIREMENTS.

Application Capabilities	Illustrative Use Cases	Required System Capabilities	Required AI Capabilities	Latency Requirements
Visualizing device state: IoT	Representing sensor readings, device status, automation dependencies	Device connectivity; state retrieval; AR display	Optional: state interpretation; visualization generation	Interactive
Visualizing device state: robotics	Displaying robots' trajectories, status, safety zones, affordances	Device connectivity; state retrieval; AR display; continuous spatial registration	Optional: motion/intent prediction; affordance estimation; spatial reasoning	Action-coupled for dynamic overlays
Translating human commands into action: IoT	Controlling devices through speech and pointing	User input; device identification; authorized control APIs	Intent recognition; device grounding; command generation	Interactive
Translating human commands into action: robotics	Directing robot movements and tasks through speech and gestures	User input; robot identification; authorized control APIs; spatial registration	Spatial intent grounding; task planning; feasibility checking	Commands: interactive; Execution: action-coupled
Teaching robots	Teaching robots manipulation tasks from egocentric human demonstrations	Egocentric capture; demonstration collection infrastructure; optional: additional sensing	Human action reconstruction; human-to-robot action transfer; robot policy learning	Relaxed

Limitations: Interoperability remains a practical hurdle, as smart glasses, robots, and IoT devices rely on disparate protocols, APIs, and data representations, requiring considerable custom integration.

Visualizing device state: IoT. AR can display otherwise invisible sensor readings, device state, automation logic, and relationships between connected devices directly within the physical environment. Such visualizations do not inherently require AI, but generative AI can aid in their creation: e.g., recent work employs advances in AI to interpret sensor readings and automatically generate AR visualizations and device interfaces [63], [113]. *Limitations*: It is often non-trivial for AR devices to associate a digital IoT entity with a physical device [113]; custom integration may be required for reliable operation.

Visualizing device state: robotics. A classical use of AR in human-robot interaction is to convey a co-located robot's otherwise hidden state and intent to human collaborators [228]; this may include visualizing planned motion, operational status and safety zones, and physical capabilities such as reachability. Many of these applications rely on glasses' ability to place virtual content in a specific 3D location relative to a robot and to keep it aligned with the robot as the robot moves; this requires advanced spatial understanding on the glasses and the robot, and

TABLE VIII
DEVICE INTERACTION AND ENVIRONMENT ACTUATION: LIMITATIONS AND DEPLOYMENT READINESS.

Application Capabilities	Representative Papers	Limitations	Readiness
Visualizing device state: IoT	[101], LO-RIOT [113]	Physical–digital device association	Situated AR visualization: user-tested prototypes. AI-generated visualization: research
Visualizing device state: robotics	RoboVisAR [148], OOHRI [244]	Spatial misalignment and stale state; unreliable predicted motion or inferred affordances; visual clutter.	User-tested research prototypes.
Translating human commands into actions: IoT	BLESelect [279], HoloHome [154], VisionClaw [143]	Ambiguous intent and device selection; invalid or unauthorized commands; conflicting automations.	Voice-based control: commercial. Situated AI-based control: research.
Translating human commands into actions: robotics	FAM-HRI [119], MRPos [86], EDITH [121]	Ambiguous intent; infeasible or unsafe plans; physical execution hazards.	AR-based commands: user-tested prototypes. Open-ended AI control: research.
Teaching robots	EgoVLA [261], EgoMimic [102], EgoDEX [76]	Incomplete and inaccurate interaction sensing; human–robot embodiment gap.	Recording demonstrations: commercial. Robot learning: AI research.

continual alignment between them, raising the required system capabilities significantly compared to the other applications. *AI transformation*: AI is actively employed in increasingly autonomous robotic operations, in both perception and action; this increases the opportunities for smart glasses as a means of visualizing robot state, which now includes AI-derived representations such as semantic affordances, capabilities and limitations [244], or generated action plans [87]. *Limitations*: Joint glasses-robot spatial registration remains brittle in practice, as aligning independently evolving spatial estimates across devices is sensitive to differences in viewpoints and sensor characteristics, environmental changes, and drift [9]. For visualizations of AI-generated robot plans and capabilities, or for establishing shared human-robot semantic understanding of a space, another limitation is the unreliability of foundation models in spatial reasoning and physical feasibility assessment [284]; progress may require more explicitly spatial and embodied representations and training, rather than model scale alone [254].

Translating human commands into actions: IoT and robots. Smart glasses provide an intuitive interface for directing nearby devices, allowing users to specify targets and actions through speech, pointing, head motion, and gaze [51], [279], [86]. *AI transformation:* Foundation models are transforming these interactions by moving beyond explicit commands to inferring partially expressed intent and translating high-level requests into action plans [87], [121]. A related line of work aims to employ smart glasses, particularly those equipped with eye trackers, to sense user attention, situation awareness, and cognitive workload and expose these states to help robots adapt to the state of their human collaborators [199], [179], [157]. *Limitations:* As with situated agents (see Section V-A), errors in glasses-based perception and intent inference can lead to unintended physical actions. In IoT, this can result in controlling the wrong device, issuing invalid commands, or triggering conflicting automations [133]; in robotics it can lead to infeasible plans or unsafe motion [85], [218]. User confirmation of intended actions can mitigate these risks, but cannot replace authorization and device- or robot-level safety checks.

Teaching robots. A distinct recent line of work employs smart glasses in teaching robots: human egocentric videos of environmental manipulations are used as training data for robots [170], [102], including for training advanced vision–language–action models [267]. Multiple systems that aim to produce human egocentric data for robotic learning are currently marketed to enterprise customers; many of them are built specifically for this purpose, to capture information that is not commonly available from commercial smart glasses such as depth and tactile data. *Limitations:* While capturing egocentric data is straightforward, using it to train policies is currently an area of active research at the intersection of AI and robotics. While conventional robot demonstrations record a robot's actions and states, egocentric human recordings need to be transformed into representations compatible with robot actions; moreover, related work explicitly mentions inaccuracies in 3D hand-pose annotations, occlusions, depth ambiguity, and lack of grasp information among the data limitations relevant to robot learning.

VI. Challenges and Future Directions

We identify several directions as particularly important for realizing advanced capabilities of AR devices with AI. Building on our categorization in the previous sections, we call for (1) AI-based runtime orchestration to manage compute resources, (2) AI-integrated AR development frameworks for fast and reproducible development cycles, and (3) multimodal and multi-device AI for improved context understanding.

A. AI Glasses with AI-Based Runtime Orchestration

Context- and computing-aware orchestration framework. Smart glasses constitute a complex "system of systems" that must satisfy stringent constraints on form factor, power, and thermal management (requiring computing awareness), while also supporting the diverse and complex requirements of real-world AI agent tasks (requiring context awareness). Previous approaches have relied on optimizing fixed system components [139], fixed computing architectures [89], or uniform task types [214]. In the new era of agentic AI, real-world tasks handled by smart glasses will become more complex and subject to multiple constraints, including context demands (e.g., whether a task only requires recognition of the current scene or long-context inputs incorporating historical context that exceed the processing capacity of on-device and edge computing), timeliness requirements (e.g., timely monitoring and correction of wound-care procedures versus deferrable meeting summaries), data-sharing constraints (e.g., local-only processing versus public sharing), and execution dependencies (e.g., guiding a user's actions in the physical world versus submitting a report in the digital world). These requirements of real-world tasks also significantly affect the selection of appropriate communication and computing architectures, including on-device, edge, and cloud computing, and the allocation of available computing resources, such as how the overall system power budget is allocated to display and sensing operations. This creates a strong need for intelligent task orchestration and calls for a future agentic system (shown in Fig. 8) that spans the distributed computing architecture of smart glasses and equips its agent orchestrator (i.e., the component responsible for managing all distributed subagents) with computing awareness and context awareness.

Silent and fast interaction. Smart glasses are expected to provide continuous egocentric perception beyond what smartphones offer, enabling longer-term memory and continuous AI involvement in users' daily lives. However, given the compromises required to fit a sophisticated computing system into the glasses' form factor with long battery life, maintaining such always-on perception remains a distant aspiration [18]. Therefore, appropriate mechanisms to trigger perception on smart glasses, e.g., audio cues, have become an important research direction [182], [153]. However, just as "Hey Siri" is rarely used anymore, people often prefer to simply tap rather than say a socially awkward wake-up phrase in public. A truly effective triggering and interaction mechanism should be silent and thus unobtrusive. Moving beyond straightforward voice commands, enabling fast and fine-grained interaction control on smart glasses, comparable to touchscreen typing on smartphones, requires advances in sensing and on-device computing capabilities to rapidly and accurately recognize users' intentions. Neither the Meta Neural Band nor smart rings can fully meet the demands for fast, fine-grained input. Future possibilities for rapidly reading users' intentions may involve exploring different interaction modalities, including the recognition of silent articulatory movements, mid-air writing, and brain–computer interfaces.

B. AI-Integrated AR Development Beneath the Hood

Compatible frameworks for AR-AI development. The software beneath the hood of AR systems falls at different levels of AR development, ranging from foundation models to support basic AR functions, to generative techniques to create 3D content, to end-to-end agentic coding of the application. However, these efforts are largely fragmented and make it difficult for researchers and developers who wish to experiment with different AI models for AR, at any level. Contrast this to AI's virtuous development cycle, enabled by a synergistic ecosystem: common programming libraries (e.g., PyTorch), model repositories (e.g., Hugging Face, GitHub), and datasets (e.g., Arena AI for LLMs, ImageNet for object detection). This is lacking for AR. Current attempts to unify XR programming focus on different pieces of the 3-part ecosystem. For the programming interfaces and libraries, XR Blocks [127] and XARP Tools [16] provide high-level abstractions for programming AR applications with integrated AI (Section IV-A). Note that there are other popular programming frameworks for AR, such as WebXR, A-Frame, and OpenXR, but they do not explicitly integrate AI. For the model repositories, GitHub is a general platform

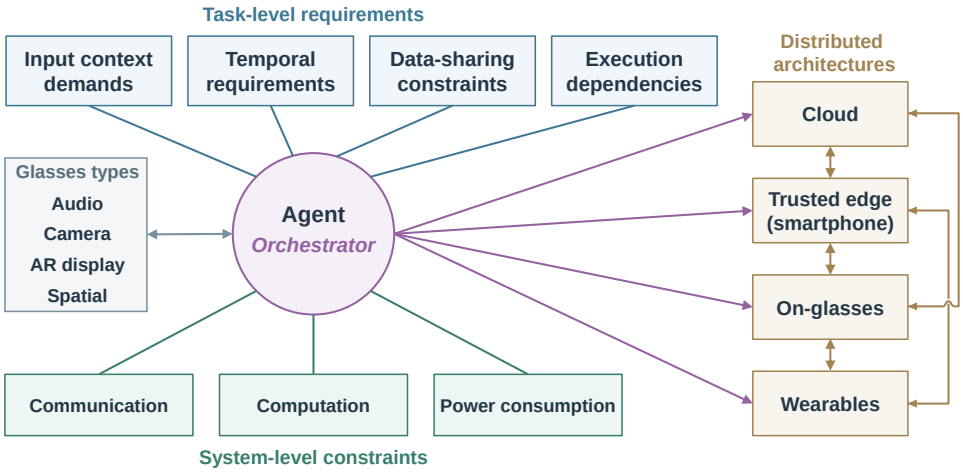

Fig. 8. Agent orchestration under task requirements and system constraints across four smart-glasses categories. The orchestrator selects wearables, on-glasses, trusted companion edge, e.g., smartphone, or cloud resources for task execution.

where AR code is often hosted, but is not AR-specific and does not allow real-time demos in a web browser, as Hugging Face does. Repositories specific to AR focus on narrow aspects of AR, like Open4D [221] that hosts 3D representations and the tools to manipulate them. Finally, the datasets piece of the ecosystem (Section IV-D) has some availability, drawn from the computer vision, robotics, and AR communities.

