Front. Comput. Sci., 2025, 19(12): 1912708
https://doi.org/10.1007/s11704-025-41123-8

REVIEW ARTICLE

# Towards spatial computing: recent advances in multimodal natural interaction for Extended Reality headsets

**Zhi-Min WANG[1], Mao-Hang RAO[1], Shang-Hua YE[1], Wei-Tao SONG[2], Feng LU (✉)[1]**

1 State Key Laboratory of VR Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China
2 School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China



**Abstract** With the widespread adoption of Extended Reality (XR) headsets, spatial computing technologies are gaining increasing attention. Spatial computing enables interaction with virtual elements through natural input methods such as eye tracking, hand gestures, and voice commands, thus placing natural human-computer interaction at its core. While previous surveys have reviewed conventional XR interaction techniques, recent advancements in natural interaction, particularly driven by artificial intelligence (AI) and large language models (LLMs), have introduced new paradigms and technologies. In this paper, we review research on multimodal natural interaction for wearable XR, focusing on papers published since 2022 in six top venues: ACM CHI, UIST, IMWUT (Ubicomp), IEEE VR, ISMAR, and TVCG. We classify and analyze these studies based on application scenarios, operation types, and interaction modalities. This analysis provides a structured framework for understanding how researchers are designing advanced natural interaction techniques in XR. Based on these findings, we discuss the challenges in natural interaction techniques and suggest potential directions for future research. This review provides valuable insights for researchers aiming to design natural and efficient interaction systems for XR, ultimately contributing to the advancement of spatial computing.

**Keywords** extended reality, multimodal, natural interaction, eye, hand, speech

## 1 Introduction

Extended Reality (XR), which includes Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR), merges the virtual and physical worlds to provide immersive experiences. In recent years, XR technologies have developed rapidly. This has led to the widespread adoption of headsets such as the Microsoft HoloLens 2 [1] and Meta Quest 3 [2], across various fields, including industrial maintenance, remote collaboration, online education and entertainment [3–5]. These developments highlight XR's potential for diverse applications and its promising market prospects.

In 2024, Apple introduced a new XR headset called the Vision Pro, which reignited public enthusiasm for XR and marked the beginning of a new era in spatial computing [6]. Spatial computing leverages advanced technologies to perceive and digitize the surrounding physical environment. It seamlessly integrates this environment with computer-generated virtual content, enabling natural interactions between humans and digital systems [7]. The advent of spatial computing is expected to bring a transformative redefinition to XR devices, often referred to as the "iPhone moment" for XR [8].

The core of spatial computing is human-computer interaction (HCI), with a particular focus on developing natural and intuitive interaction techniques [6]. Traditional methods, such as keyboards and handheld controllers, are inadequate for delivering the immersive experiences [9,10]. In light of this, research has increasingly focused on direct human input as interaction channels, including eye gaze, hand gestures, and voice commands [11–14]. Although these modalities provide more intuitive interfaces, each faces significant limitations. For example, gesture-based interaction can cause arm fatigue after prolonged use [15,16]. The performance of unimodal natural interactions requires further improvement.

To address these challenges, multimodal interaction has emerged as a promising solution by combining the strengths of individual modalities. Such systems integrate typically two modalities, e.g., Gesture + Speech [17,18], Gaze + Gesture [19,20], Gaze + Speech [21], and Gaze + Electroencephalography [22]. Each modality in these systems is assigned a specific task. For example, a user might select an object using the eye gaze and trigger an action with a hand gesture [9]. These approaches enhance the XR experience, providing users with a more natural and efficient way to interact with digital environments.

In recent years, researchers have conducted extensive

Received October 19, 2024; accepted April 17, 2025

E-mail: lufeng@buaa.edu.cn

Shang-Hua Ye participanted in this work as an intern at Beihang University.

reviews on various aspects of XR interaction techniques. These reviews have typically focused on specific areas such as AR environments [23,24], VR environments [25,26], or immersive environments involving handheld displays [1,27], highlighting the advantages, challenges, and emerging trends in each domain [1,23]. However, the rapid advancements in natural interaction, particularly driven by artificial intelligence (AI) and large language models (LLMs) [28–31], have introduced new interaction paradigms and technologies. Recently proposed new concepts such as cobodied/symbodied AI [32] have also brought new perspectives and challenges to wearable human-computer interaction. It underscores the need for updated reviews that synthesize and evaluate the latest developments in the field of XR interaction techniques.

In this paper, we aim to capture the latest trends by providing a comprehensive review of multimodal natural interaction techniques for wearable XR. Specifically, our objective is to determine and answer the following research questions (RQs):

1. What novel interaction paradigms and techniques have emerged over the past three years? (answered in Subsection 3.3)
2. What are the key evolving trends of multimodal natural interaction in XR over the past three years? (answered in Subsections 4.1 and 4.2)
3. How have recent advancements in AI and LLMs been leveraged to enhance natural interaction in XR environments? (answered in Subsection 4.3)

In order to answer these questions, we investigate papers published since 2022 across six top venues: ACM CHI, UIST, IMWUT, IEEE VR, ISMAR, and TVCG. We categorize the reviewed literature based on application context, operation types, and interaction modalities. Particularly, operation types are divided into seven categories, distinguishing between active and passive interactions, as shown in Fig. 1. Interaction modalities are discussed across nine distinct types. Additionally, we present statistical results on advanced natural interaction techniques. Based on these findings, we discuss the current challenges of natural interaction techniques and propose potential directions for future research. Our review offers valuable insights for researchers and promotes the further development of multimodal natural interaction for XR headsets.

The contributions of this review are threefold:

1. We provide a systematic review of the latest developments in multimodal interaction techniques for wearable XR, drawing from six top venues with papers published since 2022.
2. We categorize these papers based on application contexts, operation types, performance measures, and interaction modalities, and present statistical insights into advanced natural interaction techniques.
3. We identify the current challenges of natural interaction techniques and propose potential directions for future research to improve the effectiveness and usability of multimodal interactions in XR.

Several prior reviews have explored immersive interaction techniques, but with distinct scopes and focuses compared with our work. Hertel et al. [24] proposed a taxonomy for AR interaction techniques, focusing on task and modality dimensions based on works from 2016 to 2021. Spittle et al. [1] reviewed AR/VR interaction techniques from 2013 to 2020, focusing on display type, study type, input methods, and tasks. Pirker et al. [25] analyzed the educational applications of 360° videos and real VR from 2010 to 2020, including language learning, teacher education, history, and social

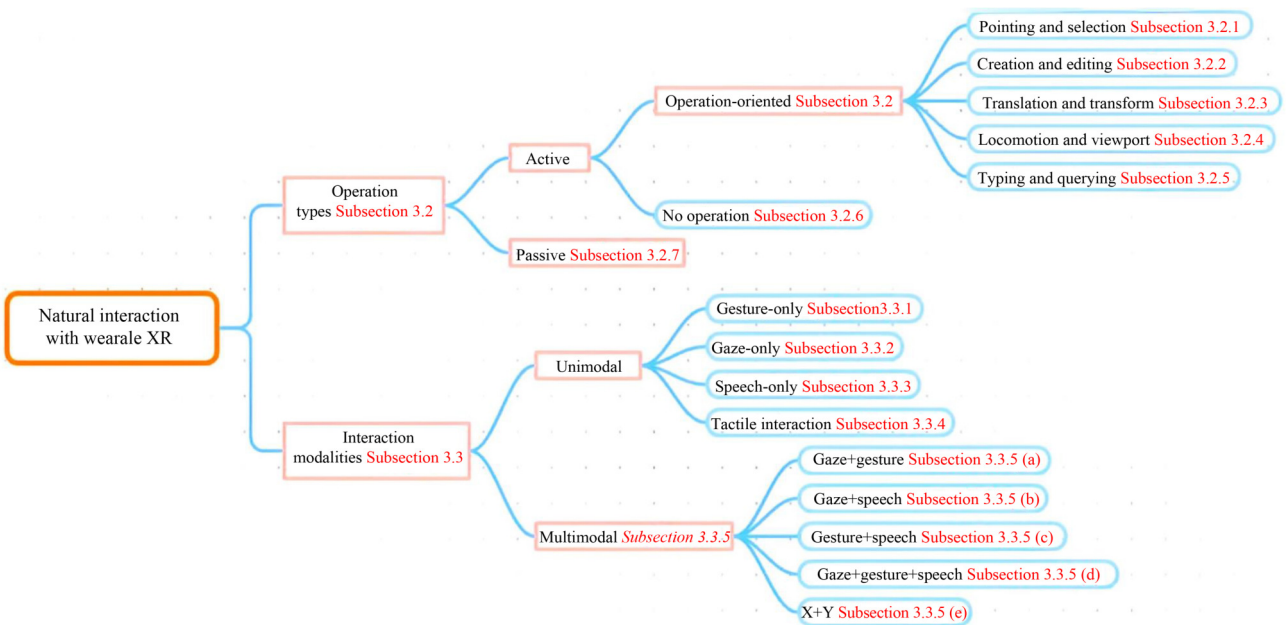



**Fig. 1** We classify the operation types into seven categories based on whether users actively input or passively receive feedback. Additionally, we divide the interaction modalities into nine categories, distinguishing between unimodal and multimodal natural interactions. (The overlap between Operation Types and Interaction Modalities is shown in Table 1)

studies, etc. Zhang et al. [27] focused on immersive visualization from 2014 to 2023, detailing multimodal perception and interaction techniques, particularly sensory modalities including vision, touch, and olfaction, while also analyzing collaborative analysis and hardware devices. Ghamandi et al. [33] reviewed 30 years of collaborative XR tasks (1993–2023) and proposed a taxonomy that classifies tasks based on actions (e.g., manipulation and navigation) and properties (e.g., temporal state and dependency). Their work focuses on understanding task structures and collaboration dynamics across the mixed reality spectrum. In contrast, our review focuses on recent developments since 2022 in multimodal natural interaction for wearable XR headsets, emphasizing the integration of AI and LLMs to enable novel interaction paradigms and providing a structured framework for advancing spatial computing.

The remainder of this paper is organized as follows. Section 2 outlines the methodology to select the reviewed literature and provides an statistic analysis of the selected papers. Section 3 presents a detailed review and analysis of natural interaction techniques in XR, categorized based on the criteria mentioned above. Section 4 discusses key challenges and offers recommendations for future research on natural interaction techniques in XR. Finally, Section 5 describes the limitations of this study, and Section 6 concludes the paper.

## 2 Survey methodology

We conduct a systematic review and analysis of natural interaction techniques in wearable XR. To this end, we survey relevant papers from the top conferences and journals in the field, focusing on recent publications. Additionally, we categorize and code these papers based on various aspects of natural interaction techniques in XR.

### 2.1 Selection criteria

This review focuses on the technologies and applications of these interaction techniques within XR. Relevant papers are selected based on the following criteria:

1. Involve wearable XR (including VR, AR, and MR). We focus on wearable XR because head-mounted displays (HMDs) offer superior mobility and support more natural interaction methods, such as gaze, gesture, and voice control, which ultimately enhance the sense of immersion.

2. Consider natural interaction techniques. We take both unimodal and multimodal interactions into consideration. Unimodal interactions include hand gesture, eye gaze, speech, and tactile input, while multimodal interactions combine these modalities. Among the unimodal interactions, hand-based ones allow precise operations and are most commonly used in daily life, such as pressing buttons and typing on a keyboard. With the widespread adoption of smartphones, voice interactions have also become progressively prevalent, with systems like Apple Siri, Microsoft Cortana, and Xiaomi XiaoAI enhancing input efficiency. In recent years, advances in eye-tracking technology have drawn increasing attention to gaze-based interaction. For tactile interactions, we focus on lightweight, wearable, and technologically advanced input devices.

3. Focus on top-tier conferences and journals. We concentrate on research published in leading venues such as ACM CHI, UIST, IMWUT (UbiComp), IEEE VR, ISMAR, and IEEE TVCG. These publications typically feature pioneering work, often marked by innovative contributions.

4. Focus on research since 2022[1)]. The release of XR devices, such as Microsoft HoloLens 2, Apple Vision Pro, HTC Vive, Meta Quest, and Pico series, has provided significant technical support to the field of XR research. We believe that recent studies offer valuable insights for researchers in this area.

### 2.2 Data collection

We search keywords in Google Scholar such as “virtual reality”, “augmented reality”, “extended reality”, “multimodal interaction”, “eye gaze”, “hand”, “speech”, and “tactile”. Article retrieval is primarily conducted following the standards outlined in Subsection 2.1. In addition, we identify some articles from other high-level journals and conferences that also meet criteria 1, 2, and 4 from Subsection 2.1. Although smaller in number, this portion of the literature serves as a valuable supplement to the collected data.

In summary, we gather a total of 104 research papers, which are summarized in Tables 1 and 2. They include 40 from CHI, 20 from IEEE VR, 14 from ISMAR, 14 from TVCG, 6 from UIST, 4 from IMWUT, 5 from other conferences and journals (IJHCS, ETRA, IUI, ISWC), and 1 from ArXiv.

### 2.3 Data analysis

The collected literature covers a range of interaction modalities and their combinations, with a focus on various application contexts, performance measures and operation types. These publications were classified and analyzed across several dimensions, as outlined in Section 3. In this section, we perform a statistical analysis of the reviewed literature from various perspectives. This high-level analysis explores the quantity and proportion of attention given to different categories of interaction research in recent years. These statistical results also support the taxonomy proposed in Section 3.

Based on whether humans act as initiators or receivers of interaction, the literature is categorized into Active interaction (84 papers) and Passive interaction (20 papers). In the Active interaction literature, the modalities include Gesture (24), Gaze (13), Speech (7), Tactile (8), and various combinations of these modalities, such as Gaze+Speech (12), Gaze+Gesture (4), Gesture+Speech (4), Gaze+Gesture+Speech (2), and other combinations (X+Y) (10). Notably, in some publications that examine multiple single or multimodal interaction schemes, the baseline methods are excluded from classification, focusing instead on the newly proposed interaction schemes. The proportion of different modalities in research since 2022 is shown in Fig. 2. Studies on Gesture and Gaze, including their combined modalities, are the most prevalent. A notable

[1)] Our final search was conducted on 16th October 2024.

