At a Glance to Your Fingertips: Enabling Direct Manipulation of Distant Objects Through SightWarp

Yang Liu
Aarhus University
Aarhus, Denmark
yangliu.hci@gmail.com

Thorbjørn Mikkelsen
Aarhus University
Aarhus, Denmark
thormik@cs.au.dk

Zehai Liu
Aarhus University
Aarhus, Denmark
202303481@post.au.dk

Gengchen Tian
Aarhus University
Aarhus, Denmark
202303478@post.au.dk

Diako Mardanbegi
American University of Beirut
Beirut, Lebanon
diako.mardanbegi@aub.edu.lb

Qiushi Zhou
Aarhus University
Aarhus, Denmark
qiushi.zhou@cs.au.dk

Hans Gellersen
Lancaster University
Lancaster, United Kingdom
Aarhus University
Aarhus, Denmark
hwg@cs.au.dk

Ken Pfeuffer
Aarhus University
Aarhus, Denmark
ken@cs.au.dk

Abstract

In 3D user interfaces, reaching out to grab and manipulate something works great until it is out of reach. Indirect techniques like gaze and pinch offer an alternative for distant interaction, but do not provide the same immediacy or proprioceptive feedback as direct gestures. To support direct gestures for faraway objects, we introduce SightWarp: an interaction technique that exploits eye-hand coordination to seamlessly summon object proxies to the user's fingertips. The idea is that after looking at a distant object, users either shift their gaze to the hand or move their hand into view—triggering the creation of a scaled near-space proxy of the object and its surrounding context. The proxy remains active until the eye–hand pattern is released. The key benefit is that users always have an option to immediately operate on the distant object through a natural, direct hand gesture. Through a user study of a 3D object docking task, we show that users can easily employ SightWarp, and that subsequent direct manipulation improves performance over gaze and pinch. Application examples illustrate its utility for 6DOF manipulation, overview-and-detail navigation, and world-in-miniature interaction. Our work contributes to expressive and flexible object interactions across near and far spaces.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **Pointing**; **Gestural input**; **User studies**.

Keywords

Input techniques, extended reality, eye-tracking, gaze interaction

ACM Reference Format:
Yang Liu, Thorbjørn Mikkelsen, Zehai Liu, Gengchen Tian, Diako Mardanbegi, Qiushi Zhou, Hans Gellersen, and Ken Pfeuffer. 2025. At a Glance to Your Fingertips: Enabling Direct Manipulation of Distant Objects Through SightWarp. In *The 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25), September 28–October 01, 2025, Busan, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3746059.3747653

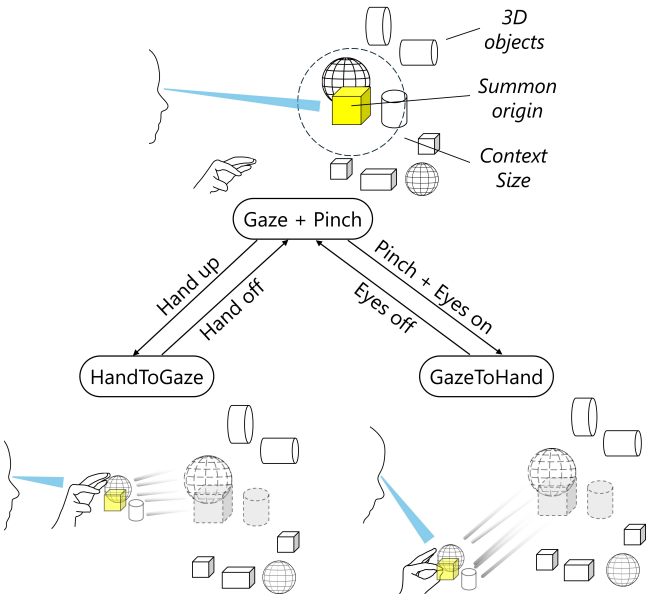

Figure 1: Direct physical manipulation of virtual objects is intuitive. To extend its use across spaces, SightWarp warps distant objects into the user's hand. By default (top), users manipulate objects at a distance using Gaze+Pinch. Alternatively, they can either move their hand up into their line of sight (left) or shift their gaze down to their hand (right), triggering a proxy that enables direct manipulation of the distant object. This tri-state model allows users to choose the most preferred mode for each gesture.