Instead, imagine an AR model repository in the future (e.g., a "Hugging Face for AR"), which allows individual components in an AR application to be swapped out as easily as object detection models in a Python script can be today. A developer could use AI to code not only the UI and interaction logic of the AR application, but also the underlying foundation models. An orchestrator agent (perhaps sharing responsibilities with the resource orchestrator in Section VI-A) could search the AR-Hugging Face for available and suitable AR foundation models based on the usage context, configure them to be deployed on the desired AR hardware, and instantiate the UI. We will elaborate on this vision in the unified case study below (Section VI-D).

Potential for AR slop. There are already major concerns about the use of AI in software development leading to code that is buggy or difficult to maintain. The rise of AI tools for AR code development (Section IV-A) leads to similar concerns, with the added risk that AR devices are not just stationary digital screens, but are physically strapped to users' heads as they move around. Unintentional or malicious AR code can cause safety or perceptual issues, and defenses are being developed (e.g., DARPA's Intrinsic Cognitive Security program). The rise of vibe-coded apps in traditional app stores [236] combined with the availability of AR vibe coding tools, like Unity AI or Google XR Blocks, hints at the potential for "AR slop" filling AR app stores, with potentially negative downstream effects on users.

C. User-facing Applications with Multimodal and Multi-device AI

Robust egocentric perception: A much-talked-about, much-desired view of AI smart glasses is of an all-seeing, all-knowing entity that perfectly captures the information humans capture – *"AI that sees what you see and hears what you hear*" [164] – and knows humans better than they know themselves. Yet the gap between this vision and what egocentric AI can deliver in practice is vast. AI-based visual and auditory perception is far from perfect [90], [245]; moreover, video and auditory streams available to glasses do not capture many phenomena that factor into human judgment, including the weight, texture, and pliability of objects humans are working with, and tactile, thermal, and olfactory cues from the broader environment. These limitations are particularly striking in applications where egocentric AI is expected to "read the mind" of its wearer, since human needs and wants are complex, ambiguous, and often cannot be readily discerned from observable behavior and measured biosignals. Extensive across-the-stack work on additional sensing modalities (see Section III-B) is needed to bring the sensing state available to AI for reasoning closer to that available to humans. It is also imperative to complement the abundant research on better AI models with research on *how to design systems and applications that operate reliably with*

imperfect AI. Environment shaping, i.e., modifying the physical environment to facilitate machine perception and reasoning [207], [237], and partial environmental instrumentation may offer practical paths toward improving the reliability of applications such as procedural guidance (see Section V-B), where the task-relevant portion of the physical environment is relatively constrained and amenable to modification. Another promising direction is to create systems that detect their own mistakes and allow users to correct AI decisions or trigger AI re-planning at key points in their tasks.

Multi-device AI-glasses-inclusive ecosystems: Partially due to the technical sophistication and the central role of eyewear, it may be tempting to envision the future where the glasses are the only device the user needs. However, it is becoming increasingly clear that this vision is not likely to be realized in practice. A substantial body of work on *hybrid user interfaces* has explored the complementary use of AR displays and conventional devices for interaction [82]. For capturing a user's control inputs, smart glasses already actively rely on wristbands, rings, and phones; for sensing the environment, external and removable cameras are increasingly proposed (see Section III-B); for computing, external pucks and phones are currently a default architecture (see Section III-A). Additionally, entire classes of promising applications, such as using smart glasses for interfacing with robots (see Section V-C), *inherently require collaboration between multiple devices* and involve multiple types of AI running simultaneously across them. We believe that building robust, easy-to-deploy, *multi-device ecosystems* covering sensing, compute, interaction and actuation is the path to enabling the next stage of AI transformation of AR. This calls for re-envisioning of platforms, applications, and AI models to include multiple devices in a flexible and adaptable manner, bring together heterogeneous data streams, and examine users' behavior, preferences, and performance in realistic multi-device settings.

D. Unified Futuristic Use Case

We envision a future use case given several of the AI transformations described above: an AI agent for computing resource orchestration, combined with a unified AR-AI development cycle that works seamlessly within a multi-device ecosystem. Suppose a user wakes up in the morning and prepares to go to work as a fruit harvester. She picks up her AI glasses, which automatically detect the context of the day, time, and weather, and know that it's a work day. Since the weather is foggy, to improve visibility, the glasses configure an egocentric computer vision model that uses an onboard infrared camera to see through the fog. At work, she collaborates with a team of robotic pickers to help reach fruits that are high overhead or low to the ground, visualizing their intent through the AI glasses and using silent and natural hand gestures to communicate with the robot team. As the fog lifts throughout the day, an agentic orchestrator on the glasses shifts processing to the RGB camera, turning off the IR camera to save energy.

On the weekend, the agent on the glasses determines from the user's email inbox that she will be going camping. Since the glasses lack cellular connectivity in the wilderness, onboard operations are key. The agent orchestrator automatically downloads appropriate models for task guidance (starting a campfire and pitching a tent). At the same time, the orchestrator extends the glasses' battery life to last for the entire 2-day trip by turning off full spatial understanding on the display, relying instead on lightweight object detection models with only audio feedback to the user. The glasses also connect to the user's smartphone camera for a non-egocentric view as needed. Realizing this futuristic use case at work and at play requires advances in AI-enabled AR in terms of resource and software orchestration, context understanding, and multi-device operation.

VII. Conclusions

Commercial AI glasses are here, in a multitude of form factors and with a myriad of intended use cases, which increasingly aim to apply AI to more and more sophisticated scene, process, and user understanding tasks, at an increasingly low latency. Our survey showcases where AI-enabled AR technology currently stands, via a close look at the state of both platforms and applications; we highlight areas of exceptional promise, such as the potential to solve important challenges in accessibility and guiding and verifying the execution of complex physical procedures, and also important shortcomings, such as unreliable AI perception and reasoning, high latency, and severe energy constraints that limit sustained, always-on operation. We present a roadmap for next-generation AI-enabled AR

applications and systems, emphasizing reliable operation with imperfect AI, context- and computing-aware runtime orchestration, and multimodal, multi-device ecosystems.

Our survey highlights that the AI transformation of AR does not stop with enriched application capabilities alone. We note another key AI transformation of AR: the shift from pre-authored to auto-generated experiences across multiple elements of AR, including 3D content, application code, on-demand procedural guidance, and dynamically generated user interactions. We describe the rapidly growing body of work on AI-enabled AR development, including frameworks, modules, and datasets, and present our vision for unifying this fragmented ecosystem through common programming libraries, model repositories, and datasets. We paint a picture of a future in which a user's task becomes the specification for an AR experience, with AI agents selecting suitable foundation models based on the usage context, configuring them for the available hardware, and instantiating the user interface.

REFERENCES

[1] AHMED, N., BANERJEE, R., GHOSE, A., AND SINHARAY, A. Feasibility analysis for estimation of blood pressure and heart rate using a smart eye wear. In *Proc. ACM WearSys* (2015).

[2] ALHARBI, R., SEN, S., NG, A., ALSHURAFA, N., AND HESTER, J. ActiSight: wearer foreground extraction using a practical RGB-thermal wearable. In *Proc. IEEE PerCom* (2022).

[3] ARORA, P., KIMMEL, E., HUANG, K., KWOK, T., SONG, Y., VEMPALA, S., LIN, G., CAKMAKCI, O., AND STARNER, T. Positioning monocular optical see through head worn displays in glasses for everyday wear. *arXiv preprint arXiv:2505.09047* (2025).

[4] ASISH, S. M., KARKI, B. B., KC, B., KOLAHCHI, N., AND SUTRADHAR, S. Synthesizing six years of AR/VR research: A systematic review of machine and deep learning applications. In *Proc. IEEE VR* (2025).

[5] BANSAL, S., ARORA, C., AND JAWAHAR, C. My view is the best view: Procedure learning from egocentric videos. In *European Conference on Computer Vision (ECCV)* (2022).

[6] BEASLEY, M. Google encourages Glass users to 'explore the world' and avoid being a 'Glasshole'. *https://9to5google.com/2014/02/18/google-encourages-glass-users-to-explore-the-world-and-avoid-being-a-glasshole/* (2014).

[7] BELL, K. Ray-Ban Meta optics review: A better fit comes with a higher price. https://www.engadget.com/2199194/ray-ban-meta-optics-review/, 2026.

[8] BHATTACHARYYA, P., MITTON, J., PAGE, R., MORGAN, O., MENZIES, B., HOMEWOOD, G., JACOBS, K., BAESSO, P., TRICKETT, D., MAIR, C., ET AL. Helios: An extremely low power event-based gesture recognition for always-on smart eyewear. In *Proc. ECCV* (2024).

[9] BLUM, H., MERCURIO, A., O'REILLY, J., ENGELBRACHT, T., DUSMANU, M., POLLEFEYS, M., AND BAUER, Z. CroCoDL: Cross-device collaborative dataset for localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), pp. 27424–27434.

[10] BOHUS, D., ANDRIST, S., SAW, N., PARADISO, A., CHAKRABORTY, I., AND RAD, M. SIGMA: An Open-Source Interactive System for Mixed-Reality Task Assistance Research, 2024.

[11] BONAZZI, P., MOOSMANN, J., CELIK, A., MAYER, P., AND MAGNO, M. OpenGlass: Ultra-low-power on-device AI eyewear with event-based vision. *arXiv preprint arXiv:2606.07431* (2026).

[12] BOROUSHAKI, T., LAM, M., DODDS, L., EID, A., AND ADIB, F. Augmenting augmented reality with Non-Line-of-Sight perception. In *Proc. USENIX NSDI* (2023).

[13] BOYD, L. E., RANGEL, A., TOMIMBANG, H., CONEJO-TOLEDO, A., PATEL, K., TENTORI, M., AND HAYES, G. R. SayWAT: Augmenting face-to-face conversations for adults with autism. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), pp. 4872–4883.

[14] BRUSIE, T., FIJAL, T., KELLER, A., LAUFF, C., BARKER, K., SCHWINCK, J., CALLAND, J. F., AND GUERLAIN, S. Usability evaluation of two smart glass systems. In *Proc. IEEE SIEDS* (2015).

[15] BURCH, B. F., KIM, N., MCPHERSON, R., ANOKYE, D., RYAN, A. S., ADDISON, O., ROY, N., HUANG, C.-M., GALIK, E., AND RESNICK, B. Smart glasses for older adults with cognitive impairment: Explanatory mixed methods study. *JMIR Aging 9*, 1 (2026), e81840.

[16] CAETANO, A., KUMARAN, R., JOU, K., HÖLLERER, T., AND SRA, M. XARP Tools: An Extended Reality Platform for Humans and AI Agents, 2025.

[17] CAI, R., JANAKA, N., KIM, H., CHEN, Y., ZHAO, S., HUANG, Y., AND HSU, D. AiGet: Transforming everyday moments into hidden knowledge discovery with ai assistance on smart glasses. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (2025), pp. 1–26.

[18] CAI, W., ZHANG, H., HUANG, Y., SUN, S., DENG, J., XU, S., SONG, J., AND ZHANG, Z. Color when it counts: Grayscale-guided online triggering for always-on streaming video sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2026).

[19] CAMPOS, C., ELVIRA, R., RODRÍGUEZ, J. J. G., MONTIEL, J. M., AND TARDÓS, J. D. ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM. *IEEE Transactions on Robotics 37*, 6 (2021), 1874–1890.