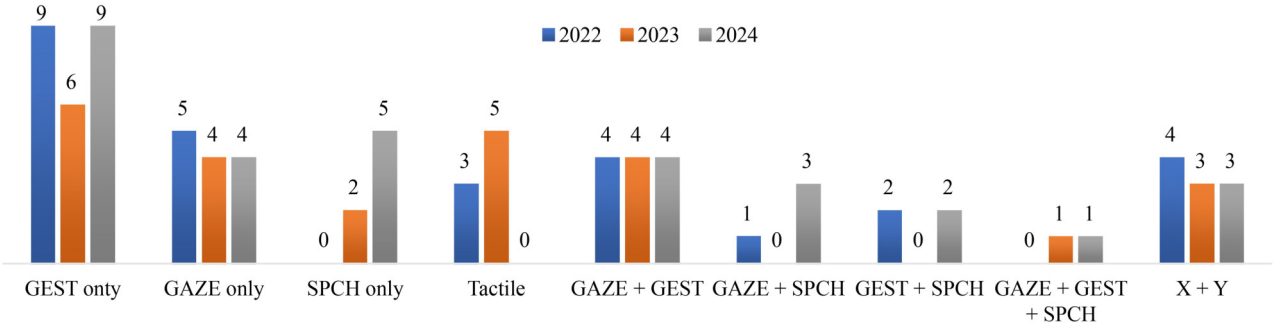



**Fig. 2** Active interaction modality data, categorized by nine modalities and three years. This figure is consistent with the literature recorded in Table 1

increase in research on Speech-only interaction is observed in 2024, accompanied by a rise in Speech-related multimodal studies, likely driven by recent advancements in LLMs. Tactile interaction-related literature is comparatively less common. In the analysis of Passive interaction modalities, tactile interaction is frequently studied as a feedback mechanism rather than as an active interaction method.

The number of publications for various operation types in the collected literature is as follows: Pointing and Selection (45), Creation and Editing (16), Translation and Transform (13), Locomotion and Viewport (9), Typing and Querying (19). The proportions of research since 2022 are illustrated in Fig. 3. Pointing and Selection remains the most focused area, although the number of studies has been decreasing annually. This decline may be due to advancements in more precise multimodal selection techniques. Significant increases in research on Locomotion and Viewport and Typing and Querying were observed in 2024. This rise may be attributed to growing attention on users' subjective experiences and the development of LLMs. These models enable users to communicate more diverse semantic information to XR systems. Further analysis of the technological developments behind these trends will be provided in Section 3.

Passive interaction types are categorized as Visual, Acoustic, Haptic, and Hybrid. Specific details are available in Table 2. We put the passive interaction into a new table (Table 2) rather than include it in Table 1 for the following reasons. Firstly, there are inherent differences between the modalities of active and passive interactions, and they are not directly equivalent. For example, while active interaction involving human's eyes is typically referred to as gaze (or eye gaze) interaction, passive interaction involving the eyes is generally categorized as visual feedback. Besides, Table 1 is primarily designed to illustrate the relationship between modalities and operation types of the reviewed papers. Passive interactions, however, do not engage with the concept of "operation" in the same way. Thus, it is not appropriate to include passive interactions in Table 1, which focuses on active interactions.

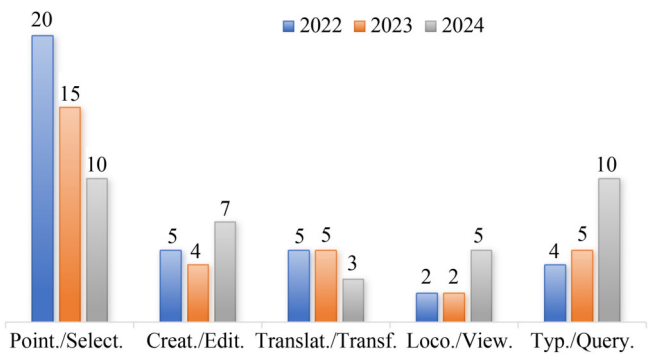



**Fig. 3** The number of publications for the five main operation types. This figure is consistent with the literature recorded in Table 1

Recent studies have increasingly focused on haptic feedback, likely due to its ability to significantly enhance users' immersive experiences. It is relatively commmon to combine visual and acoustic feedback, while standalone acoustic feedback is rarely studied. Some research also explores visual feedback alone. Moreover, the integration of the three primary sensory feedback modalities (visual, acoustic, and haptic) has emerged as a major research direction.

There are some devices that are preferred by researchers to conduct their experiment. The situation is illustrated in Table 3. Microsoft Hololens 2 and HTV Vive Pro / Vive Pro Eye are the most popular devices used for experiment in the recently two years, while Hololens 2 is more preferred in 2024. Meta Quest 2 (Oculus Quest 2) is less used after 2022 but still popular.

## 3 A Taxonomy and analysis of natural interaction techniques

Previous XR survey studies have primarily categorized research based on display type, study type, use case, input technique, and task type [1,24,27]. Since this paper focuses on head-mounted displays, there is only one display type. The study type refers to the type of user evaluation conducted (e.g., assessment or comparison), while the use case examines the conditions under which studies are conducted (e.g., static or in motion). Given the focus of RQ1 and RQ2, this paper primarily explores novel interaction paradigms and techniques, as well as evolving trends in multimodal natural interaction. Therefore, we do not discuss study type or use case in this work but retain input technique. Regarding task type, existing research often considers application scenario and operation type as part of task classification and discusses them together [1,24]. In this paper, we separate these two aspects.

As a result, we retain three classification dimensions. Our categorization follows the order of application scenario → operation type → interaction technique. This structure facilitates the analysis of inherent developmental trends in

**Table 1** The reviewed literature is categorized based on six operation types, nine interaction modalities, and year of publication ("X+Y" represents other modality combinations, where "X" denotes a modality from gesture, speech, or gaze, and "Y" denotes any other modality except these three (e.g., gaze + Head pose))

| Operation type | Year | Unimodal interaction | | | | Multimodal interaction | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | GEST only | GAZE only | SPCH only | Tactile | GAZE + GEST | GAZE + SPCH | GEST + SPCH | GAZE + GEST + SPCH | X + Y |
| Pointing and selection | 2022 | [11][34][35][36][37] | [38][39][40][41][42] | | [43] | [19][20][44] | [3] | [45] | | [46][47][48][49] |
| | 2023 | [50][51][52][53] | [54][55] | | [50] | [56][57][9][58] | | | [59] | [60][61][62] |
| | 2024 | [4][63][64] | [65][66][67] | | | [68][69] | | | | [70][71] |
| Creation and editing | 2022 | [37][72] | | | [43][73] | | | [45] | | |
| | 2023 | [74][52] | | | [75] | | | | [59] | |
| | 2024 | [4][64] | [67] | [18][12] | | | | [17][69] | | |
| Translation and transform | 2022 | [72][36][37] | | | [73] | | | [76] | | |
| | 2023 | [52][74][53] | | | | [9] | | | | [61] |
| | 2024 | [77] | | [12] | | | | [17] | | |
| Locomotion and viewport | 2022 | [37][34] | | | | | | | | |
| | 2023 | | | [78] | [79] | | | | | |
| | 2024 | [63][80] | | [81] | | [82] | | | | [83] |
| Typing and querying | 2022 | [84][85] | | | | [86] | | [45] | | |
| | 2023 | [87] | [88][89] | [90] | [75] | | | | | |
| | 2024 | [91] | [92] | [18][93][81] | | [94] | [21][95][96] | | [97] | |
| No operation | | [98][99][100][101] | | [102][93] | [103][104][105] | | | | | |

**Table 2** The literature related to passive interaction

| Passive interaction | Literature |
|---|---|
| Visual | [106][107][108][109][110][111][112][113][114][115][116][117][118] |
| Acoustic | [110][106][111] |
| Haptic | [119][120][121][122][123][124][125] |
| Hybrid | [108][106][121] |

**Table 3** Devices chosen by researchers (by years)

| Device name | 2022 | 2023 | 2024 |
|---|---|---|---|
| Microsoft hololens 2 | 7 | 6 | 13 |
| HTC Vive Pro/Vive Pro Eye | 9 | 8 | 7 |
| Meta Quest 2 (Oculus Quest 2) | 8 | 5 | 6 |

**Table 4** The literature of different application scenarios

| Application scenario | Literature |
|---|---|
| Drawing and sketching | [43,52,59,67,69,74] |
| Smart assistants | [12,21,81,96,97,109,110] |
| Virtual meetings | [17,45,62,115,117,125] |
| VR/AR navigation | [4,81,110] |
| Reading | [38,47,49] |
| Furniture assembly/maintenance | [4,110] |
| Remote collaboration | [3,95] |
| Autonomous driving | [111] |
| Enter passwords | [85,101] |

interaction modalities through the lens of application requirements. It also progresses from a broad, high-level perspective to a more detailed and specific one.

The subsequent sections are organized as follows: Subsection 3.1 presents a detailed discussion of natural interaction applications, Subsection 3.2 provides an analysis of seven distinct operation types, and Subsection 3.3 examines the implementation methodologies of various interaction modalities and the design considerations employed by researchers in developing these interaction systems.

### 3.1 Application scenarios

In the reviewed literature, nearly 70% of the research focuses on interaction design for general scenarios, without specifying particular application contexts. For example, some studies design gaze vergence control techniques [42,55,66], create more than 10 different hand gestures for interaction [52,72], and develop robust voice keyword detection methods [90,93]. However, we argue that applying multimodal natural interaction to specific application scenarios is crucial. Implementing these techniques in real-world contexts not only enhances user experience but also increase user exposure to multimodal interaction technologies. In turn, this can accelerate the adoption of these technologies and promote their use across a wider range of fields. Therefore, in this section, we discuss studies that explore specific application scenarios, as summarized in Table 4. Several exemplary application scenarios are illustrated in Fig. 4. Meanwhile, we also provide an overview of the interaction types and operation types utilized.

We find six papers apply natural interaction techniques to drawing and sketching [43,52,59,67,69,74]. In these studies, users primarily engage in sketching through gestures (bare hands, pen, touch devices, or controllers), while using eye-tracking, voice commands, or gestures for menu selection, such as switching brush colors. Eye-tracking is also employed to control 3D grids, allowing users to perceive depth more intuitively [67]. Sketching is considered a relatively complex task, involving operations like pointing, selection, creation, and editing. Two papers allow users to manipulate different geometric objects, including actions such as scaling, translation, and rotation [52,74].

Seven papers explore the application of smart assistants [12,21,81,96,97,109,110]. Among these, six papers utilize LLMs as assistants, benefiting from their superior comprehension and reasoning abilities. For example, Giunchi

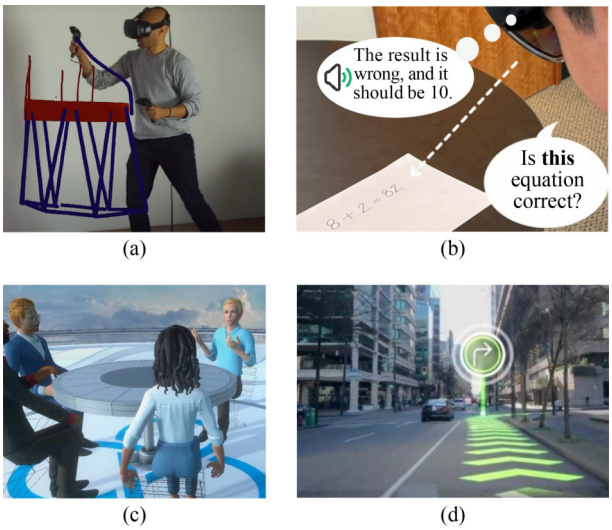



**Fig. 4** The illustrations of XR application scenarios. (a) Drawing and sketching; (b) smart assistant [97]; (c) virtual meeting; (d) AR navigation. Image courtesy of [97]

et al. allowed users to directly edit virtual objects through voice commands, such as creating or moving them [12]. Wang et al. provided intelligent guidance for virtual tours by recognizing user speech and environmental information, then delivering multimodal feedback, such as avatars, voice, text windows, and minimaps [81]. Lee et al. offered assistance with daily activities by using LLMs to answer questions related to objects of interest such as food calories or offer book recommendations, as identified by the user's gaze and speech [97].

Six papers explore the application of natural interaction techniques in video conferencing or virtual meetings [17,45,62,115,117,125]. For example, Saint-Aubert et al. focused on verbal communication with avatars in VR, introducing synchronized haptic feedback based on speech content to enhance immersion [125]. Lee et al. proposed using visual cues (lighting effects) and auditory cues (spatial audio) in VR social interactions to guide users' attention to new speakers [117]. Liao et al. and Cao et al. explored techniques for enhancing speaker presentations in video conferences [17,45]. They extract keywords from users' speech and match them with predefined images and animations, which are then synchronized with hand movements to improve expression.

Three papers investigate the application of natural interaction techniques in VR/AR navigation [4,81,110]. For example, Quere et al. used a hand-menu system to create virtual annotations for a large campus hosting a reception, which help direct participants to meeting rooms [4]. Wang et al. used LLMs to generate guidance for virtual tours through avatars, voice, and text windows [81]. It is important to note that our definition of navigation refers to guiding the user to quickly locate their position, which differs from the definition of virtual navigation in related studies that focus on controlling travel speed and direction of users [63,80,82].

Three papers explore the application of natural interaction in reading [38,47,49]. For example, Lee et al. designed a set of gaze-based interaction strategies to select text, zoom in on specific areas, and scroll through content, thereby enhancing the reading experience [38]. Meng et al. investigated head-based pointing combined with three hands-free selection mechanisms, i.e., dwell, eye blinks, and voice (hum), to select text during reading [49].

One paper discusses the use of turn-by-turn animations and text instructions displayed on AR glasses for furniture assembly [110]. One paper explores the application of AR in maintenance, such as annotating a broken video projector to guide other users in understanding how to operate it [4]. Two paper focus on remote collaboration, proposing specific methods to quickly guide collaborators to follow an expert's line of sight [3,95]. One paper explores the application of VR-based interactive feedback in autonomous driving [111]. Additionally, two papers examine how gestures can be used to quickly enter passwords in VR [85,101].

To summarize, current research on XR natural interaction techniques has explored 10 types of applications, offering a relatively diverse range of studies. However, these papers account for only 30% of the reviewed literature, indicating that further exploration is needed on how to apply natural interaction techniques to practical use cases. Additionally, the range of application types should be expanded to include fields such as medicine, industrial training, and education, all of which hold significant potential for impactful applications. Further discussion can be found in Subsection 4.3.

### 3.2 Operation types

The applications mentioned above are built on specific natural interaction operations. Researchers have developed various operations and explored diverse interaction modalities, resulting in a complex mapping between operation forms and interaction modalities [24]. This complexity underscores the importance of identifying similarities among these operations. In this section, we introduce a rational classification that reveals the relationships between different operations. This taxonomy helps to identify the core issues each type of operation must address and suggests future development trends. The paper categorizes XR operations into seven main classes, as explained below. Figure 5 demonstrates various types of operations.