UIST '25, Busan, Republic of Korea

ACM ISBN 979-8-4007-2037-6/25/09
https://doi.org/10.1145/3746059.3747653

1 Introduction

The interaction capabilities of extended reality (XR) head-worn computers–experienced by people through headsets and smart glasses–are rapidly evolving. Modern devices, for instance, support hybrid techniques that combine direct and indirect 3D hand-tracking gestures within the same user interface to benefit from the complementary strengths of both input types [16]. *Direct gestures* enables physics-oriented interaction with high angular precision, making them compelling for spatial manipulation. When a user grabs a virtual object with their hand, the object moves and rotates in immediate response—creating a tightly coupled feedback loop between motor actions and object behavior. This interaction supports proprioceptive awareness and provides rich spatial cues [32]. With the integration of eye-tracking, *indirect gestures* allow users to point using their gaze, combined with a pinch gesture for object manipulation ("Gaze+Pinch " [33, 36, 57]). The gesturing hand can remain in a comfortable position, reducing physical effort and avoiding hand occlusion of the field of view [11, 25].

A key quality of XR interfaces is the support for interaction across depth, from near to far space [31]. In near space, users can fluidly switch between direct and indirect gestures. For far space, techniques such as World-In-Miniatures [52], VoodooDolls [38] and Scaled-World-Grab [32] as well as more recent approaches [41, 59] allow users to summon proxies of faraway objects to near space. These methods enable the benefits of direct manipulation through a switching mechanism, but often require a separate invocation gesture [41] or override the existing indirect interaction mode [32, 38]. In this research, we focus on extending the default gaze-and-pinch UI semantics with a proxy summoning– but without asking users to learn extra gestures or give up indirect control. The goal is to make switching between interaction styles feel as seamlessly as in near space.

We propose **SightWarp**, a technique that exploits eye-hand coordination for summoning near-field proxies of distant objects in XR UIs. SightWarp is always available, complements Gaze+Pinch, and can be accessed on demand for each gesture. Users begin by identifying a distant object or region of interest using gaze. Then, they can transition into direct gestural manipulation by summoning a proxy of the object and its surrounding context into near space, at the location of their fingertips. This transition is triggered by coordinating gaze and hand in two ways (Figure 1):

- **GazeToHand**: After initiating a Gaze+Pinch command and holding the pinch gesture, the user directs gaze to their hand. This triggers a context warp, bringing a proxy of the selected object and its local context to the hand, where it appears from a different perspective and is ready for direct gestural manipulation.
- **HandToGaze**: Alternatively, the user can raise their hand into view while maintaining gaze on the distant object. Since users typically do not look at their hand while manipulating distant objects with indirect gestures, this distinct state—gaze focuses on a distant object, with the hand intruding into view—summons the proxy to the hand's location.

SightWarp can be useful for various applications. GazeToHand provides a new perspective by anchoring the summoned object to the user's hand (Figure 2a–b). For instance, when interacting with a city model, GazeToHand enables users to select a specific area and view it from above—facilitating occlusion management and fine-grained object placement in the summoned sub-scene. In contrast, HandToGaze warps content in the user's forward view closer, functioning like a zoom lens (Figure 2c–d). This is particularly useful for distant menus or UI subregions, allowing users to summon a proxy for detailed inspection and interaction, even before deciding what to select or manipulate.

To investigate how effectively users can perform summoning transition and the tangible benefits of proxy-based manipulation, we conducted a user study. Gaze+Pinch was the baseline indirect manipulation technique for distant objects, and we compared it to our two proposed methods, which employ distinct summoning mechanisms–GazeToHand and HandToGaze–to enable direct gestural manipulation. We selected a 3D translate+rotate docking task as a standardized, controlled method for gaining insights in user performance and experience. The study results indicate the following findings:

- Both SightWarp techniques significantly reduced task completion time compared to the baseline, with performance advantages becoming more pronounced as task complexity increased.
- Gaze+Pinch led to more clutches and erroneous gestures, reflecting late-trigger issues [22] and limited preshaping [25].
- GazeToHand resulted in significantly reduced hand movement than Gaze+Pinch, suggesting that the relaxed hand posture facilitates more efficient manipulation.

In sum, our results indicate that SightWarp offers measurable benefits over the state-of-the-art eye-hand indirect manipulation technique (Gaze+Pinch) in 3D docking tasks. Notably, the GazeToHand summoning mechanism did not incur performance penalties, suggesting a fluid, low-cost transition into direct manipulation. This makes SightWarp particularly suited for complex spatial tasks such as object docking, alignment, and manipulation–common in 3D design, modeling, and similar domains. Moreover, since SightWarp triggers summoning only through specific eye-hand coordination patterns, it is compatible with existing indirect gestures, allowing XR UIs to support the entire spectrum of direct and indirect interaction across near and far spaces.