[20] CAO, Q., KHANNA, P., LANE, N. D., AND BALASUBRAMANIAN, A. MobiVQA: Efficient on-device visual question answering. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 6*, 2 (2022).

[21] CAUDELL, T. P., AND MIZELL, D. W. Augmented reality: An application of heads-up display technology to manual manufacturing processes. In *Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences* (1992), vol. 2, pp. 659–669.

[22] CHAKRAVARTHI, B., VERMA, A. A., DANIILIDIS, K., FERMULLER, C., AND YANG, Y. Recent event camera innovations: A survey. In *Proc. ECCV* (2024).

[23] CHANDIO, Y., BASHIR, N., AND ANWAR, F. M. HoloSet: A dataset for visual-inertial pose estimation in extended reality. In *ACM SenSys* (2022).

[24] CHANG, E., HUANG, Z., LIAO, Y., BHAVSAR, S. R., PARAM, A., STARK, T., AHMADYAN, A., YANG, X., WANG, J., ABDULLAH, A., NGUYEN, G., IYER, A., HALL, D., LI, E., MOON, S., SCHEFFER, N., AHMED, K., DAMAVANDI, B., WANGA, R., KUMAR, A., PATEL, R., AND DONG, X. L. WearVQA: A visual question answering benchmark for wearables in egocentric authentic real-world scenarios. NeurIPS 2025; https://arxiv.org/abs/2511.22154, 2025.

[25] CHAPLOT, D. S., GANDHI, D., GUPTA, S., GUPTA, A., AND SALAKHUTDINOV, R. Learning To Explore Using Active Neural SLAM. In *ICLR* (2020).

[26] CHEN, J., HU, J., WANG, G., JIANG, Z., ZHOU, T., CHEN, Z., AND LV, C. TaoAvatar: Real-Time Lifelike Full-Body Talking Avatars for Augmented Reality via 3D Gaussian Splatting. *CVPR* (2025).

[27] CHEN, T., WANG, Q., HE, R., AND GOLLAKOTA, S. Spatial speech translation: Translating across space with binaural hearables. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (2025), CHI '25, Association for Computing Machinery.

[28] CHEN, Y., XIAO, B., ZHAO, Y., XIA, H., XU, B., FANG, J., LI, Y., ZHENG, Y., WANG, X., XUE, Z., ZHANG, L., LI, H., AND ZHANG, N. LightMem-Ego: Your ai memory for everyday life. *arXiv preprint arXiv:2607.11487* (2026).

[29] CHEN, Z., CHIAPPALUPI, D., LIN, T., YANG, Y., BEYER, J., AND PFISTER, H. RL-LABEL: A Deep Reinforcement Learning Approach Intended for AR Label Placement in Dynamic Scenarios. *IEEE VIS* (2024).

[30] CHEN, Z., HU, W., WANG, J., ZHAO, S., AMOS, B., WU, G., HA, K., ELGAZZAR, K., PILLAI, P., KLATZKY, R., ET AL. An empirical study of latency in an emerging class of edge computing applications for wearable cognitive assistance. In *Proc. ACM/IEEE SEC* (2017).

[31] CHENG, Y. F., CARDEN, A., CHO, H., FIDALGO, C. G., WIELAND, J., AND LINDLBAUER, D. Augmented reality in-the-wild: Usage patterns and experiences of working with AR laptops in real-world settings. *IEEE Trans. Vis. Comput. Graph.* (2025).

[32] CHO, H., KOMAR, M., AND LINDLBAUER, D. RealityReplay: Detecting and Replaying Temporal Changes In Situ Using Mixed Reality. *ACM IMWUT* (2023).

[33] CITI RESEARCH. AI Glasses: The Next Fast-Growing Edge Device. Citigroup, Dec. 2025. Published December 18, 2025. Accessed November 12, 2026.

[34] CONG, L., SMERKOUS, D., WANG, X., YIN, D., ZHANG, Z., JIN, R., WANG, Y., GERASIMIUK, M., DINESH, R. K., SMERKOUS, A., SHI, L., ZHENG, J., LAM, I., WU, X., LIU, S., LI, P., ZHU, Y., ZHAO, N., PARAKH, M., SERRAO, S., MOHAMMAD, I. A., CHEN, C.-Y., XIE, X., CHEN, T., WEINSTEIN, D., BARBONE, G., CAGLAR, B., SUNWOO, J. B., LI, F., DENG, J., WU, J. C., WU, S., AND WANG, M. Labos: The AI-XR co-scientist that sees and works with humans. *arXiv preprint arXiv:2510.14861* (2025).

[35] COSTA, G. D. M., PETRY, M. R., AND MOREIRA, A. P. Augmented Reality for Human–Robot Collaboration and Cooperation in Industrial Applications: A Systematic Literature Review. *Sensors 22*, 7 (2022), 2725.

[36] DANG, R., YUAN, Y., ZHANG, W., XIN, Y., ZHANG, B., LI, L., WANG, L., ZENG, Q., LI, X., AND BING, L. ECBench: Can Multi-modal Foundation Models Understand the Egocentric World? A Holistic Embodied Cognition Benchmark. *CVPR* (2025).

[37] DE LA TORRE, F., FANG, C. M., HUANG, H., BANBURSKI-FAHEY, A., AMORES, J., AND LANIER, J. LLMR: Real-time Prompting of Interactive Worlds using Large Language Models. In *ACM CHI* (2024).

[38] DENNING, T., DEHLAWI, Z., AND KOHNO, T. In situ with bystanders of augmented reality glasses: Perspectives on recording and privacy-mediating technologies. In *Proc. ACM CHI* (2014).

[39] DING, Y., YANG, Q., LI, Y., YANG, Z., WANG, Z., LIANG, H., AND WU, S.-T. Waveguide-based augmented reality displays: Perspectives and challenges. *eLight 3* (2023), 24.

[40] DOGAN, M. D., GONZALEZ, E. J., AHUJA, K., DU, R., COLAÇO, A., LEE, J., GONZALEZ-FRANCO, M., AND KIM, D. Augmented object intelligence with XR-objects. In *ACM UIST* (2024).

[41] DUAN, L., ROTONDO, E., XIU, Y., EOM, S., CHEN, R., LI, C., HU, Y., AND GORLATOVA, M. Probing the AR scene analysis capabilities of large multimodal models: Toward reliable real-time assessment solutions. *IEEE Internet Computing, Special Issue on Immersive Computing 29*, 6 (November 2025).

[42] DUE, B. Challenges with Google Glass in social interaction. In *Proc. PIN-C* (2015).

[43] ERKOYUNCU, J., AND KHAN, S. Olfactory-based augmented reality support for industrial maintenance. *IEEE Access 8* (2020), 30306–30321.

[44] ESSILORLUXOTTICA. Q4/Full Year 2025 Results. *https://www.essilorluxottica.com/cap/content/283060/* (2026).

[45] FANG, C. M., SAMARADIVAKARA, Y., MAES, P., AND NANAYAKKARA, S. Mirai: A wearable proactive AI "inner-voice" for contextual nudging. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (2025), ACM, pp. 1–9.

[46] FATHI, A., REN, X., AND REHG, J. M. Learning to recognize objects in egocentric activities. In *Proc. IEEE CVPR* (2011).

[47] FEINER, S., MACINTYRE, B., HÖLLERER, T., AND WEBSTER, A. A touring machine: Prototyping 3D mobile augmented reality systems for exploring the urban environment. *Pers. Technol. 1*, 4 (1997), 208–217.

[48] FERSCHA, A., AND VOGL, S. Wearable displays for everyone! *IEEE Pervasive Comput. 9*, 1 (2009), 7–10.

[49] FLABOREA, A., D'AMELY DI MELENDUGNO, G. M., PLINI, L., SCOFANO, L., DE MATTEIS, E., FURNARI, A., FARINELLA, G. M., AND GALASSO, F. PREGO: Online mistake detection in PRocedural EGocentric videos. In *Proc. IEEE/CVF CVPR* (2024).

[50] FREY, S., LUCCHINI, M. A., KARTSCH, V., INGOLFSSON, T. M., BERNARDI, A. H., SEGESSENMANN, M., OSIELENIEC, J., BENATTI, S., BENINI, L., AND COSSETTINI, A. GAPses: Versatile smart glasses for comfortable and fully-dry acquisition and parallel ultra-low-power processing of EEG and EOG. *IEEE Trans. Biomed. Circuits Syst. 19*, 3 (2024), 616–628.

[51] GADRE, S. Y., ROSEN, E., CHIEN, G., PHILLIPS, E. K., TELLEX, S., AND KONIDARIS, G. End-user robot programming using mixed reality. In *2019 International Conference on Robotics and Automation (ICRA)* (2019), IEEE, pp. 2707–2713.

[52] GAMAGE, B., DO, T.-T., PRICE, N. S. C., LOWERY, A., AND MARRIOTT, K. What do blind and low-vision people really want from assistive smart devices? Comparison of the literature with a focus study. In *Proc. ACM SIGACCESS Conference on Computers and Accessibility* (2023), ASSETS '23.

[53] GAO, Y., WU, D., SONG, J., ZHANG, X., HOU, B., LIU, H., LIAO, J., AND ZHOU, L. A wearable obstacle avoidance device for visually impaired individuals with cross-modal learning. *Nat. Commun. 16*, 1 (2025), 2857.

[54] GARDNER, M.-A., SUNKAVALLI, K., YUMER, E., SHEN, X., GAMBARETTO, E., GAGNÉ, C., AND LALONDE, J.-F. Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090* (2017).

[55] GE, L., REN, Z., LI, Y., XUE, Z., WANG, Y., CAI, J., AND YUAN, J. 3D hand shape and pose estimation from a single RGB image. In *IEEE CVPR* (2019).

[56] GEMPERLE, F., KASABACH, C., STIVORIC, J., BAUER, M., AND MARTIN, R. Design for wearability. In *Proc. IEEE ISWC* (1998).

[57] GIUDICI, A., VERONESI, C., BARTOLI, P., CALIÒ, M., TELITI, A., GERVASONI, G., TROJANIELLO, D., AND ZAPPA, F. Fully-sensorized smart-eyewear platform for on-device machine learning. *arXiv preprint arXiv:2607.16222* (2026).

[58] GONZALEZ, R. E., COLLINS, J., BENNETT, C. L., AND AZENKOT, S. Investigating use cases of AI-powered scene description applications for blind and low vision people. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2024), CHI '24, Association for Computing Machinery.

[59] GOOGLE. Create your first activity for audio glasses and display glasses. *https://developer.android.com/develop/xr/jetpack-xr-sdk/glasses/first-activity* (2026).

[60] GOOGLE. Glass Explorer Edition: Platform Overview. *https://developers.google.com/glass/develop/overview* (2026).

[61] GOPAKUMAR, M., LEE, G.-Y., CHOI, S., CHAO, B., PENG, Y., KIM, J., AND WETZSTEIN, G. Full-colour 3D holographic augmented-reality displays with metasurface waveguides. *Nature 629*, 8013 (2024), 791–797.

[62] GRUBERT, J., LANGLOTZ, T., ZOLLMANN, S., AND REGENBRECHT, H. Towards pervasive augmented reality: Context-awareness in augmented reality. *IEEE Trans. Vis. Comput. Graph. 23*, 6 (2016), 1706–1724.

[63] GUO, Y., HOU, K., YAN, Z., CHEN, H., XING, G., AND JIANG, X. Sensor2scene: Foundation model-driven interactive realities. In *2024 IEEE International Workshop on Foundation Models for Cyber-Physical Systems and Internet of Things (FMSys)* (2024), IEEE, pp. 13–19.