Before interacting with an object, the initial step is to select it. Even when creating new objects, users must first preselect a location. The ability to quickly, accurately, and stably select

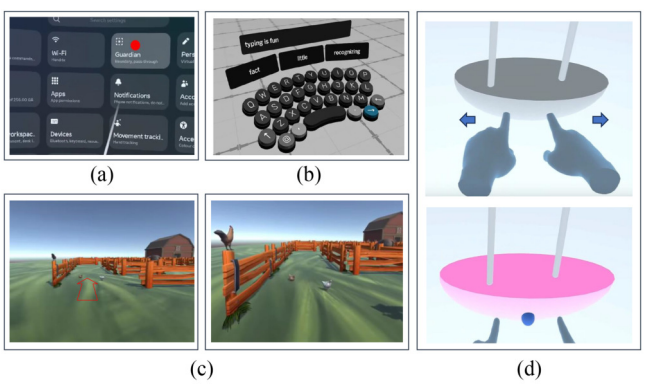



**Fig. 5** The illustrations of operation types. (a) Pointing and selection; (b) typing; (c) locomotion; (d) transform [52]. Image courtesy of [52]

desired objects is essential in nearly all scenarios [71]. This process is categorized in this paper as Pointing and Selection. XR enhances human capabilities by providing greater control over virtual environments, particularly with respect to object manipulation [98]. The ability to create objects and modify their properties is a central feature that human-centered XR should offer. This paper classifies these actions as Creation and Editing. Once objects are selected, users can perform various interactions, depending on the scenario, including controlling spatial position, size, and orientation. These actions are referred to in this paper as translation, scaling, and rotation, with the latter two collectively termed Transform. Therefore, these operations are grouped under a single category: Translation and Transform.

The aforementioned categories relate to user interactions with objects. However, users also need to navigate and observe within the virtual environment, which involves movement through space (Locomotion) and changes in viewpoint (Viewport). This paper combines these activities under the category of Locomotion and Viewport.

With the continued advancement of XR and AI technologies [126,127], users are presented with increasingly rich information within virtual environments. This development necessitates more frequent exchanges of information between users and systems [96]. Users must input task-related information into the system, while the system must accurately interpret user intentions and provide appropriate feedback. Based on recent research, we classify information interaction into two categories: Typing and Querying. The former refers to the primary method of text input, while the latter includes various forms of information retrieval for users.

The aforementioned interaction types involve active engagement of users with the environment, where users input information into the system. We categorize these operations as Active Interactions. Beyond user input, receiving feedback from the XR system is another critical component of natural interaction, as it enhances user immersion. Research that investigates methods for delivering feedback to users is categorized as Passive Interactions.

In our literature review, we also identify a distinct category of research that does not focus on specific operations but rather on improving the recognition accuracy of input modalities. We classify this body of work under the No Operation category.

This classification offers a structured framework for understanding and analyzing XR interactions, capturing the core ways in which users engage with and manipulate virtual content. It encompasses both the creation and modification of virtual objects as well as user movement and perception within the XR environment, providing a comprehensive approach to the study of XR operations.

#### 3.2.1 Pointing and selection

Pointing-and-Selection is the most common and fundamental operation in XR, essential for interacting with any object. Selection operations are often influenced by scene factors such as object size, density, and depth distribution [71]. Among recent research, 2D menu layouts are prevalent in selection operation. These studies focus on designing task-specific menu layouts for different application contexts to obtain better performance of selection [4,38,40,41,56,59,65,70]. Research focusing on selection in the 3D space primarily aims to address the problem of accuracy in high-occlusion scenarios [46,55,60]. Some studies also discuss user experience in text selection [38,49] and selection across different depth planes [42,46,66].

- Performance measures. These papers share a common set of evaluation metrics. Objective metrics predominate, including Target Delay, which describes the time for the pointing modality to align with the target object; Selection Time, measuring the total duration from object appearance to selection; False Positive Rate, indicating the frequency of false touch occurrences; Selection Accuracy, describing the overall accuracy of a selection method; Error and Abort Rate, assessing the difficulty of using a particular method; and Hand or Controller Movement, quantifying the required hand motion, closely related to arm fatigue. Common subjective metrics primarily evaluate Cognitive Load and Subjective Preferences.
- Modalities. Eye tracking is the most commonly used modality for pointing and selection because the human eye naturally focuses on objects of interest, as demonstrated in Table 1. However, gaze estimation methods still have limitations in precision, as sampled gaze points represent a spatial distribution rather than a fixed point due to tremors or involuntary saccades [55]. Pure eye-tracking pointing and selection often incorporate special interface designs or vergence control to address these issues [38–42,55,60,65,66]. Additionally, introducing secondary modalities for fine-tuning or confirming selections is a common approach [19,20,56,57,61]. A few studies explore the reliability of gesture-based selection, typically applied in scenarios where gestures align more closely with human intuition [4,35,52].

#### 3.2.2 Creation and editing

As XR technology advances, allowing users to freely create and modify objects in XR significantly expands their interaction freedom with the environment [18]. This set of operations primarily involves Creation and Editing operations. The current mainstream Creation mode uses predefined patterns, where users or developers predefine a set of virtual components that can be created by users through specific interaction modalities (e.g., [4,17,45]), mostly triggered by voice [45] or gestures [17]. In this mode, designing easily memorable and manageable Creation methods for users is a primary concern [59,72]. Additionally, some studies explore methods for users to freely create non-predefined objects [12,52,72,74]. They focus on diversifying creatable content to provide users with more freedom. Editing operations act on these created objects. As virtual elements can take various forms such as text, images, windows, and maps, there are multiple editable attributes like size, texture, and color [81].

- Performance measures. The primary concern in Editing operations is whether users can easily remember and use the relevant interaction methods, which are highly application-specific. Common metrics include both objective and subjective measures. Objective metrics include Completion Time and Success Rate. Subjective metrics mainly cover Engagement, Confidence, Expressiveness, Learnability, and Easiness of Use.
- Modalities. Hand gestures are the most common modality for Creation. Six degree-of-freedom (DoF) gestures have attracted significant research attention due to their extensive design space and the diversity of creatable objects. Numerous studies have designed intuitive creation gestures for various objects [53,59,72,74]. Given the rich semantics of human language, coupled with advanced NLP technologies, speech interaction can provide more precise information about the objects to be created and their attributes [17,45]. A few studies have incorporated eye tracking to select creation positions or specify menu items [49,59,69].

#### 3.2.3 Translation and transform

This operation corresponds to users' fundamental control capabilities on objects, which include transform, rotation, and scaling. Recent studies have defined richer interactions for precise object manipulation, extending beyond intuitive hand gestures. As XR technology advances, virtual objects have varied, necessitating simultaneous multiple basic operations on single objects. Research have focused on enhancing transformation richness and flexibility through customized gestures [17,74] and intuitive designs for various objects [52,72]. Studies also aim to extend user reach for larger-scale transformations in limited spaces [53,77], and improve interaction feedback, including haptic responses.

- Performance measures. Evaluation in these works combine objective and subjective metrics. Objective metrics include Accuracy, Speed, Stability (for measuring consistency during jittery movements and across multiple operations), and Precision (for positional accuracy). Subjective metrics focus on Learnability and Fatigue.
- Modalities. This operation typically employ intuitive hand gestures, such as pinch and grab for moving objects [4], and two-handed pull for scaling [17]. Studies have explored varied implementation details, including customizable gestures [17] and gestures inspired by everyday tool use [72]. Some research examines learning abilities and operational stability for specific gesture-based transformations [17]. However, gesture detection inevitably involves errors [17]. Some studies incorporate voice input for improved stability [17,45]. Gaze interaction, being quicker and less fatiguing, has been utilized for object transformation, with gestures often used for fine-tuning and three DoF rotation control [61]. A few studies have explored hands-free approaches using eye tracking and head-gaze for transformations [61].

#### 3.2.4 Locomotion and viewport

This operation includes both Locomotion and Viewport, which are essential for user exploration and movement in virtual environments, significantly influencing user comfort and immersion [80]. Traditional locomotion research has primarily focused on 2D planar movement, with teleportation and steering being the two most common modes. The former rarely causes discomfort, while the latter offers greater immersion [78]. Recent studies have extended this focus to 3D space exploration [63,80], aiming to improve the overall user experience. In collaborative VR scenarios, enhanced locomotion techniques help users better understand each other's movement intentions, preventing disconnection or separation [113]. However, traditional controller-based rotation and directed steering can detract from immersion, as they fail to provide natural and continuous turning experiences [78]. Improvements to the Viewport are equally important for enhancing user comfort and overall experience. One study introduces three gaze-controlled viewport methods that enable hands-free and controller-free interaction [83].

- Performance measures. Measures for locomotion primarily utilize subjective metrics, including Presence, Workload, Cybersickness, Preference, and Overall User Experience, while the main objective metric is Task Completion Time. For viewport evaluation, researchers employ similar subjective measures as locomotion, with the addition of Error Rate as an objective metric to quantify the ease of view manipulation.
- Modalities. In steering mode, except from body-leaning, recent studies have focused on gesture-based movement [80] and gaze-directed locomotion with speed and turning control using gestures [82]. Voice commands have also proven to be a viable option for steering control [78]. In teleportation mode, research has explored voice-based destination matching [78] and gesture-based alternatives to controllers [63]. Some studies focus on teleportation in collaborative VR scenarios, addressing the challenge of communicating teleportation intentions between users [113].

#### 3.2.5 Typing and querying

This kind of operation includes typing and querying, crucial for inputting information into and retrieving feedback from XR systems. Typing efficiency significantly impacts productivity [86]. Recent XR interaction advancements offer diverse modalities for information retrieval [96]. These modalities of information and their combinations have been leveraged in recent works [96,97]. User immersion in XR querying is enhanced when input methods closely resemble everyday communication, requiring systems to discern user intent [97].

- Performance measures. Evaluation for typing primarily use objective metrics such as Words per Minute (WPM), Error Rate, Deletion Count, and Prediction Count. Subjective metrics include cognitive, visual, and hand fatigue levels. Evaluation of Querying employs both subjective metrics (Simplicity, Naturalness,

Human-likeness, Personal Preferences), and objective metrics (Task Time, Times of Attempts).

- Modalities. Recent studies often retain traditional keyboard layouts while incorporating new modalities to improve speed and experience [85–88,94,128]. Gaze-tracking is used to predict characters of interest, accelerating cursor movement or providing visual cues [88,94]. He et al. proposed a gaze-selecting method combined with language models for error-tolerant blind typing [86]. Some studies achieve faster input speeds compared to midair-tapping using pure gaze input [89,92]. Gestures are used for keyboard control [85], and new 3D decoding techniques enable precise interpretation of aerial gestures [87]. For querying, speech-based interaction is fundamental, with lip-reading and other auxiliary methods improving robustness [90,93]. Eye-tracking data and gesture information can also help resolve pronoun ambiguities in speech [96,97]. One study enables gaze-only querying through an AR interface design [45].

#### 3.2.6 No operation type

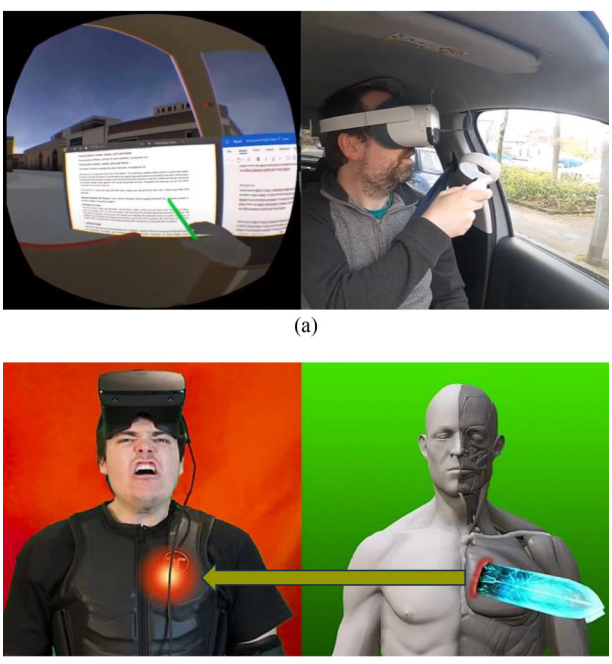
In our literature review, we identify nine studies that do not explicitly focus on specific interaction types. Instead, these studies provide optimizations aiming at improving the recognition accuracy of input modalities. These studies primarily aim to enhance the system's ability to accurately interpret and respond to users' active interaction intents. Since these papers are not discussed elsewhere in this work, we provide a detailed description here.

The majority of these studies center around gesture and touch-based interactions. Xu et al. proposed a fine-grained hand gesture detection method that leverages both visual and auditory sensors, enabling users to trigger interactions with smaller movements [103]. Kitamura et al. developed a contact-based wearable device capable of accurately recognizing micro-gestures, including continuous gesture changes and pressure inputs [104]. Lee et al. introduced a wrist-worn device that employs active acoustics to continuously capture hand movements and interactions with objects [100]. Liu et al. designed a gesture recognition method using audio signals, enabling the recognition of various hand gestures on different material surfaces, thus expanding the possibilities of gesture interaction [102]. Li et al. developed a finger-ring-based micro-gesture recognition device equipped with a miniature camera, allowing for gesture interaction on various surfaces [105]. Rupp et al. proposed a set of gesture authentication schemes that achieve the same entropy as PIN codes [101]. Shen et al. introduced a Key Gesture Spotting architecture to assist developers in rapidly developing gesture recognition systems, while simultaneously reducing gesture detection latency for a better user experience [99].Wang et al. designed a visual guidance mechanism for bare-hand interaction to help users perform interaction actions more consistently, reducing recognition errors [98]. Cai et al. developed a dual-mode keyword detection system using both speech and echo signals, enabling accurate keyword recognition in a wider range of environments [93].

#### 3.2.7 Passive interaction

In addition to the previously mentioned Active Operations and No Operation, we also examine Passive Interactions. In recent years, many studies have focused on this topic. Based on the types of feedback provided by XR environments, we categorize these studies into four groups: Visual, Acoustic, Haptic, and Hybrid. For more detailed information on this classification, please refer to Table 2. Figure 6 illustrates passive interactions.

- Performance measures. Passive Interaction primarily focuses on users' perception of the environment, making subjective metrics the primary evaluation criteria. Data in various studies is typically collected through questionnaires designed by the authors. The most commonly assessed indicators include Realism, Immersion, and Confidence. In haptic-related research, Visuo-Haptic Match is also a key metric [119]. Additionally, some studies use task completion scores in XR environments as an evaluation method [106,112].
- Modalities. The acoustic modality is not observed in isolation in the collected articles. It is typically studied in conjunction with other modalities. Visual feedback represents one of the primary feedback modalities. The impact of singular visual feedback on users' self-position perception during teleportation has been investigated [107]. Visual guidance mechanisms are employed in three articles to enhance communication efficiency among multiple users in XR environments [113,115,117]. Two articles dynamically alter environmental parameters based on users' task stages in the XR environment to improve task completion efficiency [109,112]. Three papers design feedback in VR environments based on real-world distractions,



(a)

(b)

**Fig. 6** The illustrations of passive interaction. (a) Visual feedback [106]; (b) haptic feedback. Image courtesy of [106]

aiming to reduce the impact of distractions on user immersion and comfort [106,111,116].

The haptic modality has garnered significant attention from researchers in recent years as a passive interaction modality. Five articles utilize auxiliary wearable hardware to provide additional tactile feedback, enabling users to experience a wider range of sensations in XR [119,120,122–124], such as wind [119]. Visuo-Haptic illusion with Proxies is examined in two articles, offering users perception of objects' size, weight, and motion trajectories [108,121].