The main contributions of this paper to HCI are:

- SightWarp, an XR interaction technique that (1) integrates with the existing Gaze+Pinch paradigm, though (2) enabling summoning of remote object proxies with two eye-hand coordination patterns:
  - GazeToHand: establishes a spatially-distinct perspective as a new context perspective is presented near the hand's location;
  - HandToGaze: establishes a spatially-consistent perspective as the currently-viewed context is warped to near space;
- A user study comparing GazeToHand and HandToGaze with the baseline Gaze+Pinch for object docking, showing SightWarp users are more efficient w.r.t. task completion time, clutches, and errors, revealing traits of GazeToHand and HandToGaze.
- A set of application examples, demonstrating the utility of SightWarp for cross-space object transfer, occluded and small object selection, details on-demand, and focus-and-context scenarios.

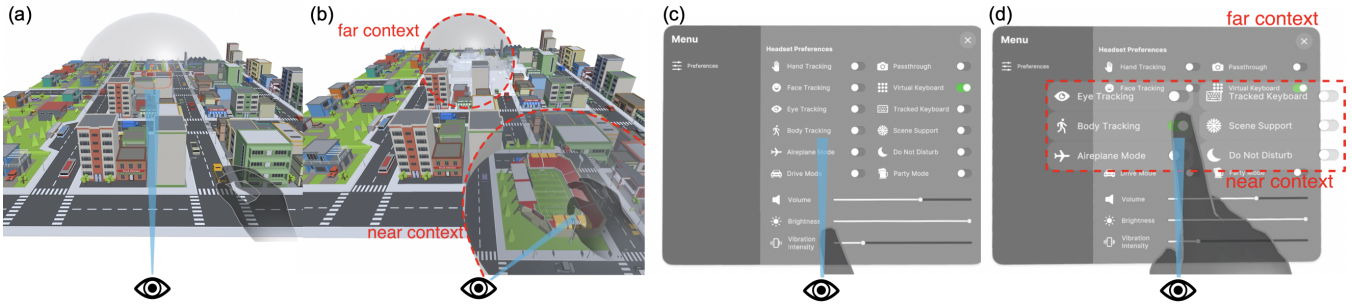

Figure 2: SightWarp applications: (a-b) overview and detail in a city planning tool where users leverage GazeToHand to obtain new perspectives to manipulate detailed or occluded objects; (c-d) HandToGaze for a magnified view of any distant UI by summoning the gaze-focused region to the same fixation distance as the hand and interacting using an air-tap. The eye icon and blue triangle indicate the user's gaze.

2 Related Work

We structured the discussion of related work into four parts, described in the following subsections.

2.1 Distant Interaction in XR

XR devices incorporate advances in hand-tracking technology that enable users to perform direct 3D manipulation of virtual objects without relying on controllers [30], fostering a deeper sense of presence [6] while improving comfort and immersion [14]. However, direct hand interaction faces two primary challenges: (1) difficulty interacting with distant or unreachable targets, and (2) fatigue induced by extended hand or arm elevation [1, 4, 15, 48]. While hand ray-casting [12, 23, 43, 50, 54] provides an effective workaround for the distance issue, studies indicate that hand-based pointing and selection can increase fatigue [5, 27, 55, 57, 62]. Because hands must handle a variety of tasks including pointing, selection via gestures, and subsequent manipulation, their risk of being overburdened further complicates the user experience.

Gaze interaction offers natural, intuitive, and efficient means of conveying user intent when selecting or interacting with distant virtual objects [19, 53], which has led to extensive research in XR contexts [3, 20, 40]. Most XR systems adopt multimodal approaches that combine gaze with other inputs to avoid visual overload from using gaze alone. Common combinations include gaze with head movements [28, 39, 49] or hand gestures [27, 33, 36, 55]. A widely explored division of labour is the "eyes select, hands manipulate" paradigm of Gaze+Pinch [36]. Studies have shown this approach improves performance for selection compared to hand-raycasting, image plane techniques [44, 57], and gaze-only methods [33, 51].

However, the benefits of gaze-hand techniques seem diminished when applied to complex manipulation tasks [25, 56, 60]. For example, recent studies on object movement [55] and asymmetric bimanual manipulation [25] found that while indirect eye-hand gestures reduced physical effort compared to direct gestures, they did not improve overall performance. Gaze was mainly beneficial for initial selection, while manipulation phases were slower—likely due to limited proprioceptive feedback and the lack of hand preshaping cues [25]. These findings suggest a potential ceiling of gaze-hand techniques in complex tasks, highlighting the need for more effective methods for manipulating distant objects.