[64] GUO, Y., ZHAO, J., DING, B., TAN, C., LING, W., TAN, Z., MIYAKI, J., DU, H., AND LU, S. Sign-to-911: Emergency call service for sign language users with assistive AR glasses. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking* (2023), pp. 1–15.

[65] GURRIN, C., SMEATON, A. F., AND DOHERTY, A. R. LifeLogging: Personal big data. *Foundations and Trends in Information Retrieval 8*, 1 (2014), 1–125.

[66] HA, K., CHEN, Z., HU, W., RICHTER, W., PILLAI, P., AND SATYANARAYANAN, M. Towards wearable cognitive assistance. In *Proc. ACM MobiSys'14* (June 2014).

[67] HAMILTON, M. A., BEUG, A., HAMILTON, H. J., AND NORTON, W. J. Augmented reality technology for people living with dementia and their care partners. In *Proc. Conference on Virtual and Augmented Reality Simulations* (2021).

[68] HAN, S., LIU, B., CABEZAS, R., TWIGG, C. D., ZHANG, P., PETKAU, J., YU, T.-H., TAI, C.-J., AKBAY, M., WANG, Z., NITZAN, A., DONG, G., YE, Y., TAO, L., WAN, C., AND WANG, R. MEgATrack: Monochrome Egocentric Articulated Hand-Tracking for Virtual Reality. *ACM Transactions on Graphics (TOG) 39*, 4 (2020), 87:1–87:13.

[69] HAN, T., LI, A., CHEN, Q., AND GAO, C. JaneEye: A 12-nm 2K-FPS 18.9-μJ/frame event-based eye tracking accelerator. In *Proc. IEEE/ACM ASP-DAC* (2026).

[70] HENDERSON, S., AND FEINER, S. Exploring the benefits of augmented reality documentation for maintenance and repair. *IEEE Transactions on Visualization and Computer Graphics 17*, 10 (2011), 1355–1368.

[71] HIGGINS, P., HAYES, C. J., LUKIN, S., AND MATUSZEK, C. Language and Gesture in Virtual Reality: Is a Gesture Worth 1000 Words? *AAAI* (2025).

[72] HIRZLE, T., MÜLLER, F., DRAXLER, F., SCHMITZ, M., KNIERIM, P., AND HORNBÆK, K. When XR and AI meet: A scoping review on extended reality and artificial intelligence. In *ACM CHI* (2023).

[73] HODGES, S., WILLIAMS, L., BERRY, E., IZADI, S., SRINIVASAN, J., BUTLER, A., SMYTH, G., KAPUR, N., AND WOODBERRY, K. SenseCam: A retrospective memory aid. In *Proceedings of the 8th International Conference on Ubiquitous Computing (UbiComp 2006)* (2006), Springer, pp. 177–193.

[74] HOLZ, C., AND WANG, E. J. Glabella: Continuously sensing blood pressure behavior using an unobtrusive wearable device. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 1*, 3 (2017), 1–23.

[75] HONG, F., GUZOV, V., KIM, H. J., YE, Y., NEWCOMBE, R., LIU, Z., AND MA, L. EgoLM: Multi-Modal Language Model of Egocentric Motions. *CVPR* (2025).

[76] HOQUE, R., HUANG, P., YOON, D. J., SIVAPURAPU, M., AND ZHANG, J. EgoDex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations (ICLR)* (2026).

[77] HSIEH, Y.-T., JYLHÄ, A., ORSO, V., GAMBERINI, L., AND JACUCCI, G. Designing a willing-to-use-in-public hand gestural interaction technique for smart glasses. In *Proc. ACM CHI* (2016).

[78] HU, E., LI, M., HONG, J., QIAN, X., OLWAL, A., KIM, D., HEO, S., AND DU, R. Thing2Reality: Enabling spontaneous creation of 3D objects from 2D content using generative AI in XR meetings. *ACM UIST* (2025).

[79] HU, X., ZHU, C., HSIA, S.-K., MA, D., JAIN, R., AND RAMANI, K. ARify: Leveraging narrated instructional videos to create augmented reality tutorials for procedural tasks. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems* (2026), ACM, pp. 1–23.

[80] HUANG, I. W., RAJBHANDARY, P., SHIU, S., HO, J. S., ZHU, J., WILSON, B., AND YE, G. Radar-based heart rate sensing on the smart glasses. *IEEE Microw. Wireless Technol. Lett. 34*, 6 (2024), 829–832.

[81] HUANG, Y., XU, J., PEI, B., YANG, L., ZHANG, M., HE, Y., CHEN, G., CHEN, X., WANG, Y., NIE, Z., LIU, J., LIN, D., FANG, F., LI, K., YUAN, C., QIAO, Y., WANG, Y., AND WANG, L. Vinci: A real-time smart assistant based on egocentric vision-language model for portable devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 9*, 3 (Sept. 2025), 1–33.

[82] HUBENSCHMID, S., SATKOWSKI, M., ZAGERMANN, J., MÉNDEZ, J., ELMQVIST, N., FEINER, S., FEUCHTNER, T., GRØNBÆK, J. E., LEE, B., SCHMALSTIEG, D., DACHSELT, R., AND REITERER, H. Hybrid user interfaces: Past, present, and future of complementary cross-device interaction in mixed reality. *IEEE Transactions on Visualization and Computer Graphics 32*, 7 (July 2026), 7380–7399.

[83] HUH, M., XUE, Z., DAS, U., ASHUTOSH, K., GRAUMAN, K., AND PAVEL, A. Vid2Coach: Transforming how-to videos into task assistants. In *Proc. ACM UIST* (2025).

[84] HUZAIFA, M., DESAI, R., GRAYSON, S., JIANG, X., JING, Y., LEE, J., LU, F., PANG, Y., RAVICHANDRAN, J., SINCLAIR, F., ET AL. Exploring Extended Reality with ILLIXR: A New Playground for Architecture Research, 2020.

[85] ICHTER, B., BROHAN, A., CHEBOTAR, Y., FINN, C., HAUSMAN, K., HERZOG, A., HO, D., IBARZ, J., IRPAN, A., JANG, E., JULIAN, R., KALASHNIKOV, D., LEVINE, S., LU, Y., PARADA, C., RAO, K., SERMANET, P., TOSHEV, A. T., VANHOUCKE, V., XIA, F., XIAO, T., XU, P., YAN, M., BROWN, N., AHN, M., CORTES, O., SIEVERS, N., TAN, C., XU, S., REYES, D., RETTINGHOUSE, J., QUIAMBAO, J., PASTOR, P., LUU, L., LEE, K.-H., KUANG, Y., JESMONTH, S., JOSHI, N. J., JEFFREY, K., RUANO, R. J., HSU, J., GOPALAKRISHNAN, K., DAVID, B., ZENG, A., AND FU, C. K. Do as I can, not as I say: Grounding language in robotic affordances. In *Proc. Conference on Robot Learning* (2023), vol. 205, pp. 287–318.

[86] IGLESIUS, E., KOBAYASHI, M., AND URANISHI, Y. MRPoS: Mixed reality-based robot navigation interface using spatial pointing and speech with large language model, 2026. Accepted to the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2026).

[87] IKEDA, B., GRAMOPADHYE, M., NEKERVIS, L., AND SZAFIR, D. MARCER: Multimodal augmented reality for composing and executing robot tasks. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (2025), IEEE, pp. 529–539.

[88] IMTIAZ, M. H., HOSSAIN, D., SENYUREK, V. Y., BELSARE, P., TIFFANY, S., AND SAZONOV, E. Wearable egocentric camera as a monitoring tool of free-living cigarette smoking: A feasibility study. *Nicotine & Tobacco Research 22*, 10 (2020), 1883–1890.

[89] IYENGAR, R., DONG, Q., NGUYEN, C., PILLAI, P., AND SATYANARAYANAN, M. Offload shaping for wearable cognitive assistance. *Electronics 13*, 20 (2024), 4083.

[90] JAIN, A., VATSA, M., AND SINGH, R. Words over pixels? rethinking vision in multimodal large language models. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)* (2025), International Joint Conferences on Artificial Intelligence Organization, pp. 10481–10489.

[91] JANG, C., BANG, K., CHAE, M., LEE, B., AND LANMAN, D. Waveguide holography for 3D augmented reality glasses. *Nat. Commun. 15*, 1 (2024), 66.

[92] JEELANI, M., AND PRIETO, M. Using the wearable acceptability range (WEAR) scale to rate social acceptability of mixed reality and AI enabled head-mounted wearables. In *Proc. IHIET* (2025).

[93] JENNINGS, N., WANG, H., LI, I., SMITH, J., AND HARTMANN, B. What's the game, then? Opportunities and challenges for runtime behavior generation. In *ACM UIST* (2024).

[94] JIANG, J., BHATNAGAR, B. L., YANG, N., MA, L., STARKE, S., KIPS, R., BERTSCH, N., HOLZ, C., AND BOGO, F. EgoExoMoCap: Distributed ego-exo human motion capture. *arXiv preprint arXiv:2607.15868* (2026).

[95] JIANG, J., AND LI, B. PartInteractor: Intent-driven part-aware 3D authoring for continuous co-creation in XR. In *ACM UIST* (2026).

[96] JIANG, Z., YUAN, X., QU, H., LIN, S., LIU, K., FAN, W., AND LI, Q. SUPERGLASSES: Benchmarking vision language models as intelligent agents for AI smart glasses. CVPR 2026 Findings; https://arxiv.org/abs/2602.22683, 2026.

[97] JIANG, Z., ZHAO, R., LIN, L., YU, Y., CHEN, H., ZHANG, X., XU, X., WANG, Y., MA, X., AND NGAI, E. C. H. DietGlance: Dietary monitoring and personalized analysis at a glance with knowledge-empowered AI assistant. *ACM Transactions on Computing for Healthcare 7*, 2 (2026), 1–46.

[98] JOCHER, G., AND QIU, J. Ultralytics yolo, 2026.

[99] JUN, H., AND NICHOL, A. Shap-E: Generating Conditional 3D Implicit Functions, 2023.

[100] KAIFOSH, P., AND REARDON, T. R. A generic non-invasive neuromotor interface for human-computer interaction. *Nature 645*, 8081 (2025), 702–711.

[101] KANELLOPOULOS, I., SIMOU, I., AND KOMNINOS, A. Evaluating the use of augmented reality and 2D interfaces for diagnosing IoT-based habitable smart spaces by non-expert users. *Proceedings of the ACM on Human-Computer Interaction 10*, 4 (2026), 1–28.

[102] KAREER, S., PATEL, D., PUNAMIYA, R., MATHUR, P., CHENG, S., WANG, C., HOFFMAN, J., AND XU, D. EgoMimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation (ICRA)* (2025), IEEE, pp. 13226–13233.

[103] KASAI, I., TANIJIRI, Y., ENDO, T., AND UEDA, H. A forgettable near eye display. In *Proc. IEEE ISWC* (2000).

[104] KHALAF, M., GARCIA, M., CHADHA, M., AND DASARI, M. A first look at multimodal mobile application performance on XR and AI smart glasses. In *Proc. ACM MobiCom* (2026).

[105] KHAWAJA, Y., NABEEL, E., HUMAYUN, S., JAVED, E., KROMBHOLZ, K., ALIZAI, M. H., AND BHATTI, N. A. Now you see me, now you don't: Consent-driven privacy for smart glasses. In *Proc. IEEE PerCom* (2026).

[106] KHAWAJA, Y., REHMAN, S., PONTICELLO, A., BHARDWAJ, D., KROMBHOLZ, K., ALIZAI, M. H., AND BHATTI, N. A. See me if you can: A multi-layer protocol for bystander privacy with consent-based restoration. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI 2026)* (2026), ACM, pp. 1–19.