### 3.3 Interaction techniques

The aforementioned operations are supported by specific interaction modalities. This section focuses on the hardware and algorithmic implementations of these modalities, as well as the design of concrete interaction techniques. Subsection 3.3.1 discusses the hardware and software implementations of gesture-only interactions, provides a summary of existing research, and analyzes the strengths and limitations of this modality. Subsection 3.3.2 covers the same aspects for gaze-only interactions, while Subsection 3.3.3 addresses speech-only interactions. Subsection 3.3.4 explores tactile interaction in a similar manner. Subsection 3.3.5 focuses on multimodal interaction techniques, specifically covering Gaze + Gesture, Gaze + Speech, Gesture + Speech, and a triple-modal technique (Gaze + Gesture + Speech). Lastly, we discuss various X+Y interactions, which compare different multimodal interaction techniques in these studies.

#### 3.3.1 Gesture only

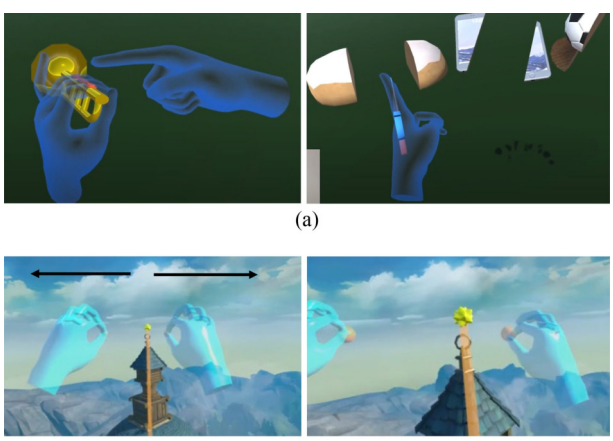
We begin by introducing the implementation methods for hand gestures.

- Hardware. Currently, hand tracking mainly relies on sensors integrated into VR/AR HMDs, primarily utilizing infrared (IR) cameras. The latest devices, such as the Meta Quest 3, additionally use two RGB cameras for enhanced tracking [2]. The hand-tracking performance of XR devices used in recent research is summarized in Table 5. Among these, the HTC Vive Pro shows lower accuracy compared to the HoloLens 2 and Meta Quest 2 [129]. Although the latency and sampling rate of the HoloLens 2 and HTC Vive Pro have not been documented, users have not reported experiencing any noticeable delays during use. Typically, users rely on the HTC Vive Pro's hand controller for gesture-based interactions. Two papers [100,102] explore low-power sensing modules, such as speakers and microphones, to detect hand-reflected sound waves as an alternative to image-based tracking.
- Algorithm. Hand tracking in VR/AR HMDs typically employs computer vision techniques to output hand skeleton data [130,131]. Four papers [52,72,99,100] aim to classify a larger variety of gestures (e.g., 10 or more) using neural networks or machine learning. Other studies focus on simpler gestures like pinching or clicking, which only require distance detection between the index finger and thumb.
- Summary of current research. In the reviewed literature, gesture only interaction is the most frequently studied modality for spatial computing, with 24 papers focusing on this topic, shown in Fig. 2. The hands are natural and intuitive interfaces for interacting with objects in daily life. For example, 20 papers explored using hand gestures for object selection or translation. Hand movements in 6DoF have been used to create complex 3D scenes [74,84] and for VR locomotion [63,80]. Gestures also enable fine-grained control of small objects. Three papers investigate text input in VR, either via direct virtual keyboard interactions [87,101] or gesture-to-text conversions [85]. Gesture transformations can convey complex semantic information. For instance, two papers use 10 or more gestures to imitate a wide range of objects, such as a telescope, scissors, or camera [52,72]. Additionally, three papers explore hand redirection techniques, leveraging the perceptual phenomenon of change blindness in hand and arm movements [11,36,68]. This allows users to operate within a limited physical space while enabling broader interactions in a virtual environment. Figure 7 illustrates hand gesture only interactions.
- Advantages and disadvantages of hand gesture only interaction. Hand-gesture only interaction offers several advantages. It is natural and intuitive, as gestures align with everyday interactions. Gestures also allow for complex semantic representation and providing fine-grained control for precise manipulation of small objects. However, there are notable drawbacks. For the use of complex gestures, the cognitive load increases as users need considerable time to master them [72].

**Table 5** Comparison of hand tracking performance across devices

| Device name | Accuracy/mm | Latency/ms | Sampling rate/Hz |
|---|---|---|---|
| Microsoft HoloLens 2 | around 15 | – | – |
| Meta Quest 2 | around 11 | 45 | 60 |
| HTC Vive Pro | around 37 | – | – |
| Lee et al., 2024 [100] | 4.81 | 500 | – |

Note: – indicates that no reports are found regarding this item.



(a)

(b)

**Fig. 7** The illustrations of hand gesture only interaction. (a) Use gestures to imitate a wide range of objects; (b) move forward through zooming out with both hands. Images courtesy of [72,80]

Gesture recognition accuracy still requires improvement, particularly in cases of occlusion, as mentioned in papers [52,72]. Additionally, prolonged use of gestures can lead to arm fatigue [4,63], and the lack of tangible support means that users do not receive physical feedback [35,87]. Social acceptance can also pose a challenge, as obvious hand gesture interaction may not be appropriate in public [105]. To summarize, hand gesture interactions often require users to adapt to complex paradigms defined by researchers, which increases the cognitive load. Future developments should focus on improving ease of use, making hand gesture interaction more user-friendly for novices, and minimizing tracking accuracy requirements.

#### 3.3.2 Gaze only

Gaze-based interaction is a key focus in extended reality, leveraging humans' intuitive eye movements. Eye gaze rapidly reaches objects of interest, making it ideal for pointing and selection tasks [132]. Gaze-only methods eliminate the need for body movements, benefiting users in various situations, including those with disabilities or in socially awkward scenarios [83].

**Hardware and algorithm**. There are two main eye-tracking strategies in XR environments [133]. The first is the Pupil Center Corneal Reflection (PCCR) method [134,135]. This approach uses 1–2 near-infrared cameras positioned close to the eye, along with multiple near-infrared light sources (typically 6–12) directed at the eye. Based on the corneal reflection patterns, the system reconstructs an accurate eye model. The advantage of this method lies in its high precision and its ability to support slippage detection and compensation for device slippage [136]. However, it has drawbacks, including higher hardware costs and a more complex process for calibrating the optical positions of the light sources and cameras [137]. Currently, commercial XR devices such as Microsoft HoloLens 2, HTC Vive Pro Eye, and Meta Quest Pro utilize the PCCR-based eye-tracking approach, as shown in Table 6. These devices typically achieve eye-tracking accuracy within the range of 0.5°–1.6°. Furthermore, because they are robust against device slippage, the differences in eye-tracking performance during interaction can be negligible. The latency falls within an acceptable range for interaction purposes, and a sampling rate of 30 Hz is generally sufficient for gaze-based interaction. Overall, the eye-tracking performance of these XR devices is comparable.

The second method is glint-free, which relies on temporal pupil information rather than corneal reflections [138,139]. This technique typically uses a single camera and a single light source primarily for illumination. By analyzing pupil data over multiple frames, it reconstructs the eye model. The advantage of this method is its simplicity and lower hardware requirements, although the accuracy of the eye model reconstruction is generally lower compared to PCCR [42]. Early implementations, such as the Microsoft HoloLens 1 integrated with Pupil Labs' eye-tracking system [139], uses this glint-free approach. While its accuracy is comparable to the PCCR method under ideal conditions, it is highly sensitive to device slippage, resulting in a rapid deterioration of gaze accuracy over time.

**Table 6** Comparison of eye-tracking performance across XR devices

| Device name | Accuracy | Latency/ms | Sampling rate/Hz | Slippage-robust |
|---|---|---|---|---|
| Microsoft HoloLens 2 | 1.5° | - | 30 | √ |
| HTC Vive Pro Eye | 0.5°–1.1° | 50 | 120 | √ |
| Meta Quest Pro | 1.6° | 58 | 90 | √ |
| Microsoft HoloLens 1 | 1° | 8.5 | 120 | × |

Note: Microsoft HoloLens 1 is integrated with Pupil Labs' eye tracker.

**Summary of current research.** Gaze-only interactions with virtual user interfaces (UIs) often encounter the Midas touch problem due to lack of confirmation modalities. To address this, researchers have explored peripheral vision areas and auxiliary UIs. Choi et al. [39] proposed the Kuiper Belt concept, while Yi et al. [41] studied optimal virtual menu layouts. Orlosky et al. [65] and Kim et al. [40] designed auxiliary interfaces to enhance pure gaze selection operations. Besides, several techniques have been developed for heavily-occluded scenarios. Sidenmark et al. [55] matched object depth motion with gaze vergence changes, while Wei et al. [60] used probabilistic models based on head and gaze endpoints. Yi et al. [46] combined planar and depth information analysis. Vergence estimation technology has introduced vergence control as a novel gaze-only interaction mode. Zhang et al. [66] designed visual depth control methods for object selection, while Wang et al. [42] developed depth control schemes for see-through vision. In typing applications, pure eye movement input methods have shown promising results. Cui et al. [89] designed a word prediction algorithm based on eye movement trajectories. Similar systems integrated with a LLM are also proposed by Hu et al. [92].

Gaze-based interactions have also been applied to enhance user experience in various contexts. Chen et al. [54] explored activating hidden objects in virtual films, Turkmen et al. [67] investigated gaze-activated auxiliary grids in virtual sketching, and Lee et al. [38] introduced gaze-activated magnification in VR reading. Additional applications include perspective switching control using different eye movement modes [83] and text operations such as Selection-and-Snap and Gaze Scroll (Lee et al. [38]). These diverse applications demonstrate the potential of gaze-based interactions to significantly enhance user experiences across various XR scenarios. Figure 8 illustrates gaze-only interactions.

**Advantages and disadvantages of gaze-only interaction.** The main advantages of eye movement interaction include speed, ease of use, and intuitiveness. Its maximum utility is demonstrated when large-scale physical movements are impossible or when both hands are occupied with tasks. Furthermore, eye movements possess dual selection capabilities in both 2D planes and depth directions. These capabilities can be combined to achieve more diverse interactions. Richer semantic information can also be extracted from human eye movement patterns. Pure eye movement modality has the potential to enable more powerful interaction functions. In eye movement interaction, the human eye serves as both the medium for initiating interactive behaviors and the primary sensory organ for observation in

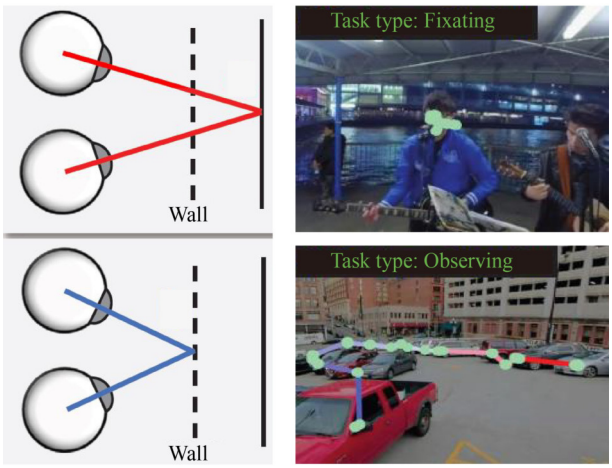



**Fig. 8** The illustrations of gaze only interaction. (a) Gaze vergence control [42]; (b) gaze patterns [118]. Images courtesy of [42,118]

XR. Consequently, the Midas touch problem persists. This issue remains a primary consideration for researchers in subsequent studies. Similarly, many eye movement interactions involve the design of new interaction interfaces and visual cues to assist users in completing interactions. The visual disturbances and cognitive load imposed on users by these new interfaces are significant issues that cannot be overlooked.

### 3.3.3 Speech only

**Hardware and algorithm.** Speech-based interactions in XR environments are typically enabled by a single microphone. While some researchers utilize an external microphone [125,128], the majority depend on the built-in microphone of VR/AR HMDs, such as Hololens 2 [59], HTC Vive Pro Eye [3,49], and Valve Index [78]. For speech recognition, most researchers rely on established speech-to-text systems or APIs, such as Windows DictationRecognizer [3] and WebSpeech [45]. To summarize, these devices and methods have been well established in recent years.

In addition to traditional method above, there is a growing interest in silent speech recognition, which enables speech detection without audible vocalization. This approach often involves additional devices, such as those used in active acoustic sensing. For example, EchoSpeech [90] leverages inaudible echo wave that are emitted by two speakers and received by two microphones mounted on a glass-frame. Cai et al. [93] further enhanced this by combining both vocal and echoic modalities for more robust speech recognition across various scenarios. Depth cameras have also been explored for silent speech recognition. Wang et al. [140] used TrueDepth camera at three different locations to capture lip movement depth data, enabling silent speech recognition via point cloud video analysis.

**Summery of current research.** Recently, speech-based interactions can be generally classified into two categories: keyword-based and LLM-based. Keyword-based speech interactions identify keywords provided by users as interaction cues or commands, typically including pronouns (e.g., "this" and "here") [3], commands (e.g., "teleport" and "select") [59,78], and user-defined keywords [17,45]. For instance, Hombeck et al. [78] introduced three speech-only interaction techniques for locomotion, by keywords of direction (e.g., "*left*"), landmark (e.g., "jump to *bed*") or grid number (e.g., "teleport to *fifteen*"). However, keyword-based interaction imposes significant limitations on the vocabulary available for interaction [78].

In contrast, LLM-based speech interactions overcome this restriction. With the increasing sophistication of large language models (LLMs), more researchers are leveraging LLMs as intelligent agents for interaction, allowing users to issue commands without vocabulary constraints. In this approach, the entire speech input is transcribed into text, which is then combined with prompt words or other contextual information as input to the LLM for further inference [12,18,81,97], as shown in Fig. 9. DreamCodeVR [12] exemplifies the use of LLMs to translate spoken language into code, enabling users to modify the behaviour of a running VR application irrespective of their programming skills. LLM-based speech interactions have significantly expanded the scope of operations and applications of speech-based interaction.

**Advantages and disadvantages of speech only interaction.** Speech only interaction is widely acknowledged for its minimal physical effort, as speaking in natural language requires little exertion [78]. Moreover, speech interactions for discrete selection are robust and do not require physical movements or gestures [59]. For text input task, speech input aided by speech-to-text technology is much quicker and more intuitive than keyboard input [78]. However, speech-only interaction has notable drawbacks. The weaknesses cited most frequently are high latency and inaccuracy. Several studies [3,45,49,78,125] report that delays or recognition errors have impacted the user experience during interaction. Additionally, speech interactions, especially keyword-based interactions, often increase cognitive load and raises learning curve due to the requirement of memorizing specific commands [59]. Furthermore, speech interactions lack subtlety and may face social acceptance challenges [49,81]. Fortunately, recent advancements mentioned above in silent speech recognition offer potential solutions of this problem.