2.2 Image Plane Interaction Techniques

Our goal is to explore how distant object manipulation can be more effectively based on the notion of direct gestures. This is closely related to Pierce et al.'s image plane interaction techniques, which treat the 3D scene from the user's perspective as a 2D image plane, enabling users to interact with distant content through direct hand gestures [37]. For instance, the HeadCrusher technique allows users to select a distant object by occluding it with a pinching hand. This makes it plausible to follow up with pinch-based rotation, scaling, and translation (RST) gestures. Such methods could complement Gaze+Pinch, especially because indirect gestures typically keep the hand out of the line of sight. However, depth differences between the hand and distant targets introduce parallax effects, which can cause visual misalignment (e.g., doubled or offset fingers) and create ambiguity in the perceived depth and precise selection point. Later work showed Gaze&Finger to be more efficient for close targets than distant ones due to parallax [57]. This limitation was also noted in Pierce et al.'s discussion of image plane techniques [37].

To improve occlusion-based selection, researchers have explored multimodal combinations with gaze. EyeSeeThrough, introduced by Mardanbegi et al., leverages spatially-coupled eye-hand coordination to eliminate explicit mode switching [28]. Gaze&Finger extended this idea to selection tasks by aligning the index finger with the user's gaze in view space [27], achieving performance comparable to Gaze+Pinch [57]. This approach has also been applied to region selection in AR [46] and finger typing in VR [26], where it reduced finger movement compared to standard mid-air typing.

We initially considered such gaze-based image plane techniques as a pathway to direct object manipulation. However, relying on gaze selection for every gesture departs from the directness of hand gestures. It requires users to constantly shift focus to the target, and the need to synchronize gaze and hand gesture timing can lead to the late-trigger problem [22].

2.3 Hand Teleporting and Object Summoning

To address the limitations of distant manipulation, one class of techniques focuses on transporting the user's virtual hand to the distant target. An early example was the Scaled-World Grab locomotion variant proposed by Mine et al. [32], where users are virtually

transported toward an object through a single grabbing action. Similarly, the Go-Go technique employs a non-linear mapping between the physical and virtual hand, effectively extending reach beyond physical constraints [42]. This inspired a range of "virtual-hand teleportation" techniques that enable distant interaction without full-body locomotion [5, 9, 21, 42, 64].

An alternative approach tackles the challenge from the opposite direction: by bringing a representation of the distant object (also known as near-field metaphor, proxy, or replica) into the user's reachable space. A seminal work is Stoakley et al.'s World in Miniature [52], where a handheld mini-world represents the entire virtual scene, allowing users to indirectly manipulate the objects in the scene. Bringing objects closer enables users to manipulate them directly while benefiting from proprioception, stereopsis, head-motion parallax, and improving manipulation accuracy [32].

To better make this concept scale, subsequent research focused on interaction techniques to trigger the proxy creation on the fly for any given virtual context. From an input-theoretical point of view, These can be classified into two categories. First, using a dedicated, additional input command to trigger the proxy creation. For instance, specific gestures (e.g., Poros [41]) or occlusion selection and bimanual input (e.g., VoodooDolls [38]). These allow a clear separation of intent, at the expense of adding an additional step before one can use the proxy, and an additional command to learn for the user. Second, completely replacing the remote control method. For instance, Scaled-World Grab [32] warps the selected object and its context to the user's hand at every gesture. However, this takes away the option to fall back to the default remote control paradigm. Our work extends the prior art through exploring a warping technique without new commands due to exploiting eye-hand coordination patterns.

2.4 Direct/Indirect Mode Switching

Historically, computer systems have relied on distinct direct and indirect input devices, each offering complementary interaction properties [16]. To harness the strengths of both, hybrid techniques have been developed that allow users to switch between input modes. For example, HybridPointing supports direct pen input on large displays, but transitions to an indirect cursor mode when interacting with a trailing widget [10]. Similarly, ARCPad extends a touchpad's relative pointing with an absolute mode, differentiating between tap and drag gestures [29].