[107] KIM, J., LEE, C.-J., JUNG, H., YU, T. C., ZHANG, R., OAKLEY, I., AND ZHANG, C. WatchHand: Enabling continuous hand pose tracking on off-the-shelf smartwatches. In *Proc. ACM CHI* (2026).

[108] KIM, K., YANG, Y., KIM, S., YEO, W., LEE, Y., REN, M., AND HWANG, S. J. MA-EgoQA: Question answering over egocentric videos from multiple embodied agents. *arXiv preprint arXiv:2603.09827* (2026).

[109] KIM, L. The lessons of Google Glass: Aligning key benefits and sociability. In *Proc. HIMI* (2018).

[110] KIM, Y., JADEJA, D., PRADHAN, D., YANG, Y., AND KAUFMAN, A. E. SpeechLess: Micro-utterance with personalized spatial memory-aware assistant in everyday augmented reality. In *Proc. IEEE VR* (2026).

[111] KLEIN, G., AND MURRAY, D. Parallel tracking and mapping for small AR workspaces. In *IEEE ISMAR* (2007).

[112] KOHNO, T., KOLLIN, J., MOLNAR, D., AND ROESNER, F. Display leakage and transparent wearable displays: Investigation of risk, root causes, and defenses. Tech. rep., Microsoft Co., USA, Research, 2015.

[113] KOLBERG, J., PABST, M., BIENER, V., MORI, S., AND SCHMALSTIEG, D. Semantic scene graphs for creating a localization-ready internet of things. *IEEE Transactions on Visualization and Computer Graphics 32*, 8 (2026), 7517–7531.

[114] KORTUEM, G., SEGALL, Z., AND BAUER, M. Context-aware, adaptive wearable computers as remote interfaces to "intelligent" environments. In *Proceedings of the 2nd International Symposium on Wearable Computers (ISWC)* (1998), IEEE Computer Society, pp. 58–65.

[115] KOSMYNA, N., MORRIS, C., SARAWGI, U., NGUYEN, T., AND MAES, P. AttentivU: A wearable pair of EEG and EOG glasses for real-time physiological processing. In *Proc. IEEE BSN* (2019).

[116] KUDITHIPUDI, D., SCHUMAN, C., VINEYARD, C. M., PANDIT, T., MERKEL, C., KUBENDRAN, R., AIMONE, J. B., ORCHARD, G., MAYR, C., BENOSMAN, R., ET AL. Neuromorphic computing at scale. *Nature 637*, 8047 (2025), 801–812.

[117] KWON, H., NAIR, K., SEO, J., YIK, J., MOHAPATRA, D., ZHAN, D., SONG, J., CAPAK, P., ZHANG, P., VAJDA, P., BANBURY, C., MAZUMDER, M., LAI, L., SIRASAO, A., KRISHNA, T., KHAITAN, H., CHANDRA, V., AND REDDI, V. J. XRBench: An Extended Reality (XR) Machine Learning Benchmark Suite for the Metaverse. *MLSys* (2023).

[118] KWON, Y., LIU, L., FUCHS, H., HABERMANN, M., AND THEOBALT, C. DELIFFAS: Deformable Light Fields for Fast Avatar Synthesis. In *NeurIPS* (2023).

[119] LAI, Y., YUAN, S., LI, P., ZHANG, B., KIEFER, B., DENG, T., AND ZELL, A. FAM-HRI: Foundation-model assisted multimodal human-robot interaction combining gaze and speech. *IEEE Transactions on Automation Science and Engineering 23* (2026), 10521–10534.

[120] LAMMING, M., BROWN, P., CARTER, K., ELDRIDGE, M., FLYNN, M., LOUIE, G., ROBINSON, P., AND SELLEN, A. The design of a human memory prosthesis. *The Computer Journal 37*, 3 (1994), 153–163.

[121] LEE, D., CHOI, J., SHIN, D. K., KANG, S., AND LEE, K. Hierarchical policies from verbal and egocentric human signals for natural human-robot interaction, 2026.

[122] LEE, J., WANG, J., BROWN, E., CHU, L., RODRIGUEZ, S. S., AND FROEHLICH, J. E. GazePointAR: A context-aware multimodal voice assistant for pronoun disambiguation in wearable augmented reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI)* (2024), pp. 1–20.

[123] LEE, S., AND HUA, H. Effects of focal distance on near-field depth perception and accommodative response in a vari-focal optical see-through augmented reality display. *IEEE Trans. Vis. Comput. Graph. 31*, 9 (2024), 4695–4711.

[124] LEE, S.-P., LU, Z., ZHANG, Z., HOAI, M., AND ELHAMIFAR, E. Error detection in egocentric procedural task videos. In *IEEE CVPR* (June 2024), pp. 18655–18666.

[125] LEGENDRE, C., MA, W.-C., FYFFE, G., FLYNN, J., CHARBONNEL, L., BUSCH, J., AND DEBEVEC, P. DeepLight: Learning Illumination for Unconstrained Mobile Mixed Reality. In *IEEE CVPR* (2019).

[126] LI, C., WU, G., CHAN, G. Y.-Y., TURAKHIA, D. G., CASTELO, S., LI, D., WELCH, L., SILVA, C. T., AND QIAN, J. Satori: Towards proactive AR assistant with belief-desire-intention user modeling. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)* (2025), ACM.

[127] LI, D., DU, R., ET AL. XR Blocks: Accelerating Human-Centered AI + XR Innovation, 2025. Google XR Labs.

[128] LI, J., CHEN, Y., PIAO, D., PAN, P., YU, Y., WANG, D., YAN, H., YUE, L., WANG, S., CHEN, Y., HUANG, S., AND LIU, M. EgoProx: Evaluating MLLMs on Egocentric 3D Proximity Reasoning Across a Cognitive Hierarchy. *CVPR* (2026).

[129] LI, J. N., ZHANG, Z. J., AND MA, J. OmniQuery: Contextually augmenting captured multimodal memories to enable personal question answering. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI 2025)* (2025), ACM, pp. 1–20.

[130] LI, K., KIPS, R., ARCADU, F., SCHIKORA, M., HO, J., GRAF, C., CHEN, J., CAI, X., SIVAPALAN, G., ZAREI, A., ET AL. Head and body pose estimation using ultrasound sensing on smart glasses. In *Proc. ACM CHI EA* (2026).

[131] LI, K., ZHANG, R., CHEN, B., CHEN, S., YIN, S., MAHMUD, S., LIANG, Q., GUIMBRETIÈRE, F., AND ZHANG, C. GazeTrak: Exploring acoustic-based eye tracking on a glass frame. In *Proc. ACM MobiCom* (2024).

[132] LI, K., ZHANG, R., CHEN, S., CHEN, B., SAKASHITA, M., GUIMBRETIÈRE, F., AND ZHANG, C. EyeEcho: Continuous and low-power facial expression tracking on glasses. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI 2024)* (2024), ACM.

[133] LI, S., GUO, Y., YAO, J., LIU, Z., AND WANG, H. HomeBench: Evaluating LLMs in smart homes with valid and invalid instructions across single and multiple devices. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2025), Association for Computational Linguistics, pp. 12230–12250.

[134] LI, S., LIN, J., AND YE, S. Grounding generative verification in 3D content. *arXiv preprint arXiv:2610.22360* (Oct. 2026).

[135] LI, Y., NIU, J., MIAO, Z., GE, C., ZHOU, Y., HE, Q., DONG, X., DUAN, H., DING, S., QIAN, R., ZHANG, P., ZANG, Y., CAO, Y., HE, C., AND WANG, J. OVO-Bench: How Far is Your Video-LLMs from Real-World Online Video Understanding? *CVPR* (2025).

[136] LI, Z., LI, H., WANG, J., HUI, P., AND LIANG, H.-N. FlexiCamAR: Enhancing everyday camera interactions on AR glasses with a flexible additional viewpoint. In *Proc. IEEE VR* (2026).

[137] LIKAMWA, R., WANG, Z., CARROLL, A., LIN, F. X., AND ZHONG, L. Draining our Glass: An energy and heat characterization of Google Glass. In *Proc. ACM APSys* (2014).

[138] LIN, H., CHEN, S., LIEW, J. H., CHEN, D. Y., LI, Z., SHI, G., FENG, J., AND KANG, B. Depth anything 3: Recovering the visual space from any views. *arXiv:2511.10647* (2025).

[139] LIN, W., ZHAO, S., ROSS, I., MARSHALL, C., KONDGULI, S., AND ZHU, Y. LowPowAR: Power-constrained tone mapping for augmented reality. *arXiv preprint arXiv:2607.19509* (2026).

[140] LIU, F., CHEN, J., XU, K., LIU, Y., GUAN, H., LU, X., YANG, B., HANCKE, G., LIU, R., AND LAU, R. W. GuideMe: Multi-domain task guidance and intervention in streaming video. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2026).

[141] LIU, J. AI glasses are aiding cheating in exams. test-obsessed Asia is ground zero. CNN; https://edition.cnn.com/2026/06/26/asia/ai-glasses-cheating-exams-intl-hnk, 2026.

[142] LIU, S., ZENG, Z., REN, T., LI, F., ZHANG, H., YANG, J., JIANG, Q., LI, C., YANG, J., SU, H., ET AL. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Proc. ECCV* (2024).

[143] LIU, X., LEE, D., GONZALEZ, E. J., GONZALEZ-FRANCO, M., AND SUZUKI, R. VisionClaw: Always-on AI agents through smart glasses. *arXiv preprint arXiv:2604.03486* (2026).

[144] LIU, Y., AND KWON, H. Efficient Depth Estimation for Unstable Stereo Camera Systems on AR Glasses. *CVPR* (2025).

[145] LOMBARDI, S., SARAGIH, J., SIMON, T., AND SHEIKH, Y. Deep appearance models for face rendering. *ACM Transactions on Graphics (TOG) 37*, 4 (2018), 1–13.

[146] LU, X., CHEN, Y., CHEN, C., TAN, H., CHEN, B., XIE, Y., HU, R., TAN, G., WU, R., HU, Y., ET AL. BlueLM-V-3B: Algorithm and system co-design for multimodal large language models on mobile devices. In *Proc. IEEE/CVF CVPR* (2025).

[147] LUNDING, R., HUBENSCHMID, S., FEUCHTNER, T., AND GRØNBÆK, K. ARTHUR: authoring human–robot collaboration processes with augmented reality using hybrid user interfaces. *Virtual Reality 29*, 2 (2025), 73.

[148] LUNDING, R. S., LUNDING, M. S., FEUCHTNER, T., PETERSEN, M. G., GRØNBÆK, K., AND SUZUKI, R. RoboVisAR: Immersive authoring of condition-based AR robot visualisations. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (2024), Association for Computing Machinery, pp. 462–471.

[149] LV, Y., ZHANG, T., WANG, Z., SONG, Y., DING, H., HAN, J., AND WANG, F. mmEgoHand: Egocentric hand pose estimation and gesture recognition with head-mounted millimeter-wave radar and IMU. *arXiv preprint arXiv:2501.13805* (2025).

[150] MA, K., MAO, Y., LI, W., TAN, S., ZHONG, Y., CHEN, H., WANG, H., AND WANG, M. R4DSG: Relative 4d scene graph memory for object-centric question answering in long egocentric video. In *Proceedings of the 34th ACM International Conference on Multimedia (MM 2026)* (2026), ACM.

[151] MA, R., MORIMOTO, Y., HO, J. S., SHIU, S., AND ZHU, J. mmET: mmWave radar-based eye tracking on smart glasses. In *Proc. ACM SenSys* (2025).