### 3.3.4 Tactile interaction

In discussing Tactile Interaction, we refer to interactions with additional input devices that are light-weight, wearable and technologically advanced, for example those illustrated in

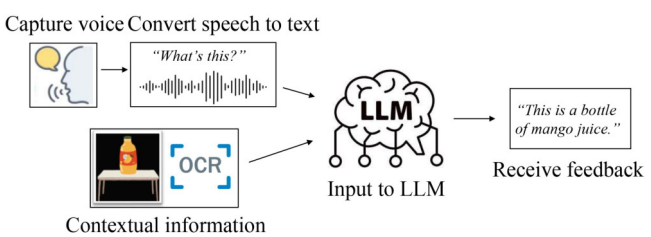
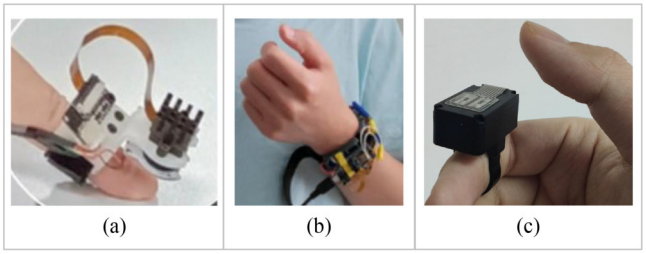



**Fig. 9** A common process in LLM-based speech interactions. The user's voice is captured and transcribed into text, which is then combined with contextual information to form prompts for the LLM. The LLM's feedback is subsequently provided to the user or processed further by the system



(a) (b) (c)

**Fig. 10** Different devices and locations of tactile interactions: (a) NailRing [105]; (b) AO-Finger [103]; (c) GazeRing [141]. Images courtesy of [103, 105,141]

Fig. 10. As such, interactions involving controllers or cumbersome tangible objects are excluded. This section focuses on two primary interaction techniques: finger gesture interaction and touch-based interaction. Additionally, studies investigating tactile feedback to enhance user immersion are also presented.

Finger gesture interactions have been extensively studied in recent years. Researchers use various input modalities, such as cameras [105], optical sensors [103,104], and electric field sensors [43], to recognize predefined finger gestures. The positioning of input devices also differs: NailRing [105] and TouchLog [104] place their devices above the user's index fingernail, whereas AO-Finger [103] positions its devices around the user's wrist. NailRing [105] utilizes a micro close-focus camera to capture changes of color in the nail and finger. Rather than cameras, TouchLog [104] employs a nail-type device with photo-reflective sensors for privacy, to detect the skin deformation of the fingertip during gestures. Beyond discrete gestures, EFRing [43] explores continuous 1D finger micro-movement tracking using a ring-shaped device through electric-field sensing.

Finger gesture interactions offer several advantages. Their primary advantage lies in their subtlety and greater social acceptance [43,103–105]. Notably, the pre-defined finger gestures in recent works typically involve no more than two fingers (thumb and index). Thus, these techniques are much more effortless and light-weight compared with hand gesture and other methods [103]. However, finger gesture interactions face challenges with generalization accuracy due to individual differences, as reported by [104,105]. Both studies suggest that individual calibration before use can mitigate this issue. Additionally, due to the complexity of recognition algorithms and the computational limitations of mobile devices, finger gesture recognition exceeds the capabilities of the minimal input devices and thus often confined to PC platforms [43,103].

Touch-based interactions involve works with the need for a surface or device to be physically touched. Researchers propose numerous devices for touch-based interactions, such as flexible sensors [75], packaged microphones [102], and even robots [79]. For text editing task in speech-unfriendly AR environments, TouchEditor [75] equips the user's arm with a flexible touchpad composed of flexible pressure sensors. They design numerous operations including text selection, cursor positioning, text retyping, and editing commands. Similarly focusing on text entry in AR, TapGazer [86] explores typing with TapStrap (finger-worn accelerometers), touch-sensitive gloves or touchpads, supported by gaze and language model. Buttons can also be wearable to offer a touch-based solution. FingerButton [50] allows users to seamlessly transition between the real world and VR with a finger-worn button device. Satriadi et al. [73] demonstrated greater creativity by leveraging tangible globes to explore the design of immersive data visualization in AR.

Touch-based interactions seem to represent a compromise between traditional device (e.g., controller and keyboard) and finger or hand gesture. Their performance of speed and accuracy is comparable to traditional input devices like keyboards [86], while offering greater convenience and subtlety akin to finger or hand gestures [50,75]. Moreover, they cause less arm and hand fatigue compared with gesture-based interactions. On the other hand, they still require additional hardware [86] and face the challenges related to the larger form factors of these devices [75]. Furthermore, touch-based interactions are dependent on surfaces (e.g., arm) for input.

Feedback from the interactive device is necessary to enhance the interactive experience [105]. Since wearable tactile devices is always available to access [50], several studies have investigated these devices to provide tactile feedback. For example, Saint-Aubert et al. [125] utilized a HapCoil-Plus actuator to generate speech-based tactile vibrations, aiming to enhance users' persuasiveness, co-presence, and leadership. Jingu et al. [124] introduced an electrotactile device with a thin and flexible form factor to enable double-sided simulation feedback within pinched fingerpads.

### 3.3.5 Multimodal interaction

Unimodal interactions, as discussed above, have been explored across various spatial computing scenarios, but each comes with its own limitations. By leveraging the complementary strengths of these modalities, multimodal interactions seek to enhance usability. Wang et al. [142] argued that eye gaze-based interaction is more suitable for primary target pointing, hand gestures provide capabilities for transformation and editing, and descriptive voice input improves system controllability, e.g., commands such as open/close or up/down. Inspired by these insights, recent research has developed various multimodal interactions for different XR tasks [3,59]. In this paper, we primarily focus on dual-modal techniques (e.g., Gaze + Gesture, Gaze + Speech, Gesture + Speech) and a triple-modal technique (Gaze + Gesture + Speech).

(a) Gaze + Gesture. Many studies adhere to the principle of "gaze selects, hand manipulates" [9], as eye gaze, which can rapidly locate a target, complements the fine-grained control provided by hand gestures. Apple Vision Pro identified this dual-modal interaction as the primary access method for spatial computing [6]. In the reviewed literature, two papers focused on text input tasks in XR environments [86,94]. Ren et al. [94] found that gaze reaches the target before the hand, and leveraged this to enlarge and highlight the possible following key, providing users quicker and more accurate

selection. He et al. [86] enabled users to input text by tapping the approximate area of a character on a keyboard, without looking at their hands or the keyboard. Additionally, users can resolve ambiguities by selecting the intended word via eye movement. Researchers also explore by using gaze as a directional reference, combined with hand movements along the gaze ray or swipes to navigate in VR environments [82].

Besides above tasks, many studies have examined menu selection and 3D object manipulation in XR [9,20,56–58,69]. Wagner et al. [20,56] designed a Fitts' Law study to evaluate the efficiency of two different gaze-hand alignment techniques for target selection. Shi et al. [57] utilized eye gaze and hand gestures for region selection in AR. Rodríguez et al. [69] enabled artists to draw VR sketches using this dual-modal interaction. Bao et al. [9] explored how hand-eye coordination can facilitate object selection and manipulation in occluded environments. Caillet et al. [58] designed a two-phase interaction technique with gaze and hand gesture to reduce fatigue in 3D selection.

(b) Gaze + Speech. Eye gaze and speech are typically integrated to offer hands-free interaction: gaze provides contextual information, while speech functions as a command or query. Jing et al. [3] visualized shared gaze cues using contextual speech input to enhance the efficiency of XR remote collaboration. When the XR system detects keywords such as "this" or "that", it guides collaborators to follow the speaker's gaze direction. Li et al. [21] proposed a method in AR environments that processes multimodal sensory inputs (visual and auditory) from daily activities using LLMs, predicting potential digital follow-up actions. G-VOILA [96] leveraged gaze data and visual field as contextual information to enhance LLMs' ability to respond to users' daily speech queries. In addition to serving as inputs, visual and audio modalities can also function as outputs of XR systems to enhance immersion. Lee et al. [117] designed a multi-modal attention guidance system that utilizes visual cues (lighting effects) and auditory cues (spatial audio) in VR social group conversations, significantly improving users' response times and conversation satisfaction during turn-taking interaction.

(c) Gesture + Speech. Gesture and speech are the two most commonly used modalities in delivering presentations. Therefore, they are naturally integrated for augmented presentations. For example, Liao et al. [45] proposed augmented live presentations using visuals and animation for expressive storytelling. To achieve this, they used speech recognition combined with keyword extraction. These keywords are mapped to predefined display contents, such as images and videos, which are displayed on the user's hand in real time. However, this method presents inferior visual artifacts due to imperfect system recognition [17]. Cao et al. [17] predefined the layouts and effects of graphics and used a set of rules to allow speech, gesture, and graphics to be elastically connected for visual quality. Williams et al. [76] investigated the effect of referent display methods (text vs. animation) on gesture and speech elicitation. In terms of animated referents, they found users elicited interaction proposals with gesture + speech and suffered lower workload than speech-only interactions.

(d) Gaze + Gesture + Speech. Several studies have also explored the integration of gaze, gesture, and speech to leverage the benefits of all three modalities. Chen et al. [59] proposed the Compass+Ring menu that integrates three modalities: gaze to point at menu items, speech to confirm selections, and gestures to adjust parameters by rotating the wrist or pinching fingers to navigate menu levels. The Ring menu allows for quick adjustments using simple gestures like "putting on" and "rotating" a virtual ring. Lee et al. [97] leveraged eye gaze, hand pointing and conversation history to achieve pronoun disambiguation in users' speech. When the user queries with pronouns, the proposed system replace them with identified objects and text within the field-of-view (FoV). This approach enables the voice assistant to provide more accurate responses by incorporating contextual information.

(e) X + Y. In the reviewed literature, numerous studies explore multimodal interactions beyond the modalities discussed previously, which here are referred to "X + Y". Due to space constraints, only three types of them are presented in this section:

- X + head pose.

Several studies focus on integrating head pose with other modalities [48,60,61,71,83]. For example, Clenchclick [48] explored the combination of head movements and teeth clenching for target selection in AR.

- X + facial expression.

Some other studies emphasize combining facial expressions with additional modalities [46,70]. GazePuffer [70], for instance, introduced cheek puffing gestures combined with gaze to offer an innovative hands-free interaction.

- Multimodal Comparison.

Additionally, some researches compare unimodal and multimodal interaction techniques to determine which modalities are better suited for specific tasks or scenarios [47,49,62]. In terms of the text selection task in VR, Meng et al. [49] designed three hands-free selection mechanisms, including dwell, eye Blinks and voice (hum) to complement head-based pointing. Yan et al. [62] introduced ConeSpeech, a VR-based directional speech interaction method allowing users to selectively communicate with target listeners, minimizing disturbance to bystanders. The authors compare five control modalities, i.e., head, gaze, torso, hand, and controller, and ultimately select head orientation for its balance of speed, accuracy, and intuitiveness.

## 4 Discussion and recommendations

Based on the previous taxonomy and analysis of natural interaction techniques in wearable XR, this section focuses on the challenges these techniques face and offers recommendations for future research directions. Objectively, researchers aim to develop more accurate and efficient natural interaction techniques. Subjectively, the goal is to provide more comfortable and immersive interaction experiences. Additionally, to enhance the usability of these techniques, researchers should apply them in real-world scenarios to improve the overall user experience in XR.

Accordingly, the discussion is organized into three main areas. Subsection 4.1 presents recommendations for toward more accurate and reliable natural interactions in XR. Subsection 4.2 discusses recommendations for toward more natural, comfortable, and immersive interactions in XR. Lastly, Subsection 4.3 offers suggestions for bridging interaction design and practical XR applications.

### 4.1 Toward more accurate and reliable natural interactions in XR

As above section mentioned, current natural interaction such as gaze, gesture, and speech, still face the insufficient accuracy during interaction, especially in the complex situations, e.g., outdoors and walking. For example, for reducing the effects of Midas touch problem on gaze-only interaction, Subsection 3.3.2 mentions many researchers that optimize the virtual menu layouts [39,41], or require extra eye movements to confirm the selections [40,65]. For the false hand gesture recognition caused by occlusion, researchers usually require users to manipulate objects in certain postures and orientations to avoid occlusion [52,72], or use the wearable device based hand-tracking [86]. While these methods have shown some success in improving interaction accuracy, there is still significant room for advancement. We propose that the future of achieving more accurate XR interaction lies in two promising research directions: multimodal interaction and error recovery mechanisms.

**Multimodal interaction.** Integrating multiple modalities into a unified interaction system significantly enhances the accuracy and reliability of user inputs. For instance, Bao et al. addressed the issue of low gaze-pointing accuracy by allowing the hand to refine the pointing direction, thereby facilitating target selection [9]. Moreover, combining gaze with hand gesture interaction follows the principle of "gaze select, hand manipulate" [20,56]. This approach reduces hand operations and thus decreases the likelihood of hand recognition error. Multimodal integration can also resolve ambiguities in speech. For instance, Lee et al. used gaze or hand pointing to identify specific objects, clarifying vague verbal descriptions such as "What is this?" [97]. This highlights the effectiveness of multimodal systems in improving interaction accuracy and reducing ambiguity.

**Error recovery.** Error recovery mechanisms are crucial for ensuring robust and user-friendly interactions in XR environments. Despite advancements in tracking accuracy, unintended actions or misrecognitions such as clicking the wrong button due to inattention, remain inevitable. For example, Sendhilnathan et al. categorized gesture-based interaction events into three types: correctly recognized input actions, input recognition errors, and user errors, observing that these categories were consistent across tasks. They then applied a deep learning method to differentiate these events using only eye movement input, achieving promising results [44]. Sidenmark et al. designed an error-aware mechanism that adaptively switches to fallback modalities (e.g., head pointing or a controller) when errors or noise occur during gaze interactions. They also adjusted the weighting between the gaze modality and fallback options based on the error ratio [19]. These approaches highlight the vital role of error recovery in enhancing the reliability and overall usability of XR systems.

### 4.2 Toward more natural, comfortable and immersive interactions in XR

Interaction techniques aim not only for accuracy but also for subjective factors such as comfort, immersion, and usability. In XR environments, the subjective experience of a particular method may be more important to certain users than objective metrics (e.g., temporal performance) [78]. In the reviewed literature, over 50 studies assess users' subjective experiences with interaction techniques through measures like task load, immersion, preference, and usability. These assessments are typically conducted via post-study questionnaires, such as the NASA Task Load Index (NASA-TLX), System Usability Scale (SUS), and Simulator Sickness Questionnaire (SSQ). This emphasizes the importance of user-friendliness as a critical quality of interaction techniques. We anticipate that future XR interactions will continue to prioritize improvements in reducing task load, enhancing immersion, and improving subtlety.

**Reduce task load.** Physical and cognitive load are critical factors in user's interactive experience. The NASA-TLX is the most widely used tool for assessing the task load of interaction techniques. As discussed in Subsection 3.3, each modality has its own limitations in terms of task load: hand-gesture-only interactions can lead to arm fatigue [4]; speech-only interactions increase cognitive load due to the requirement of memorizing keywords [59]; prolonged use of gaze-only interactions can also cause eye fatigue [66]. Our review identifies several studies focusing on reducing users' burden. One common approach is to effectively integrate multiple modalities. For instance, Bao et al. [9] proposed a better method of combining gaze and hand gestures, significantly alleviating arm fatigue. Compass+Ring [59] introduces a multimodal menu integrating gaze, speech, and gesture to mitigate eye fatigue. Additionally, some studies innovate new devices and modalities, offering more effortless interactions such as finger gestures. AO-finger [103] introduces a wristband that recognizes fine-grained finger gestures that require little exertion. We believe that reducing task load will remain a key focus in future interaction design.

**Enhance immersion.** Immersion is a critical component of the XR experience and is one of the 3Is of VR [143]. Interaction, a bidirectional process, also aims to provide users with an immersive experience. Numerous studies mainly focus on offering feedback with different devices and technologies to enhance the immersion of XR. For instance, Jang et al. [144] utilized ultrasonic devices to deliver mid-air haptic feedback, allowing users to touch and explore volume-rendered hologram with their bare hands. Saint-Aubert et al. [125] investigated tactile vibrations speech from users or virtual avatars to enhance persuasiveness, co-presence, and leadership. Besides, several studies have proposed novel well-designed interaction techniques to enhance users' immersion. Illumotion [80] is an example to increase presence and reduce cybersickness by designing a hand-gesture-based interaction

technique inspired by photo manipulation. This suggests that future research will continue to focus on enhancing immersion in XR environments.

**Improve subtlety.** For XR devices to be integrated into daily life, interaction techniques must be adaptable across various scenarios without raising concerns. In public environments, conspicuous interactions may lead to social awkwardness and privacy challenges [105]. Thus, enhancing the subtlety of interaction techniques is crucial for improving social acceptance. Some existing methods, such as hand gestures and speech interactions, have been criticized for their lack of subtlety. Hand gestures are restricted by the FoV of HMD cameras [103], while speech interactions may disrupt the experiences of others [81]. Recent research has increasingly focused on developing more discreet interaction techniques. Tactile interactions, including micro finger gestures [43,103–105] and wearable devices [50,86], are considered to offer greater subtlety and social acceptance compared to hand gestures. Additionally, to enable speech interactions in noisy or speech-unfriendly environments, several studies have proposed methods for silent speech recognition [90,93,140]. Therefore, improving the subtlety of interaction techniques will remain a prominent and ongoing topic in XR research.

### 4.3 Empowering multimodal XR natural interaction with AI and LLMs

Recent advancements in AI and LLMs have significantly expanded the potential of natural interactions in XR environments. By enhancing input recognition, enabling semantic reasoning, and facilitating multimodal integration, these technologies improve the efficiency, intuitiveness, and usability of XR systems. This section discusses the roles of AI and LLMs, supported by recent research, and explores potential future directions for their application.

AI enhances input recognition and contextual understanding. AI plays a pivotal role in addressing challenges related to multimodal input recognition and resolving ambiguities in XR interactions. Machine learning models have significantly improved input accuracy by combining diverse data sources. For example, gaze direction and head orientation are integrated to predict user targets in complex scenes [60]. Similarly, error-aware systems powered by deep learning can identify user mistakes or inconsistencies during interactions and adaptively switch to alternative modalities, such as voice commands or hand gestures [19,44]. Beyond input recognition, AI enhances contextual understanding by integrating sensory data. For instance, fine-grained gesture recognition systems leveraging visual and audio sensors enable precise interactions even in occluded or noisy environments [103]. In lifelogging applications, AI models can process egocentric video data to automatically segment and tag key events based on temporal and spatial patterns [145]. These advancements increase the reliability and robustness of XR systems, making them more responsive and adaptable to user needs.

LLMs enable semantic reasoning and natural language interaction. LLMs bring advanced semantic reasoning capabilities to XR systems, enabling more natural, intuitive, and flexible user interactions. Unlike traditional keyword-based approaches, LLMs can process open-ended user commands and resolve ambiguous references by incorporating contextual cues, such as gaze direction, gestures, or spatial information. For example, recent systems have demonstrated how LLMs can resolve pronoun ambiguities in speech by linking them to objects within the user's field of view [96,97]. Beyond disambiguation, LLMs empower non-technical users to modify virtual environments dynamically through natural language commands [12]. Additionally, they enhance interactivity by supporting intelligent virtual assistants that integrate multimodal inputs, such as speech and gaze, to deliver context-aware and reasoning-driven responses [81]. These capabilities make XR systems more intuitive, accessible, and adaptable to diverse user needs.

Looking forward, the integration of AI and LLMs presents transformative opportunities for developing adaptive, user-centric XR systems. A key area of future exploration involves leveraging AI to dynamically model user behavior and preferences, enabling systems to personalize interactions over time. For example, advancements in lifelogging technologies, such as egocentric vision models and wearable AR devices, facilitate the continuous capture and encoding of users' daily experiences, including what they see, hear, and do [97]. By combining AI for event detection with LLMs for semantic reasoning, future systems could enable effortless memory retrieval. Users could query their past experiences with questions such as, "What did I read during yesterday's meeting?" or "Who did I talk to at lunch?" These capabilities align with the broader objective of creating more intuitive, human-centric, and context-aware XR systems.

### 4.4 Bridging interaction design and practical XR applications

As discussed in Subsection 3.1, nearly 70% of the research focuses on the design of interaction techniques without specifying their application in concrete scenarios. While these techniques could potentially be adapted to various applications, they often require significant modifications for specific contexts. This gap presents challenges that limit the widespread adoption of natural interaction techniques in XR environments. Therefore, we argue that researchers should focus more on application scenarios when developing natural interaction techniques.

- Existing applications of natural interaction techniques. Subsection 3.1 highlights over 20 papers that explore natural interaction techniques in different application contexts, such as sketching [43,52], virtual meetings [117,125], XR navigation [4,81], reading [38,49], and maintenance [4]. These applications have significant potential for further expansion. For example, in sketching, combining brain-computer interfaces (BCIs) with eye-tracking technology could allow users to control the shape and color of design objects directly through thought, rather than relying solely on gestures or voice commands. In virtual meetings, real-time emotion recognition technology can capture participants' facial expressions, tone of voice, and heart

rate to generate personalized avatars.

- Natural interaction techniques can also be applied to new domains. Beyond the previously mentioned applications, in healthcare, these techniques offer more convenient care by integrating eye-tracking, gestures, and voice commands. For instance, patients wearing lightweight AR headsets can use eye-tracking to control devices like lighting or the TV, and adjust the bed angle with voice commands. In entertainment, natural interaction transforms experiences by allowing players to interact with virtual characters through gestures and speech, while the game adjusts its environment or difficulty based on the player's emotions in real time.
- Application beyond the lab. Currently, most studies on natural interaction techniques are confined to laboratory or controlled environments. Only a limited number of studies explore their use in more complex, real-world contexts such as shopping malls or city streets [21,97]. To enable the broader adoption of XR technologies, it is crucial to design natural interaction techniques that can address the challenges posed by these complex, uncontrolled environments.
- Breaking down barriers between applications. Existing research often develops interaction modalities that are specific to each application, which increases the learning curve for users. If interactions across different XR applications are standardized, similar to the unified operation model of smartphone apps, it would greatly reduce the learning curve for users. This standardization could, in turn, encourage wider adoption of these technologies.

## 5 Limitations and future work

While this review of recent papers from top venues offers valuable insights into the latest trends in XR natural interaction techniques, there are several limitations that need to be addressed in future work.

Firstly, due to the vast amount of relevant research, we limited our scope to publications from six major venues since 2022. However, other venues, such as those within the ACM and IEEE digital libraries, also contain valuable research, which is not included. Additionally, extending the review to earlier years could provide a clearer understanding of the broader development trends in the field.

Furthermore, in this paper, we categorized the collected literature into four main categories: application context, operation types, performance measures, and interaction modalities. Future work could explore additional categories, such as study types or use cases [1], to provide a more detailed understanding.

Finally, our review primarily focuses on papers that explicitly mention natural interaction techniques for wearable XR. However, many studies on non-wearable natural interaction could potentially be adapted for wearable XR. For example, Ahmad Khan et al. [146] explored the synchronization between gaze and speech to implicitly link voice notes with digital text content. Although their study was conducted in a desktop environment, this approach could be applied to wearable XR systems as well.

## 6 Conclusion

In this paper, we reviewed research papers on natural interaction techniques for wearable XR, published since 2022 in six top venues. We categorized this literature based on application context, operation types, performance measures, and interaction modalities. Specifically, we classified operation types into seven categories, distinguishing between active and passive interactions. Interaction modalities are further broken down into nine distinct types. In addition, we presented statistical analyses of advanced natural interaction techniques. Building on these insights, we identified key challenges in natural interaction systems and suggested potential avenues for future research. This review offers valuable insights for researchers aiming to design natural and efficient interaction systems for XR.

**Acknowledgements** This work was supported by Beijing Natural Science Foundation (L242019).

**Competing interests** The authors declare that they have no competing interests or financial conflicts to disclose.



## References

1. Spittle B, Frutos-Pascual M, Creed C, Williams I. A review of interaction techniques for immersive environments. IEEE Transactions on Visualization and Computer Graphics, 2023, 29(9): 3900−3921
2. Aros M, Tyger C L, Chaparro B S. Unraveling the meta quest 3: an out-of-box experience of the future of mixed reality headsets. In: Proceedings of the 26th International Conference on Human-Computer Interaction. 2024, 3−8
3. Jing A, Lee G A, Billinghurst M. Using speech to visualise shared gaze cues in MR remote collaboration. In: Proceedings of 2022 IEEE Conference on Virtual Reality and 3D User Interfaces. 2022, 250−259
4. Quere C, Menin A, Julien R, Wu H Y, Winckler M. HandyNotes: using the hands to create semantic representations of contextually aware real-world objects. In: Proceedings of 2024 IEEE Conference Virtual Reality and 3D User Interfaces. 2024, 265−275
5. Barteit S, Lanfermann L, Bärnighausen T, Neuhann F, Beiersmann C. Augmented, mixed, and virtual reality-based head-mounted devices for medical education: systematic review. JMIR Serious Games, 2021, 9(3): e29080
6. Hrycak C, Lewakis D, Krüger J. Investigating the apple vision pro spatial computing platform for GPU-based volume visualization. In: Proceedings of 2024 IEEE Visualization and Visual Analytics. 2024, 181−185
7. Yenduri G, M R, Maddikunta P K R, Gadekallu T R, Jhaveri R H,

Bandi A, Chen J, Wang W, Shirawalmath A A, Ravishankar R, Wang W. Spatial computing: concept, applications, challenges and future directions. 2024, arXiv preprint arXiv: 2402.07912

8. Hackl C, Cronin I. Spatial Computing: An AI-Driven Business Revolution. Hoboken: John Wiley & Sons, 2024
9. Bao Y, Wang J, Wang Z, Lu F. Exploring 3D interaction with gaze guidance in augmented reality. In: Proceedings of 2023 IEEE Conference Virtual Reality and 3D User Interfaces. 2023, 22−32
10. Bérard F, Ip J, Benovoy M, El-Shimy D, Blum J R, Cooperstock J R. Did ”minority report” get it wrong? Superiority of the mouse over 3D input devices in a 3D placement task. In: Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction. 2009, 400−414
11. Matthews B J, Thomas B H, Von Itzstein G S, Smith R T. Shape aware haptic retargeting for accurate hand interactions. In: Proceedings of 2022 IEEE Conference on Virtual Reality and 3D User Interfaces. 2022, 625−634
12. Giunchi D, Numan N, Gatti E, Steed A. DreamCodeVR: towards democratizing behavior design in virtual reality with speech-driven programming. In: Proceedings of 2024 IEEE Conference Virtual Reality and 3D User Interfaces. 2024, 579−589
13. Huang Y, Yang L, Chen G, Zhang H, Lu F, Sato Y. Matching compound prototypes for few-shot action recognition. International Journal of Computer Vision, 2024, 132(9): 3977−4002
14. Wang Z, Gu X, Lu F. DEAMP: dominant-eye-aware foveated rendering with multi-parameter optimization. In: Proceedings of 2023 IEEE International Symposium on Mixed and Augmented Reality. 2023, 632−641
15. Chaconas N, Höllerer T. An evaluation of bimanual gestures on the Microsoft HoloLens. In: Proceedings of 2018 IEEE Conference on Virtual Reality and 3D User Interfaces. 2018, 33−40
16. Hincapié-Ramos J D, Guo X, Moghadasian P, Irani P. Consumed endurance: a metric to quantify arm fatigue of mid-air interactions. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2014, 1063−1072
17. Cao Y, Kazi R H, Wei L Y, Aneja D, Xia H. Elastica: adaptive live augmented presentations with elastic mappings across modalities. In: Proceedings of 2024 CHI Conference on Human Factors in Computing Systems. 2024, 599
18. De La Torre F, Fang C M, Huang H, Banburski-Fahey A, Fernandez J A, Lanier J. LLMR: real-time prompting of interactive worlds using large language models. In: Proceedings of 2024 CHI Conference on Human Factors in Computing Systems. 2024, 600
19. Sidenmark L, Parent M, Wu C H, Chan J, Glueck M, Wigdor D, Grossman T, Giordano M. Weighted pointer: error-aware gaze-based interaction through fallback modalities. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(11): 3585−3595
20. Lystbæk M N, Rosenberg P, Pfeuffer K, Grønbæk J E, Gellersen H. Gaze-hand alignment: combining eye gaze and mid-air pointing for interacting with menus in augmented reality. Proceedings of the ACM on Human-Computer Interaction, 2022, 6(ETRA): 145
21. Li J N, Xu Y, Grossman T, Santosa S, Li M. OmniActions: predicting digital actions in response to real-world multimodal sensory inputs with LLMs. In: Proceedings of 2024 CHI Conference on Human Factors in Computing Systems. 2024
22. Wang H, Dong X, Chen Z, Shi B E. Hybrid gaze/EEG brain computer interface for robot arm control on a pick and place task. In: Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2015, 1476−1479
23. Tran T T M, Brown S, Weidlich O, Billinghurst M, Parker C. Wearable augmented reality: research trends and future directions from three major venues. IEEE Transactions on Visualization and Computer Graphics, 2023, 29(11): 4782−4793
24. Hertel J, Karaosmanoglu S, Schmidt S, Bräker J, Semmann M, Steinicke F. A taxonomy of interaction techniques for immersive augmented reality based on an iterative literature review. In: Proceedings of 2021 IEEE International Symposium on Mixed and Augmented Reality. 2021, 431−440
25. Pirker J, Dengel A. The potential of 360° virtual reality videos and real VR for education—A literature review. IEEE Computer Graphics and Applications, 2021, 41(4): 76−89
26. Katona J. A review of human−computer interaction and virtual reality research fields in cognitive InfoCommunications. Applied Sciences, 2021, 11(6): 2646
27. Zhang Y, Wang Z, Zhang J, Shan G, Tian D. A survey of immersive visualization: focus on perception and interaction. Visual Informatics, 2023, 7(4): 22−35
28. Guan H, Song C, Zhang Z. GRAMO: geometric resampling augmentation for monocular 3D object detection. Frontiers of Computer Science, 2024, 18(5): 185706
29. Liu F, Zheng Z, Shi Y, Tong Y, Zhang Y. A survey on federated learning: a perspective from multi-party computation. Frontiers of Computer Science, 2024, 18(1): 181336
30. Tang J, Song R, Huang Y, Gao S, Yu Z. Semantic-aware entity alignment for low resource language knowledge graph. Frontiers of Computer Science, 2024, 18(4): 184319
31. Li J, Lei Y, Bian Y, Cheng D, Ding Z, Jiang C. RA-CFGPT: Chinese financial assistant with retrieval-augmented large language model. Frontiers of Computer Science, 2024, 18(5): 185350
32. Lu F, Zhao Q. Towards cobodied/symbodied AI: concept and eight scientific and technical problems. SCIENTIA SINICA Informationis, 2025, 55(2): 444−448
33. Ghamandi R K, Hmaiti Y, Nguyen T T, Ghasemaghaei A, Kattoju R K, Taranta E M, LaViola J J. What and how together: a taxonomy on 30 years of collaborative human-centered XR tasks. In: Proceedings of 2023 IEEE International Symposium on Mixed and Augmented Reality. 2023, 322−335
34. Chowdhury S, Ullah A K M A, Pelmore N B, Irani P, Hasan K. WriArm: leveraging wrist movement to design wrist+arm based teleportation in VR. In: Proceedings of 2022 IEEE International Symposium on Mixed and Augmented Reality. 2022, 317−325
35. Schmitz M, Günther S, Schön D, Müller F. Squeezy-feely: investigating lateral thumb-index pinching as an input modality. In: Proceedings of 2022 CHI Conference on Human Factors in Computing Systems. 2022, 61
36. Ban R, Matsumoto K, Narumi T, Kuzuoka H. Wormholes in VR: teleporting hands for flexible passive haptics. In: Proceedings of 2022 IEEE International Symposium on Mixed and Augmented Reality. 2022, 748−757
37. Yu D, Zhou Q, Dingler T, Velloso E, Gonçalves J. Blending on-body and mid-air interaction in virtual reality. In: Proceedings of 2022 IEEE International Symposium on Mixed and Augmented Reality. 2022, 637−646
38. Lee G, Healey J, Manocha D. VRDoc: gaze-based interactions for VR reading experience. In: Proceedings of 2022 IEEE International Symposium on Mixed and Augmented Reality. 2022, 787−796
39. Choi M, Sakamoto D, Ono T. Kuiper belt: utilizing the ”out-of-natural angle” region in the eye-gaze interaction for virtual reality. In: Proceedings of 2022 CHI Conference on Human Factors in Computing Systems. 2022, 357
40. Kim T, Ham A, Ahn S, Lee G. Lattice menu: a low-error gaze-based marking menu utilizing target-assisted gaze gestures on a lattice of visual anchors. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022, 277
41. Yi X, Lu Y, Cai Z, Wu Z, Wang Y, Shi Y. GazeDock: gaze-only menu selection in virtual reality using auto-triggering peripheral menu. In: Proceedings of 2022 IEEE Conference on Virtual Reality and 3D User Interfaces. 2022, 832−842