[152] MAHMUD, S., AGARWAL, D., AJIT, A., LIANG, Q., VIRANDA, T., GUIMBRETIERE, F., AND ZHANG, C. MunchSonic: Tracking fine-grained dietary actions through active acoustic sensing on eyeglasses. In *Proc. ACM ISWC* (2024).

[153] MAHMUD, S., PARIKH, V., LIANG, Q., LI, K., ZHANG, R., AJIT, A., GUNDA, V., AGARWAL, D., GUIMBRETIÈRE, F., AND ZHANG, C. Actsonic: Recognizing everyday activities from inaudible acoustic wave around the body. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (2024).

[154] MAHROO, A., GRECI, L., MONDELLINI, M., AND SACCO, M. Assessment of a mixed reality smart home controller: HoloHome pilot study on healthy adults. *Virtual Reality 27*, 3 (2023), 2673–2690.

[155] MANIAR, N., CHAN, S. W. T., ZULFIKAR, W., REN, S., XU, C., AND MAES, P. MemPal: Leveraging multimodal ai and llms for voice-activated object retrieval in homes of older adults. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI 2025)* (2025), ACM, pp. 993–1015.

[156] MANN, S. Wearable computing: A first step toward personal imaging. *Computer 30*, 2 (1997), 25–32.

[157] MANZARDO, M., FRABONI, F., MORANDINI, S., GUALTIERI, L., VIDONI, R., AND PIETRANTONI, L. Dynamic velocity scaling for industrial collaborative robots: A gaze-driven approach. *Scientific Reports 16* (2026).

[158] MARURI, H. A. C., LOPEZ-MEYER, P., HUANG, J., BELTMAN, W. M., NACHMAN, L., AND LU, H. V-speech: Noise-robust speech capturing glasses using vibration sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 2*, 4 (2018), 1–23.

[159] MATSUHASHI, K., KANAMOTO, T., AND KUROKAWA, A. Thermal model and countermeasures for future smart glasses. *Sensors 20*, 5 (2020), 1446.

[160] MEDINA, R. J., BOTROS, S., GANDHI, P., CHOUDHURY, R., DAS, K., GHOBRIL, A. B., AND SHIELDS, C. L. Artificial intelligence smart eyeglasses for the detection and description of stationary objects. *JAMA Ophthalmology 144*, 8 (2026), 692–702.

[161] META. AI Building Blocks — Overview. https://developers.meta.com/vr/documentation/unity/unity-ai-building-blocks-overview/, 2026.

[162] META. Introducing Meta VR Glasses: A Cinema, Courtside Seat, and Workspace in Just 100 Grams. *https://about.fb.com/news/2026/09/introducing-meta-vr-glasses-3d-movies-immersive-live-sports-100-grams/* (2026).

[163] META PLATFORMS, INC. Create more natural movements using inside-out body tracking and generative legs. Meta Horizon OS Developers Blog, 2023. Dec. 20, 2023; https://developers.meta.com/vr/blog/inside-out-body-tracking-and-generative-legs/.

[164] META PLATFORMS, INC. Third Quarter 2023 Results – Prepared Remarks, October 2023. Q3 2023 earnings conference call prepared remarks, October 25, 2023.

[165] META PLATFORMS, INC. Agentic Tools: Agent Skills for Meta Quest/Horizon OS VR Development. https://github.com/meta-quest/agentic-tools, 2026.

[166] METZ, C. A.I. agents: They're fun. they're useful. but don't give them the credit card. *The New York Times* (2026).

[167] MILLERDURAI, C., AKADA, H., WANG, J., LUVIZON, D., PAGANI, A., STRICKER, D., THEOBALT, C., AND GOLYANIK, V. EventEgo3D++: 3D human motion capture from a head-mounted event camera. *Int. J. Comput. Vis. 133*, 9 (2025), 6432–6455.

[168] MOOSMANN, J., BONAZZI, P., LI, Y., BIAN, S., MAYER, P., BENINI, L., AND MAGNO, M. Ultra-efficient on-device object detection on AI-integrated smart glasses with TinyissimoYOLO. In *Proc. ECCV* (2024).

[169] MURAI, R., DEXHEIMER, E., AND DAVISON, A. J. MASt3R-SLAM: Real-Time Dense SLAM with 3D Reconstruction Priors. *IEEE CVPR* (2025).

[170] NAIR, S., RAJESWARAN, A., KUMAR, V., FINN, C., AND GUPTA, A. R3M: A universal visual representation for robot manipulation. In *Proceedings of the 6th Conference on Robot Learning (CoRL)* (2022), vol. 205 of *Proceedings of Machine Learning Research*, PMLR, pp. 892–909.

[171] NEWCOMBE, R. A., IZADI, S., HILLIGES, O., MOLYNEAUX, D., KIM, D., DAVISON, A. J., KOHLI, P., SHOTTON, J., HODGES, S., AND FITZGIBBON, A. KinectFusion: Real-time dense surface mapping and tracking. In *IEEE ISMAR* (2011).

[172] NGUYEN, A., AND ARDAYFIO, C. I-XRAY. 404 Media; https://www.404media.co/someone-put-facial-recognition-tech-onto-metas-smart-glasses-to-instantly-dox-strangers/, 2024.

[173] NICHOL, A., JUN, H., DHARIWAL, P., MISHKIN, P., AND CHEN, M. Point-E: A System for Generating 3D Point Clouds from Complex Prompts, 2022.

[174] NIE, J., LIU, Y., HU, Y., WANG, Y., XIA, S., PREINDL, M., AND JIANG, X. SPIDERS+: A light-weight, wireless, and low-cost glasses-based wearable platform for emotion sensing and bio-signal acquisition. *Pervasive Mob. Comput. 75* (2021), 101424.

[175] OMDIA. Omdia: Global AI Glasses Shipments More Than Doubled in 1H26, Hitting a Record 4.2 Million. Omdia Press Release, Sept. 2026. Published September 22, 2026. Accessed November 12, 2026.

[176] ORESKOVIC, A., MCBRIDE, S., AND NAYAK, M. Google Glass future clouded as some early believers lose faith. *https://www.marketscreener.com/quote/stock/GOOGLE-INC-9469/news/Google-Glass-future-clouded-as-some-early-believers-lose-faith-Insight-19395178/* (2014).

[177] OWEN, D. Subtitling your life. *The New Yorker* (April 2025).

[178] PADMANABHAN, A., STANKOSKI, S., PANCHEVSKI, F., ARMITAGE, R. M., COX, S., BAERT, C., SAZDOV, B., DARKOVSKA, M., KIPRIJANOVSKA, I., FATOORECHI, M., SOFRONIEVSKI, B., INDOVSKA, E., NIKODINOVSKA, A., KULEBANOVA, S., CLEAL, A., GJORESKI, M., SEVENOAKS, T., NDUKA, C., GJORESKI, H., AND YEOMANS, M. R. Validating OCOsense smart glasses in a three-week home-based study: Assessing detection of eating, food identification and the use of haptic feedback to aid behaviour modification. *Appetite 218* (2026), 108382.

[179] PALETTA, L., DINI, A., MURKO, C., YAHYANEJAD, S., AND AUGSDÖRFER, U. H. Estimation of situation awareness score and performance using eye and head gaze for human–robot collaboration. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications* (2019), Association for Computing Machinery.

[180] PAPI, S., POLÁK, P., MACHÁČEK, D., AND BOJAR, O. How "real" is your real-time simultaneous speech-to-text translation system? *Transactions of the Association for Computational Linguistics 13* (2025), 281–313.

[181] PAREDES, L., IPSITA, A., MESA, J. C., MARTINEZ GARRIDO, R. V., AND RAMANI, K. StretchAR: Exploiting touch and stretch as a method of interaction for smart glasses using wearable straps. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 6*, 3 (2022), 1–26.

[182] PARUCHURI, A., HERSEK, S., AGGARWAL, L., YANG, Q., LIU, X., KULSHRESTHA, A., COLACO, A., FUCHS, H., AND CHATTERJEE, I. Egotrigger: Toward audio-driven image capture for human memory enhancement in all-day energy-efficient smart glasses. *IEEE Transactions on Visualization and Computer Graphics* (2025). Presented at ISMAR 2025.

[183] PEDDI, R., ARYA, S., CHALLA, B., PALLAPOTHULA, L., VYAS, A., GOURIPEDDI, B., WANG, J., ZHANG, Q., KOMARAGIRI, V., RAGAN, E., RUOZZI, N., XIANG, Y., AND GOGATE, V. CaptainCook4D: A dataset for understanding errors in procedural activities. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track* (2024), vol. 37. arXiv:2312.14556.

[184] PEI, Y., HUANG, R., ZHA, M., WANG, G., WANG, P., KANG, Q., YANG, Y., AND SHEN, H. T. AttentionAR: AR adaptation and warning for real-world safety via attention modeling and MLLM reasoning. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST 2025)* (2025), ACM, pp. 1–19.

[185] PENG, Y.-H., HSI, M.-W., TAELE, P., LIN, T.-Y., LAI, P.-E., HSU, L., CHEN, T.-C., WU, T.-Y., CHEN, Y.-A., TANG, H.-H., AND CHEN, M. Y. SpeechBubbles: Enhancing captioning experiences for deaf and hard-of-hearing people in group conversations. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), pp. 1–10.

[186] PEREZ, S. Meta CTO explains why the smart glasses demos failed at Meta Connect — and it wasn't the Wi-Fi. TechCrunch, Sept. 2025.

[187] PETTENELLA, A., CRAFA, D. M., SPAGNOLI, J., PEZZOLI, C., PARACCHINI, M., DI GIACOMO, S., FIORINI, C., BYRNE, S., ONGARELLO, T., MERIGO, L., ET AL. Development of a low-power wearable eye tracker based on hidden photodetectors. In *Proc. ACM ETRA* (2025).

[188] PETTENELLA, A., PALMIERI, G., COSTANTINI, F. R., MELLONI, F., MERIGO, L., AND CARMINATI, M. Capacitive eye tracking: a first fully-embedded demonstrator. In *Proc. ACM ETRA* (2026).

[189] PIRSIAVASH, H., AND RAMANAN, D. Detecting activities of daily living in first-person camera views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012)* (2012), IEEE, pp. 2847–2854.

[190] PLINI, L., SCOFANO, L., DE MATTEIS, E., D'AMELY DI MELENDUGNO, G. M., FLABOREA, A., SANCHIETTI, A., FARINELLA, G. M., GALASSO, F., AND FURNARI, A. TI-PREGO: Chain of thought and in-context learning for online mistake detection in PRocedural EGocentric videos. *arXiv preprint arXiv:2411.02570* (2024).

[191] PLIZZARI, C., TONIONI, A., XIAN, Y., KULSHRESTHA, A., AND TOMBARI, F. Omnia de EgoTempo: Benchmarking Temporal Understanding of Multi-Modal LLMs in Egocentric Videos. *CVPR* (2025).

[192] POOLE, B., JAIN, A., BARRON, J. T., AND MILDENHALL, B. DreamFusion: Text-to-3D using 2D Diffusion, 2022.

[193] PRAKASH, S., BAHREMAND, A., NGUYEN, L. D., AND LIKAMWA, R. GLEAM: An Illumination Estimation Framework for Real-time Photorealistic Augmented Reality on Mobile Devices. In *ACM MobiSys* (2019).

[194] PROJECT ARIA, META REALITY LABS RESEARCH. Aria dataset explorer. https://explorer.projectaria.com/, 2026. Interactive tool for browsing, previewing, and downloading Project Aria public datasets (e.g., Aria Digital Twin, Nymeria).

[195] PRYTZ, E., NILSSON, S., AND JÖNSSON, A. The importance of eye-contact for collaboration in AR systems. In *Proc. IEEE ISMAR* (2010).

[196] QIAN, Y., YANG, Z., CHEN, S.-C., MA, Y., CHEN, Y.-C., CHEN, H.-S., LIN, C.-L., AND WU, S.-T. Power consumption of light engines for emerging augmented reality glasses: perspectives and challenges. *Adv. Photonics 7*, 3 (2025), 034001–034001.

[197] QIN, T., LI, P., AND SHEN, S. VINS-Mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics 34*, 4 (2018), 1004–1020.
[198] QIU, L., LI, P., LI, H., ZUO, Q., GU, X., DONG, Y., YUAN, W., PENG, R., ZHU, S., HAN, X., CHEN, G., AND DONG, Z. LHM++: An efficient large human reconstruction model for pose-free images to 3d. *arXiv preprint arXiv:2506.13766* (2025).
[199] QIU, S., LIU, H., ZHANG, Z., ZHU, Y., AND ZHU, S.-C. Human–robot interaction in a shared augmented reality workspace. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2020), IEEE, pp. 11413–11418.
[200] RAYNEO. RayNeo X3 Pro launch, 2025.
[201] RHODIN, H., ERSOY, I., AYGUN, S., PFEIFFER, C. J., VOLLMER, A. L., AND TODT, I. Comparison of augmented reality glasses for the assistive communication support of hearing loss. *Frontiers in Neurology 16* (2025), 1635699.
[202] ROESNER, F., MOLNAR, D., MOSHCHUK, A., KOHNO, T., AND WANG, H. J. World-driven access control for continuous sensing. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS 2014)* (2014), ACM, pp. 1169–1181.
[203] SAADATZI, M. N., TAFAZZOLI, F., WELCH, K. C., AND GRAHAM, J. H. EmotiGO: Bluetooth-enabled eyewear for unobtrusive physiology-based emotion recognition. In *Proc. IEEE CASE* (2016).
[204] SAMSUNG ELECTRONICS. [Interview] [Galaxy Unpacked July 2026] Intelligent Eyewear: The First Step Toward the Next Mobile AI Interface. Samsung Mobile Press, July 2026. Published July 31, 2026. Accessed November 12, 2026.
[205] SAVVA, M., KADIAN, A., MAKSYMETS, O., ZHAO, Y., WIJMANS, E., JAIN, B., STRAUB, J., LIU, J., KOLTUN, V., MALIK, J., PARIKH, D., AND BATRA, D. Habitat: A Platform for Embodied AI Research, 2019. ICCV 2019.
[206] SCARGILL, T., EOM, S., CHEN, Y., AND GORLATOVA, M. Ambient intelligence for next-generation AR. In *Handbook of the Metaverse*. Springer, 2026, pp. 3–53.
[207] SCARGILL, T., JANAMSETTY, R., FRONK, C., EOM, S., AND GORLATOVA, M. Environment texture optimization for mobile augmented reality. *Proc. ACM IMWUT 8*, 3 (2024).
[208] SCHACTER, D. L. The seven sins of memory: Insights from psychology and cognitive neuroscience. *American Psychologist 54*, 3 (1999), 182–203.
[209] SCHENKLUHN, M., KNIERIM, M. T., KISS, F., AND WEINHARDT, C. Connecting home: Human-centric setup automation in the augmented smart home. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024), Association for Computing Machinery.
[210] SEMPIONATTO, J. R., NAKAGAWA, T., PAVINATTO, A., MENSAH, S. T., IMANI, S., MERCIER, P., AND WANG, J. Eyeglasses based wireless electrolyte and metabolite sensor platform. *Lab Chip 17*, 10 (2017), 1834–1842.
[211] SHEN, T., BAHMANI, S., HE, K., SRINIVASAN, S. G., CAO, T., REN, J., LI, R., WANG, Z., SHARP, N., GOJCIC, Z., FIDLER, S., HUANG, J., LING, H., GAO, J., AND REN, X. Lyra 2.0: Explorable generative 3D worlds, 2026. NVIDIA; project page: https://research.nvidia.com/labs/sil/projects/lyra2/.
[212] SIM, I., CHOI, K., BAEK, Y., CHOI, J. H., JO, J., YEOM, J., KIM, B., CHO, Y., LEE, H., BANG, H., ET AL. Microdisplay technologies in augmented reality and virtual reality headsets. *Nat. Rev. Electr. Eng. 2*, 9 (2025), 634–650.
[213] SINGH, A., BHANUSHALI, M. A., LUO, J., LUO, G., AND PUNDLIK, S. Assisting the blind to reach daily objects using smart glasses. *Displays 93* (2026), 103390.
[214] SINGH, R., RAY, D., SMITH, C., AND ADVE, S. SemanticXR: Low power, real-time queryable semantic mapping with a device-cloud architecture. *IEEE Transactions on Visualization and Computer Graphics* (2026).
[215] SNAP INC. Introducing SPECS Augmented Reality Glasses. *https://newsroom.snap.com/introducing-specs-augmented-reality-glasses* (2026).
[216] SNAP INC. SnapML on Spectacles. *https://developers.snap.com/spectacles/about-spectacles-features/snapML* (2026).
[217] SOMANATH, G., AND KURZ, D. HDR Environment Map Estimation for Real-Time Augmented Reality, 2020.
[218] SON, Y., KIM, M., KIM, S., HAN, S., KIM, J., JANG, D., YU, Y., AND PARK, C. Y. Subtle risks, critical failures: A framework for diagnosing physical safety of LLMs for embodied decision making. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (2025), Association for Computational Linguistics, pp. 25692–25733.
[219] SONG, H., YOON, B., YANG, S., KANG, S., KIM, H., METZMACHER, H., AND WOO, W. VRGaussianAvatar: Integrating 3D Gaussian Avatars into VR. *IEEE VR* (2026).
[220] SONG, V. Let's talk about the battery in smart glasses. The Verge, Feb. 2025.
[221] SPATIAL INTELLIGENCE RESEARCH GROUP. Open4D: An open platform for time-varying 4D geometry data, 2026. SINRG Lab, Northeastern University; https://open4dfoundation.github.io/.
[222] SRIVASTAVA, T., KHANNA, P., PAN, S., NGUYEN, P., AND JAIN, S. MuteIt: Jaw motion based unvoiced command recognition using earable. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 6*, 3 (2022), 1–26.
[223] SRIVASTAVA, T., KHANNA, P., PAN, S., NGUYEN, P., AND JAIN, S. Unvoiced: Designing an LLM-assisted unvoiced user interface using earables. In *Proc. ACM SenSys* (2024).
[224] STEARNS, L., FINDLATER, L., AND FROEHLICH, J. E. Design of an augmented reality magnification aid for low vision users. In *Proc. ACM ASSETS* (2018).
[225] STOVER, D., AND BOWMAN, D. A. TAGGAR: General-purpose task guidance from natural language in augmented reality using vision-language models. In *Proceedings of the 2024 ACM Symposium on Spatial User Interaction (SUI '24)* (2024), ACM.
[226] SUKUMAR, V. Dense TOPS vs. sparse TOPS: What's the difference? *https://www.qualcomm.com/news/onq/2025/07/dense-tops-vs-sparse-tops-whats-the-difference* (2025).
[227] SUO, S., WONG, K., XU, J., TU, J., CUI, A., CASAS, S., AND URTASUN, R. MixSim: A Hierarchical Framework for Mixed Reality Traffic Simulation. In *IEEE CVPR* (2023).

[228] SUZUKI, R., KARIM, A., XIA, T., HEDAYATI, H., AND MARQUARDT, N. Augmented reality and robotics: A survey and taxonomy for AR-enhanced human-robot interaction and robotic interfaces. In *Proc. CHI* (2022).

[229] SZALAVÁRI, Z., SCHMALSTIEG, D., FUHRMANN, A., AND GERVAUTZ, M. "Studierstube": An environment for collaboration in augmented reality. *Virtual Real. 3*, 1 (1998), 37–48.

[230] TANG, J., CHEN, Z., CHEN, X., WANG, T., ZENG, G., AND LIU, Z. LGM: Large multi-view Gaussian model for high-resolution 3D content creation. In *European Conference on Computer Vision* (2024), Springer.

[231] TANG, Y., SITU, J., CUI, A. Y., WU, M., AND HUANG, Y. LLM integration in extended reality: A comprehensive review of current trends, challenges, and future perspectives. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (2025).

[232] TANG, Y., WU, M., SITU, J., LIU, J., CUI, A. Y., AND HUANG, Y. Empowered XR through generative AI: Balancing superpowers and risks. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)* (2026), ACM, pp. 1–23.

[233] TENCENT HUNYUAN3D TEAM. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3D assets generation, 2025.

[234] TORAL-LOPEZ, V., CRIADO, S., ROMERO, F. J., MORALES, D. P., CASTILLO, E., GARCÍA, A., TAHMASSEBI, A., AND MEYER-BAESE, A. Wearable biosignal acquisition system for decision aid. In *Proc. SPIE* (2018), vol. 10662.

[235] TOWERS, M., KWIATKOWSKI, A., BALIS, J., DE COLA, G., DELEU, T., GOULÃO, M., ANDREAS, K., KRIMMEL, M., KG, A., PEREZ-VICENTE, R., TERRY, J., PIERRÉ, A., SCHULHOFF, S., TAI, J. J., TAN, H., AND YOUNIS, O. G. Gymnasium: A standard interface for reinforcement learning environments. In *Advances in Neural Information Processing Systems* (2025), D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, Eds., vol. 38, Main Conference, Curran Associates, Inc.

[236] TREMAYNE-PENGELLY, A. Vibe coders are clogging up the App Store. *The Atlantic* (October 2026).

[237] TUNG, Y.-S., LUEBBERS, M. B., RONCONE, A., AND HAYES, B. Workspace optimization techniques to improve prediction of human motion during human-robot collaboration. In *Proc. ACM/IEEE HRI* (2024), ACM, pp. 743–751.

[238] TUSSA, Y., HEREDIA, A., AND ROY, N. Lessons learned from developing a privacy-preserving multimodal wearable for local voice-and-vision inference. *arXiv preprint arXiv:2511.11811* (2025).

[239] UNITY TECHNOLOGIES. Unity Skills: A Collection of Reusable Skills for AI Coding Agents. https://github.com/Unity-Technologies/skills, 2026.

[240] UNITY TECHNOLOGIES. Unity's AI Tools for Game Development. https://unity.com/features/ai, 2026.

[241] VEERABADRAN, V., XIAO, F., KAMRA, N., MATIAS, P., CHEN, J., DROOFF, C., ROADS, B. D., WILLIAMS, R. J., HENDERSON, E., ZHAO, X., CARLBERG, K., TIGHE, J., AND RIDGEWAY, K. Benchmarking egocentric multimodal goal inference for assistive wearable agents. In *Advances in Neural Information Processing Systems* (2025), vol. 38, pp. 28469–28503.

[242] WANG, J., DABRAL, R., LUVIZON, D., CAO, Z., LIU, L., BEELER, T., AND THEOBALT, C. Ego4o: Egocentric human motion capture and understanding from multi-modal input. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2025), pp. 22668–22679.

[243] WANG, J., SCHARSTEIN, D., BAPAT, A., BLACKBURN-MATZEN, K., YU, M., LEHMAN, J., ALSISAN, S., WANG, Y., TSAI, S., FRAHM, J.-M., HE, Z., VAJDA, P., COHEN, M. F., AND UYTTENDAELE, M. A Practical Stereo Depth System for Smart Glasses. *CVPR* (2023).

[244] WANG, L. W., KARI, M., AND ABTAHI, P. Explainable OOHRI: Communicating robot capabilities and limitations as augmented reality affordances. In *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction* (2026), Association for Computing Machinery, pp. 427–437.