42. Wang Z, Zhao Y, Lu F. Gaze-vergence-controlled see-through vision in augmented reality. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(11): 3843−3853
43. Chen T, Li T, Yang X, Zhu K. EFRing: enabling thumb-to-index-finger microgesture interaction through electric field sensing using single smart ring. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2022, 6(4): 161
44. Sendhilnathan N, Zhang T, Lafreniere B, Grossman T, Jonker T R. Detecting input recognition errors and user errors using gaze dynamics in virtual reality. In: Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology. 2022, 38
45. Liao J, Karim A, Jadon S S, Kazi R H, Suzuki R. RealityTalk: real-time speech-driven augmented presentation for AR live storytelling. In: Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology. 2022, 17
46. Yi X, Qiu L, Tang W, Fan Y, Li H, Shi Y. DEEP: 3D gaze pointing in virtual reality leveraging eyelid movement. In: Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology. 2022, 3
47. Xu W, Meng X, Yu K, Sarcar S, Liang H N. Evaluation of text selection techniques in virtual reality head-mounted displays. In: Proceedings of 2022 IEEE International Symposium on Mixed and Augmented Reality. 2022, 131−140
48. Shen X, Yan Y, Yu C, Shi Y. ClenchClick: hands-free target selection method leveraging teeth-clench for augmented reality. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2022, 6(3): 139
49. Meng X, Xu W, Liang H N. An exploration of hands-free text selection for virtual reality head-mounted displays. In: Proceedings of 2022 IEEE International Symposium on Mixed and Augmented Reality. 2022, 74−81
50. Das S, Nasser A, Hasan K. FingerButton: enabling controller-free transitions between real and virtual environments. In: Proceedings of 2023 IEEE International Symposium on Mixed and Augmented Reality. 2023, 533−542
51. Zhu F, Sidenmark L, Sousa M, Grossman T. PinchLens: applying spatial magnification and adaptive control-display gain for precise selection in virtual reality. In: Proceedings of 2023 IEEE International Symposium on Mixed and Augmented Reality. 2023, 1221−1230
52. Song Z, Dudley J J, Kristensson P O. HotGestures: complementing command selection and use with delimiter-free gesture-based shortcuts in virtual reality. IEEE Transactions on Visualization and Computer Graphics, 2023, 29(11): 4600−4610
53. Tseng W J, Huron S, Lecolinet E, Gugenheimer J. FingerMapper: mapping finger motions onto virtual arms to enable safe virtual reality interaction in confined spaces. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 2023, 874
54. Chen Y S, Hsieh C E, Jie M T Y, Han P H, Hung Y P. Leap to the eye: implicit gaze-based interaction to reveal invisible objects for virtual environment exploration. In: Proceedings of 2023 IEEE International Symposium on Mixed and Augmented Reality. 2023, 214−222
55. Sidenmark L, Clarke C, Newn J, Lystbæk M N, Pfeuffer K, Gellersen H. Vergence matching: inferring attention to objects in 3D environments for gaze-assisted selection. In: Proceedings of 2023 CHI Conference on Human Factors in Computing Systems. 2023, 257
56. Wagner U, Lystbæk M N, Manakhov P, Grønbæk J E S, Pfeuffer K, Gellersen H. A fitts' law study of gaze-hand alignment for selection in 3D user interfaces. In: Proceedings of 2023 CHI Conference on Human Factors in Computing Systems. 2023, 252
57. Shi R, Wei Y, Qin X, Hui P, Liang H N. Exploring gaze-assisted and hand-based region selection in augmented reality. Proceedings of the ACM on Human-Computer Interaction, 2023, 7(ETRA): 160
58. Caillet A C, Goguey A, Nigay L. 3D selection in mixed reality: designing a two-phase technique to reduce fatigue. In: Proceedings of 2023 IEEE International Symposium on Mixed and Augmented Reality. 2023, 800−809
59. Chen X, Guo D, Feng L, Chen B, Liu W. Compass+ring: a multimodal menu to improve interaction performance and comfortability in one-handed scenarios. In: Proceedings of 2023 IEEE International Symposium on Mixed and Augmented Reality. 2023, 473−482
60. Wei Y, Shi R, Yu D, Wang Y, Li Y, Yu L, Liang H N. Predicting gaze-based target selection in augmented reality headsets based on eye and head endpoint distributions. In: Proceedings of 2023 CHI Conference on Human Factors in Computing Systems. 2023, 283
61. Hou B J, Newn J, Sidenmark L, Khan A A, Bækgaard P, Gellersen H. Classifying head movements to separate head-gaze and head gestures as distinct modes of input. In: Proceedings of 2023 CHI Conference on Human Factors in Computing Systems. 2023, 253
62. Yan Y, Liu H, Shi Y, Wang J, Guo R, Li Z, Xu X, Yu C, Wang Y, Shi Y. ConeSpeech: exploring directional speech interaction for multi-person remote communication in virtual reality. IEEE Transactions on Visualization and Computer Graphics, 2023, 29(5): 2647−2657
63. Sindhupathiraja S R, Ullah A K M A, Delamare W, Hasan K. Exploring bi-manual teleportation in virtual reality. In: Proceedings of 2024 IEEE Conference Virtual Reality and 3D User Interfaces. 2024, 754−764
64. Dupré C, Appert C, Rey S, Saidi H, Pietriga E. TriPad: touch input in AR on ordinary surfaces with hand tracking only. In: Proceedings of 2024 CHI Conference on Human Factors in Computing Systems. 2024, 754
65. Orlosky J, Liu C, Sakamoto K, Sidenmark L, Mansour A. EyeShadows: peripheral virtual copies for rapid gaze selection and interaction. In: Proceedings of 2024 IEEE Conference Virtual Reality and 3D User Interfaces. 2024, 681−689
66. Zhang C, Chen T, Shaffer E, Soltanaghai E. FocusFlow: 3D gaze-depth interaction in virtual reality leveraging active visual depth manipulation. In: Proceedings of 2024 CHI Conference on Human Factors in Computing Systems. 2024, 372
67. Turkmen R, Gelmez Z E, Batmaz A U, Stuerzlinger W, Asente P, Sarac M, Pfeuffer K, Machuca M D B. EyeGuide & EyeConGuide: gaze-based visual guides to improve 3D sketching systems. In: Proceedings of 2024 CHI Conference on Human Factors in Computing Systems. 2024, 178
68. Zenner A, Karr C, Feick M, Ariza O, Krüger A. Beyond the blink: investigating combined saccadic & blink-suppressed hand redirection in virtual reality. In: Proceedings of 2024 CHI Conference on Human Factors in Computing Systems. 2024, 750
69. Rodriguez R, Sullivan B T, Barrera Machuca M D, Batmaz A U, Tornatzky C, Ortega F R. An artists' perspectives on natural interactions for virtual reality 3D sketching. In: Proceedings of 2024 CHI Conference on Human Factors in Computing Systems. 2024, 163
70. Lai Y, Sun M, Li Z. GazePuffer: hands-free input method leveraging puff cheeks for VR. In: Proceedings of 2024 IEEE Conference Virtual Reality and 3D User Interfaces. 2024, 331−341
71. Marquardt A, Steininger M, Trepkowski C, Weier M, Kruijff E. Selection performance and reliability of eye and head gaze tracking under varying light conditions. In: Proceedings of 2024 IEEE Conference Virtual Reality and 3D User Interfaces. 2024, 546−556
72. Pei S, Chen A, Lee J, Zhang Y. Hand interfaces: using hands to imitate objects in AR/VR for expressive interactions. In: Proceedings of 2022 CHI Conference on Human Factors in Computing Systems. 2022, 429
73. Satriadi K A, Smiley J, Ens B, Cordeil M, Czauderna T, Lee B, Yang Y, Dwyer T, Jenny B. Tangible globes for data visualisation in augmented reality. In: Proceedings of 2022 CHI Conference on Human Factors in Computing Systems. 2022, 505
74. Xu X, Zhou Y, Shao B, Feng G, Yu C. GestureSurface: VR sketching through assembling scaffold surface with non-dominant hand. IEEE Transactions on Visualization and Computer Graphics, 2023, 29(5): 2499−2507
75. Zhan L, Xiong T, Zhang H, Guo S, Chen X, Gong J, Lin J, Qin Y. TouchEditor: interaction design and evaluation of a flexible touchpad

for text editing of head-mounted displays in speech-unfriendly environments. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2023, 7(4): 198

76. Williams A S, Ortega F R. The impacts of referent display on gesture and speech elicitation. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(11): 3885−3895
77. Deng C L, Sun L, Zhou C, Kuai S G. Dual-gain mode of head-gaze interaction improves the efficiency of object positioning in a 3D virtual environment. International Journal of Human−Computer Interaction, 2024, 40(8): 2067−2082
78. Hombeck J, Voigt H, Heggemann T, Datta R R, Lawonn K. Tell me where to go: voice-controlled hands-free locomotion for virtual reality systems. In: Proceedings of 2023 IEEE Conference Virtual Reality and 3D User Interfaces. 2023, 123−134
79. Mortezapoor S, Vasylevska K, Vonach E, Kaufmann H. CoboDeck: a large-scale haptic VR system using a collaborative mobile robot. In: Proceedings of 2023 IEEE Conference Virtual Reality and 3D User Interfaces. 2023, 297−307
80. Sin Z P T, Jia Y, Li R C, Va Leong H, Li Q, Ng P H F. Illumotion: an optical-illusion-based VR locomotion technique for long-distance 3D movement. In: Proceedings of 2024 IEEE Conference Virtual Reality and 3D User Interfaces. 2024, 924−934
81. Wang Z, Yuan L P, Wang L, Jiang B, Zeng W. VirtuWander: enhancing multi-modal interaction for virtual tour guidance through large language models. In: Proceedings of 2024 CHI Conference on Human Factors in Computing Systems. 2024, 612
82. Kang S, Jeong J, Lee G A, Kim S H, Yang H J, Kim S. The RayHand navigation: a virtual navigation method with relative position between hand and gaze-ray. In: Proceedings of 2024 CHI Conference on Human Factors in Computing Systems. 2024, 634
83. Lee H S, Weidner F, Sidenmark L, Gellersen H. Snap, pursuit and gain: virtual reality viewport control by gaze. In: Proceedings of 2024 CHI Conference on Human Factors in Computing Systems. 2024, 375
84. Stemasov E, Wagner T, Gugenheimer J, Rukzio E. ShapeFindAR: exploring in-situ spatial search for physical artifact retrieval using mixed reality. In: Proceedings of 2022 CHI Conference on Human Factors in Computing Systems. 2022, 292
85. Song Z, Dudley J J, Kristensson P O. Efficient special character entry on a virtual keyboard by hand gesture-based mode switching. In: Proceedings of 2022 IEEE International Symposium on Mixed and Augmented Reality. 2022, 864−871
86. He Z, Lutteroth C, Perlin K. TapGazer: text entry with finger tapping and gaze-directed word selection. In: Proceedings of 2022 CHI Conference on Human Factors in Computing Systems. 2022, 337
87. Shen J, Dudley J J, Kristensson P O. Fast and robust mid-air gesture typing for AR headsets using 3D trajectory decoding. IEEE Transactions on Visualization and Computer Graphics, 2023, 29(11): 4622−4632
88. Zhao M, Pierce A M, Tan R, Zhang T, Wang T, Jonker T R, Benko H, Gupta A. Gaze speedup: eye gaze assisted gesture typing in virtual reality. In: Proceedings of the 28th International Conference on Intelligent User Interfaces. 2023, 595−606
89. Cui W, Liu R, Li Z, Wang Y, Wang A, Zhao X, Rashidian S, Baig F, Ramakrishnan I V, Wang F, Bi X. GlanceWriter: writing text by glancing over letters with gaze. In: Proceedings of 2023 CHI Conference on Human Factors in Computing Systems. 2023, 719
90. Zhang R, Li K, Hao Y, Wang Y, Lai Z, Guimbretière F, Zhang C. EchoSpeech: continuous silent speech recognition on minimally-obtrusive eyewear powered by acoustic sensing. In: Proceedings of 2023 CHI Conference on Human Factors in Computing Systems. 2023, 852
91. Shen J, Boldu R, Kalla A, Glueck M, Surale H B, Karlson A. RingGesture: a ring-based mid-air gesture typing system powered by a deep-learning word prediction framework. IEEE Transactions on Visualization and Computer Graphics, 2024, 30(11): 7441−7451
92. Hu J, Dudley J J, Kristensson P O. SkiMR: dwell-free eye typing in mixed reality. In: Proceedings of 2024 IEEE Conference Virtual Reality and 3D User Interfaces. 2024, 439−449
93. Cai Z, Ma Y, Lu F. Robust dual-modal speech keyword spotting for XR headsets. IEEE Transactions on Visualization and Computer Graphics, 2024, 30(5): 2507−2516
94. Ren Y, Zhang Y, Liu Z, Xie N. Eye-hand typing: eye gaze assisted finger typing via Bayesian processes in AR. IEEE Transactions on Visualization and Computer Graphics, 2024, 30(5): 2496−2506
95. Jadon S S, Faridan M, Mah E, Vaish R, Willett W, Suzuki R. Augmented conversation with embedded speech-driven on-the-fly referencing in AR. 2024, arXiv preprint arXiv: 2405.18537
96. Wang Z, Shi Y, Wang Y, Yao Y, Yan K, Wang Y, Ji L, Xu X, Yu C. G-VOILA: gaze-facilitated information querying in daily scenarios. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2024, 8(2): 78
97. Lee J, Wang J, Brown E, Chu L, Rodriguez S S, Froehlich J E. GazePointAR: a context-aware multimodal voice assistant for pronoun disambiguation in wearable augmented reality. In: Proceedings of 2024 CHI Conference on Human Factors in Computing Systems. 2024, 408
98. Wang X, Lafreniere B, Zhao J. Exploring visualizations for precisely guiding bare hand gestures in virtual reality. In: Proceedings of 2024 CHI Conference on Human Factors in Computing Systems. 2024, 636
99. Shen J, Dudley J J, Mo G B, Kristensson P O. Gesture spotter: a rapid prototyping tool for key gesture spotting in virtual and augmented reality applications. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(11): 3618−3628
100. Lee C J, Zhang R, Agarwal D, Yu T C, Gunda V, Lopez O, Kim J, Yin S, Dong B, Li K, Sakashita M, Guimbretiere F, Zhang C. EchoWrist: continuous hand pose tracking and hand-object interaction recognition using low-power active acoustic sensing on a wristband. In: Proceedings of 2014 CHI Conference on Human Factors in Computing Systems. 2024, 403
101. Rupp D, GrieBer P, Bonsch A, Kuhlen T W. Authentication in immersive virtual environments through gesture-based interaction with a virtual agent. In: Proceedings of 2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops. 2024, 54−60
102. Liu T, Xiao Y, Hu M, Sha H, Ma S, Gao B, Guo S, Liu Y, Song W. AudioGest: gesture-based interaction for virtual reality using audio devices. IEEE Transactions on Visualization and Computer Graphics, 2025, 31(2): 1569−1581
103. Xu C, Zhou B, Krishnan G, Nayar S K. AO-finger: hands-free fine-grained finger gesture recognition via acoustic-optic sensor fusing. In: Proceedings of 2023 CHI Conference on Human Factors in Computing Systems. 2023, 306
104. Kitamura R, Yamamoto T, Sugiura Y. TouchLog: finger micro gesture recognition using photo-reflective sensors. In: Proceedings of the 2023 International Symposium on Wearable Computers. 2023, 92−97
105. Li T, Liu Y, Ma S, Hu M, Liu T, Song W. NailRing: an intelligent ring for recognizing micro-gestures in mixed reality. In: Proceedings of 2022 IEEE International Symposium on Mixed and Augmented Reality. 2022, 178−186
106. Pöhlmann K M T, Li G, McGill M, Markoff R, Brewster S A. You spin me right round, baby, right round: examining the impact of multi-sensory self-motion cues on motion sickness during a VR reading task. In: Proceedings of 2023 CHI Conference on Human Factors in Computing Systems. 2023, 712
107. Medlar A, Lehtikari M T, Glowacka D. Behind the scenes: adapting cinematography and editing concepts to navigation in virtual reality. In: Proceedings of 2024 CHI Conference on Human Factors in Computing Systems. 2024, 545
108. Feick M, Regitz K P, Tang A, Krüger A. Designing visuo-haptic illusions with proxies in virtual reality: exploration of grasp, movement trajectory and object mass. In: Proceedings of 2022 CHI Conference on Human Factors in Computing Systems. 2022, 635

109. Wu G, Qian J, Quispe S C, Chen S, Rulff J, Silva C. ARTiST: automated text simplification for task guidance in augmented reality. In: Proceedings of 2024 CHI Conference on Human Factors in Computing Systems. 2024, 939
110. Yang J J, Qiu L, Corona-Moreno E A, Shi L, Bui H, Lam M S, Landay J A. AMMA: adaptive multimodal assistants through automated state tracking and user model-directed guidance planning. In: Proceedings of 2024 IEEE Conference Virtual Reality and 3D User Interfaces. 2024, 892−902
111. Elsharkawy A I A M, Ataya A A S, Yeo D, An E, Hwang S, Kim S. SYNC-VR: synchronizing your senses to conquer motion sickness for enriching in-vehicle virtual reality. In: Proceedings of 2024 CHI Conference on Human Factors in Computing Systems. 2024, 257
112. Li Y, Liu Z, Yuan L, Tang H, Fan Y, Xie N. Dynamic scene adjustment mechanism for manipulating user experience in VR. In: Proceedings of 2024 IEEE Conference Virtual Reality and 3D User Interfaces. 2024, 179−188
113. Rasch J, Rusakov V D, Schmitz M, Müller F. Going, going, gone: exploring intention communication for multi-user locomotion in virtual reality. In: Proceedings of 2023 CHI Conference on Human Factors in Computing Systems. 2023, 785
114. Tan F F Y, Xu P, Ram A, Suen W Z, Zhao S, Huang Y, Hurter C. AudioXtend: assisted reality visual accompaniments for audiobook storytelling during everyday routine tasks. In: Proceedings of 2024 CHI Conference on Human Factors in Computing Systems. 2024, 83
115. Wang X, Zhang W, Fu H. A3RT: attention-aware AR teleconferencing with life-size 2.5D video avatars. In: Proceedings of 2024 IEEE Conference Virtual Reality and 3D User Interfaces. 2024, 211−221
116. Tao Y, Lopes P. Integrating real-world distractions into virtual reality. In: Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology. 2022, 5
117. Lee G, Lee D Y, Su G M, Manocha D. “May I speak?”: multi-modal attention guidance in social VR group conversations. IEEE Transactions on Visualization and Computer Graphics, 2024, 30(5): 2287−2297
118. Wang Z, Lu F. Tasks reflected in the eyes: egocentric gaze-aware visual task type recognition in virtual reality. IEEE Transactions on Visualization and Computer Graphics, 2024, 30(11): 7277−7287
119. Shen V, Shultz C, Harrison C. Mouth haptics in VR using a headset ultrasound phased array. In: Proceedings of 2022 CHI Conference on Human Factors in Computing Systems. 2022, 275
120. Tatzgern M, Domhardt M, Wolf M, Cenger M, Emsenhuber G, Dinic R, Gerner N, Hartl A. AirRes mask: a precise and robust virtual reality breathing interface utilizing breathing resistance as output modality. In: Proceedings of 2022 CHI Conference on Human Factors in Computing Systems. 2022, 274
121. Kim M J, Ofek E, Pahud M, Sinclair M J, Bianchi A. Big or small, it’s all in your head: visuo-haptic illusion of size-change using finger-repositioning. In: Proceedings of 2024 CHI Conference on Human Factors in Computing Systems. 2024, 751
122. Yamazaki Y, Hasegawa S. Providing 3D guidance and improving the music-listening experience in virtual reality shooting games using musical vibrotactile feedback. In: Proceedings of 2023 IEEE Conference Virtual Reality and 3D User Interfaces. 2023, 276−285
123. Shen V, Rae-Grant T, Mullenbach J, Harrison C, Shultz C. Fluid reality: high-resolution, untethered haptic gloves using electroosmotic pump arrays. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. 2023, 8
124. Jingu A, Withana A, Steimle J. Double-sided tactile interactions for grasping in virtual reality. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. 2023, 9
125. Saint-Aubert J, Argelaguet F, Macé M, Pacchierotti C, Amedi A, Lécuyer A. Persuasive vibrations: effects of speech-based vibrations on persuasion, leadership, and co-presence during verbal communication in VR. In: Proceedings of 2023 IEEE Conference Virtual Reality and 3D User Interfaces. 2023, 552−560
126. Shi C, Chen J, Liu J, Yang C. Graph foundation model. Frontiers of Computer Science, 2024, 18(6): 186355
127. Guo W, Zhuang F, Zhang X, Tong Y, Dong J. A comprehensive survey of federated transfer learning: challenges, methods and applications. Frontiers of Computer Science, 2024, 18(6): 186356
128. Hedeshy R, Kumar C, Menges R, Staab S. Hummer: text entry by gaze and hum. In: Proceedings of 2021 CHI Conference on Human Factors in Computing Systems. 2021, 741
129. Schneider D, Biener V, Otte A, Gesslein T, Gagel P, Campos C, Pucihar K Č, Kljun M, Ofek E, Pahud M, Kristensson P O, Grubert J. Accuracy evaluation of touch tasks in commodity virtual and augmented reality head-mounted displays. In: Proceedings of 2021 ACM Symposium on Spatial User Interaction. 2021, 7
130. Cai M, Lu F, Sato Y. Generalizing hand segmentation in egocentric videos with uncertainty-guided model adaptation. In: Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, 14392−14401
131. Cai M, Lu F, Gao Y. Desktop action recognition from first-person point-of-view. IEEE Transactions on Cybernetics, 2019, 49(5): 1616−1628
132. Cheng Y, Wang H, Bao Y, Lu F. Appearance-based gaze estimation with deep learning: a review and benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 7509−7528
133. Wang Z, Zhao Y, Liu Y, Lu F. Edge-guided near-eye image analysis for head mounted displays. In: Proceedings of 2021 IEEE International Symposium on Mixed and Augmented Reality. 2021, 11−20
134. Guestrin E D, Eizenman M. General theory of remote gaze estimation using the pupil center and corneal reflections. IEEE Transactions on Biomedical Engineering, 2006, 53(6): 1124−1133
135. Wu Z, Rajendran S, van As T, Zimmermann J, Badrinarayanan V, Rabinovich A. MagicEyes: a large scale eye gaze estimation dataset for mixed reality. 2020, arXiv preprint arXiv: 2003.08806
136. Santini T, Niehorster D C, Kasneci E. Get a grip: slippage-robust and glint-free gaze estimation for real-time pervasive head-mounted eye tracking. In: Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications. 2019, 17
137. Narcizo F B, de Queiroz J E R, Gomes H M. Remote eye tracking systems: technologies and applications. In: Proceedings of the 26th Conference on Graphics, Patterns and Images Tutorials. 2013, 15−22
138. Dierkes K, Kassner M, Bulling A. A fast approach to refraction-aware eye-model fitting and gaze prediction. In: Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications. 2019, 23
139. Kytö M, Ens B, Piumsomboon T, Lee G A, Billinghurst M. Pinpointing: precise head- and eye-based target selection for augmented reality. In: Proceedings of 2018 CHI Conference on Human Factors in Computing Systems. 2018, 81
140. Wang X, Su Z, Rekimoto J, Zhang Y. Watch your mouth: silent speech recognition with depth sensing. In: Proceedings of 2024 CHI Conference on Human Factors in Computing Systems. 2024, 323
141. Wang Z, Sun J, Hu M, Rao M, Song W, Lu F. GazeRing: enhancing hand-eye coordination with pressure ring in augmented reality. In: Proceedings of 2024 IEEE International Symposium on Mixed and Augmented Reality. 2024, 534−543
142. Wang Z, Wang H, Yu H, Lu F. Interaction with gaze, gesture, and speech in a flexibly configurable augmented reality system. IEEE Transactions on Human-Machine Systems, 2021, 51(5): 524−534
143. Burdea G C, Coiffet P. Virtual Reality Technology. 2nd ed. Hoboken: Wiley, 2003

144. Jang J, Frier W, Park J. Multimodal volume data exploration through mid-air haptics. In: Proceedings of 2022 IEEE International Symposium on Mixed and Augmented Reality. 2022, 243−251
145. Shiota T, Takagi M, Kumagai K, Seshimo H, Aono Y. Egocentric action recognition by capturing hand-object contact and object state. In: Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. 2024, 6527−6537
146. Khan A A, Newn J, Bailey J, Velloso E. Integrating gaze and speech for enabling implicit interactions. In: Proceedings of 2022 CHI Conference on Human Factors in Computing Systems. 2022, 349

Zhi-Min WANG is a postdoctoral researcher with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, China. He received his PhD from Beihang University in 2024 and his BS from Chang'an University, China in 2019. His research focuses on VR/AR, human-computer interaction, and eye-tracking technologies. He serves as a program committee member for AAAI 2025 and as a regular reviewer for leading international venues, including IEEE VR, ISMAR, TVCG, CVPR, and IJHCI.

Mao-Hang RAO is currently a PhD student with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, China. He obtained his bachelor degree in computer science and technology from Beihang University, China in 2023. His research interests include human-computer interaction, computer vision, and virtual reality.

Shang-Hua YE is currently an intern student with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, China. He obtained his bachelor degree in computer science and technology from Beijing Institute of Technology, China in 2023. His research interests include human-computer interaction, virtual reality, and computer graphics.

Wei-Tao SONG received his PhD degree in optical engineering, from Beijing Institute of Technology, China in 2016. He worked as a research fellow at Nanyang Technological University, Singapore from 2016 to 2020. He is currently a professor of Optical Engineering in Beijing Institute of Technology, China. His research interests include color science, human-computer interaction, and virtual and augmented reality.

Feng LU received his BS and MS degrees in Automation from Tsinghua University, China in 2007 and 2010, respectively, and his PhD degree in Information Science and Technology from The University of Tokyo, Japan in 2013. He is currently a full professor with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, China. His research interests include computer vision, natural interaction, and VR/AR. He is a distinguished member of CCF and CSIG. He is serving/has served as an Area Chair for prestigious international conferences such as CVPR, ICCV, ECCV, NeurIPS, and ACM MM.