[245] WANG, S., JIN, Z., TANG, C., LI, Q., LI, B., CHEN, C., HU, Y., YU, W., LI, Y., ZHUANG, J., YANG, Y., WANG, M., HAN, M., DING, Y., BAI, J., OUYANG, T., CHANG, S.-Y., CHEN, X., TIAN, X., ZHANG, J., LU, L., SUN, G., CHEN, Z., WU, J., ZHOU, B., WANG, Y., SAINATH, T., WU, Y., AND ZHANG, C. Towards general auditory intelligence for machine listening and speaking. *Nature Machine Intelligence 8* (2026), 1201–1220.

[246] WANG, X., KWON, T., RAD, M., PAN, B., CHAKRABORTY, I., ANDRIST, S., BOHUS, D., FENIELLO, A., TEKIN, B., FRUJERI, F. V., JOSHI, N., AND POLLEFEYS, M. HoloAssist: An egocentric human interaction dataset for interactive AI assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), pp. 20270–20281.

[247] WANG, Z., SHI, Y., WANG, Y., YAO, Y., YAN, K., WANG, Y., JI, L., XU, X., AND YU, C. G-VOILA: Gaze-facilitated information querying in daily scenarios. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 8*, 2 (2024), 78:1–78:33.

[248] WASHINGTON, P., VOSS, C., KLINE, A., HABER, N., DANIELS, J., FAZEL, A., DE, T., FEINSTEIN, C., WINOGRAD, T., AND WALL, D. P. SuperpowerGlass: A wearable aid for the at-home therapy of children with autism. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1*, 3 (2017), 1–22.

[249] WHITMIRE, E., STRASNICK, E., BOLDU, R., SODHI, R., GODWIN, N., NG, S., LEVI, A., KARLSON, A., TAN, R., FALLER, J., ET AL. TouchFusion: Multimodal wristband sensing for ubiquitous touch interactions. *arXiv preprint arXiv:2602.15011* (2026).

[250] WORLD LABS TEAM. Atlas: A world model for spatial intelligence. *World Labs Blog* (2026). https://www.worldlabs.ai/blog/atlas.

[251] XIAO, J., HUANG, N., QIU, H., TAO, Z., YANG, X., HONG, R., WANG, M., AND YAO, A. EgoBlind: Towards egocentric visual assistance for the blind. In *Advances in Neural Information Processing Systems (NeurIPS 2025), Datasets and Benchmarks Track* (2025).

[252] XIU, Y., SCARGILL, T., AND GORLATOVA, M. ViDDAR: Vision language model-based detrimental content detection for augmented reality. *IEEE Transactions on Visualization and Computer Graphics 31*, 05 (2025).

[253] XU, L., YANG, B., JIANG, S., LIU, K., HOU, K., FAN, Y., CHEN, H., YAN, Z., AND JIANG, X. Pro^{2}Assist: Continuous step-aware proactive assistance with multimodal egocentric perception for long-horizon procedural tasks. *arXiv preprint arXiv:2605.04227* (2026).

[254] XU, R., WANG, W., TANG, H., CHEN, X., WANG, X., CHU, F.-J., FEISZLI, M., AND LIANG, K. J. Multi-SpatialMLLM: Multi-frame spatial understanding with multi-modal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2026), pp. 31078–31088.

[255] YAN, Z., WU, Y., ZHANG, Y., AND CHEN, X. A. EmoGlass: An end-to-end AI-enabled wearable platform for enhancing self-awareness of emotional health. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022), pp. 1–19.

[256] YANG, B., GUO, Y., XU, L., YAN, Z., CHEN, H., XING, G., AND JIANG, X. SocialMind: LLM-based proactive AR social assistive system with human-like perception for in-situ live interactions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 9*, 1 (2025), 1–30.

[257] YANG, B., XU, L., ZENG, L., GUO, Y., JIANG, S., LU, W., LIU, K., LI, Y., JIANG, X., XING, G., AND YAN, Z. ProAgent: Harnessing On-Demand Sensory Contexts for Proactive LLM Agent Systems, 2025.

[258] YANG, B., XU, L., ZENG, L., LIU, K., JIANG, S., LU, W., CHEN, H., JIANG, X., XING, G., AND YAN, Z. ContextAgent: Context-aware proactive LLM agents with open-world sensory perceptions. In *Advances in Neural Information Processing Systems* (2025), vol. 38, pp. 185637–185671.

[259] YANG, C., WANG, Q., WANG, C., AKOTO-YEBOAH, A., CHENG, S., AMOAH, B., JIA, Y., CHEN, S., CANUMALLA, A., ZHAO, K., ET AL. Scentronic: An artificial intelligence–assisted olfactory interface for remote odor sharing and transmission. *Sci. Adv. 12*, 43 (2026), eaee9266.

[260] YANG, J., LIU, S., GUO, H., DONG, Y., ZHANG, X., ZHANG, S., WANG, P., ZHOU, Z., XIE, B., WANG, Z., OUYANG, B., LIN, Z., COMINELLI, M., CAI, Z., LI, B., ZHANG, Y., ZHANG, P., HONG, F., WIDMER, J., GRINGOLI, F., YANG, L., AND LIU, Z. EgoLife: Towards egocentric life assistant. In *Proc. IEEE/CVF CVPR* (2025).

[261] YANG, R., YU, Q., WU, Y., YAN, R., LI, B., CHENG, A.-C., ZOU, X., FANG, Y., CHENG, X., QIU, R.-Z., YIN, H., LIU, S., HAN, S., LU, Y., AND WANG, X. EgoVLA: Learning vision-language-action models from egocentric human videos. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2026).

[262] YANG, S., HUANG, Y., CAI, W., SUN, S., FANG, F., HE, Y., XIE, Y., DENG, J., ZHANG, H., SONG, J., ET AL. Egocentric co-pilot: Web-native smart-glasses agents for assistive egocentric AI. In *Proc. ACM WWW* (2026).

[263] YANG, Z., WANG, T., LIN, Y., CHEN, Y., ZENG, H., PEI, J., WANG, J., LIU, X., ZHOU, Y., ZHANG, J., ET AL. A vision chip with complementary pathways for open-world sensing. *Nature 629*, 8014 (2024), 1027–1033.

[264] YE, H., HU, T., AND GORLATOVA, M. No guide, no cheat: detecting smart glasses via AR optical signatures. In *Proc. ACM HotMobile* (2026).

[265] YE, M., DANELLJAN, M., YU, F., AND KE, L. Gaussian Grouping: Segment and Edit Anything in 3D Scenes. *ECCV* (2024).

[266] YOSHIDA, T., KURITA, S., NISHIMURA, T., AND MORI, S. Generating 6DoF Object Manipulation Trajectories from Action Description in Egocentric Vision. *CVPR 2025* (2025).

[267] YOSHIDA, T., KURITA, S., NISHIMURA, T., AND MORI, S. Developing vision-language-action model from egocentric videos. In *2026 IEEE International Conference on Robotics and Automation (ICRA)* (Vienna, Austria, 2026).

[268] YU, A., CHANG, Y., XIE, C., AND CARLONE, L. Pandora: Articulated 3D scene graphs from egocentric vision. *arXiv preprint arXiv:2603.28732* (2026).

[269] YU, J. Developing Meta's Orion AR Glasses. *https://www.infoq.com/presentations/meta-orion-ar-glasses/* (2025).

[270] YU, J., FERNIANY, W., GUTHRIE, B., PAREKH, S. G., AND PONCE, B. Lessons learned from Google Glass: telemedical spark or unfulfilled promise? *Surg. Innov. 23*, 2 (2016), 156–165.

[271] YU, S., SHU, L., YANG, A., FU, Y., SUNKARA, S., WANG, M., CHEN, J., BANSAL, M., AND GONG, B. Ego2web: A web agent benchmark grounded in egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2026), pp. 25633–25643.

[272] YUAN, X., WANG, Y., JIANG, Z., QU, H., DING, Y., LIN, S., XING, G., YANG, H., CAO, J., LI, Q., AND FAN, W. AI smart glasses for wearable intelligence: From egocentric sensing to agentic personalization. *arXiv preprint arXiv:2609.19793* (Sept. 2026).

[273] ZHANG, C., HAN, D., QIAO, Y., KIM, J. U., BAE, S.-H., LEE, S., AND HONG, C. S. Faster Segment Anything: Towards Lightweight SAM for Mobile Applications, 2023.

[274] ZHANG, J., CHEN, H., AND LIU, Y. From seeing to acting: Smart glasses as first-person intelligence platforms. *arXiv preprint arXiv:2608.24877* (Aug. 2026).

[275] ZHANG, L., CUI, K., CHEN, H., ROZNYATOVSKIY, V., ZHANG, J., GUO, A., SHIN, K. G., AND SUN, K. EventHOI: Event-driven resource-efficient hand-object interaction logging on smartglasses, 2026.

[276] ZHANG, R., AND AMFT, O. Monitoring chewing and eating in free-living using smart eyeglasses. *IEEE J. Biomed. Health Inform. 22*, 1 (2017), 23–32.

[277] ZHANG, R., LI, K., HAO, Y., WANG, Y., LAI, Z., GUIMBRETIÈRE, F., AND ZHANG, C. EchoSpeech: continuous silent speech recognition on minimally-obtrusive eyewear powered by acoustic sensing. In *Proc. ACM CHI* (2023).

[278] ZHANG, S., LI, S., AND LI, Q. Understood: Real-time communication support for adults with ADHD using mixed reality. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology* (2025), pp. 1–23.

[279] ZHANG, T., LAN, Z., XU, C., LI, Y., AND CHEN, Y. BLEselect: Gestural IoT device selection via bluetooth angle of arrival estimation from smart glasses. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 6*, 4 (2022), 1–28.

[280] ZHANG, X., VU, D. N. T., SHEN, C., QIN, Z., HONG, R., WANG, Y., SHI, Y., AND CHAN, J. VergeIO: Depth-aware eye interaction on glasses. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 10*, 3 (2026), 1–31.

[281] ZHANG, Y., DONG, X. L., LIN, Z., MADOTTO, A., KUMAR, A., DAMAVANDI, B., CHAI, J., AND MOON, S. Proactive assistant dialogue generation from streaming egocentric videos. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2025), Association for Computational Linguistics, pp. 12044–12068.

[282] ZHAO, A. Y., GUNTURU, A., DO, E. Y.-L., AND SUZUKI, R. Guided reality: Generating visually-enriched AR task guidance with LLMs and vision models. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)* (2025).

[283] ZHAO, N., ALDRICH, M., REINHART, C. F., AND PARADISO, J. A. A multidimensional continuous contextual lighting control system using Google Glass. In *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments (BuildSys)* (2015), Association for Computing Machinery, pp. 235–244.

[284] ZHOU, W., TAO, M., ZHAO, C., GUO, H., DONG, H., TANG, M., AND WANG, J. PhysVLM: Enabling visual language models to understand robotic physical reachability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), pp. 6940–6949.

[285] ZHU, C., HSIA, S.-K., HU, X., LIU, Z., SHI, J., AND RAMANI, K. agentAR: Creating augmented reality applications with tool-augmented LLM-based autonomous agents. In *Proc. ACM UIST* (2025).

[286] ZHU, M., SHANG, D., AND ZHANG, S. Q. ESCA: Enabling Seamless Codec Avatar Execution through Algorithm and Hardware Co-Optimization for Virtual Reality, 2025.

[287] ZHU, X. T., LIU, X., HARTEVELD, C., DESAI, S., AND MCGIVNEY, E. Conversational successes and breakdowns in everyday smart glasses use. Extended Abstracts of CHI 2026 (CHI EA '26); https://arxiv.org/abs/2602.22340, 2026.