The multimodal combination of eye-tracking and direct input devices supports both direct and indirect gesture modes modulated by eye-hand coordination. Some approaches explored explicit mode switching, e.g., FingerSwitches [34] uses Gaze+Pinch micro-gestures to switch UI windows across static, dynamic, and self entities. A distinct category is implicit mode switches without explicit manual input. For instance, Gaze-Shifting [35] uses implicit modulation based on eye-hand coordination patterns. This approach defines direct manipulation when manual input falls within a predefined gaze-centric range in a 2D interface, determined by the distance between the gaze point and input position. For instance, a pen's input can seamlessly transition from the default direct drawing mode to indirect menu operation when the user's focus shifts to a distant menu. Such a co-existence of direct and indirect input modes at the granularity of each input command minimizes mode-switching costs [18]. This principle has been extended to 3D user interfaces, using hand-tracking input device and visual-angle-based range definitions, enabling users to shift between direct and indirect modes for each pinch gesture [25, 36], and is a core feature of the Apple Vision Pro's UI. We extend the prior art by considering how the direct-indirect flexibility can be brought to the manipulation of objects at a distance.

3 SightWarp Interaction Design

SightWarp is a novel technique to summon proxies of distant objects into reach through exploiting eye-hand coordination patterns. In the following, we detail the design and parameters of the method.

3.1 Phases of SightWarp Interaction

Our method integrates the act of proxy summoning with subsequent direct gestural operations into a cognitively unified interaction flow. The interaction procedure includes the following states (Figure 3):

1. **Start**: A target is identified based on gaze direction and fixation.
2. **Trigger**: Summoning proxies of the target and its context to the user's hand.
3. **Manipulation**: Then, users directly operate on the proxy of target or of other objects within the context.
4. **Release**: The target and its context returns to the far space with the results of the direct interaction once the trigger condition is no long maintained.

For **Trigger**, we adopt a simple eye-hand coordination pattern that avoids interfering with indirect gestures in far space. Gaze+Pinch is typically used with the hand held ergonomically away from the gaze direction, e.g., near waist level, aligning with prior observations [25]. This makes gaze-hand alignment–either looking at the hand or bringing it into the line of sight–an expressive and yet unused eye-hand coordination pattern during Gaze+Pinch. Albeit prior work has exploited this pattern for various object selection use cases [26–28, 45], we re-imagine this pattern for a new purpose: to trigger proxy summoning for direct gestural control. The mechanism includes two simple ways of eye-hand coordination, which we define as two modes in SightWarp's input model:

- GazeToHand: Summoning is triggered by users explicitly directing their gaze to the hand after selecting an object with Gaze+Pinch. GazeToHand offers a distinct perspective of the far context by summoning it to the hand (e.g., a top view).
- HandToGaze: HandToGaze is is triggered by bringing the hand to pinch on the line of sight that focuses on a target. HandToGaze creates the visual perception of directly manipulating distant objects while preserving its viewing angle. The near-space proxy can be summoned either at the hand's position or along the gaze line. Summoning to the hand ensures that every pinch can exactly select the gazed target. Summoning along the gaze, while not guaranteeing a successful direct selection upon initial pinch, allows for more flexible hand repositioning to interact with objects other than the initially gazed target within the proxy.

For summoning the proxy of a distant object for direct gestural manipulation, its immediate surroundings (**context**) must also be summoned for contextual reference, following established practice

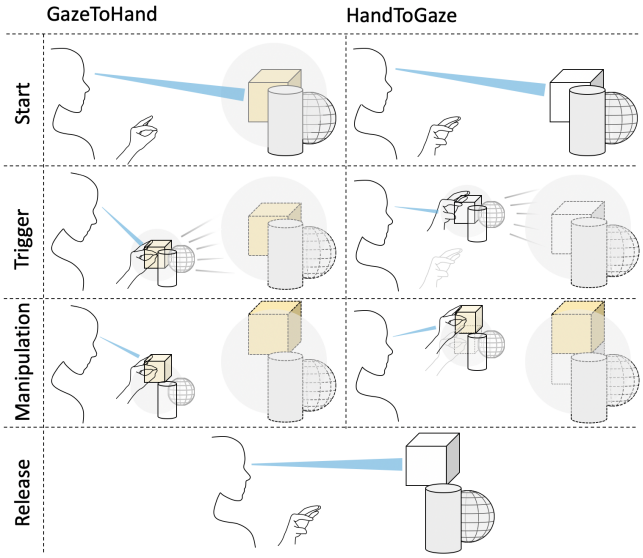

Figure 3: Four-phase illustration of GazeToHand and HandToGaze. The blue triangle indicates the user's gaze. An object is colored yellow when it is being manipulated with indirect or direct gestures. GazeToHand requires users to first pre-select a target via Gaze+Pinch, then triggers summoning by directing their gaze toward their pinch hand. Users then apply direct gestures to move the target in near space, and finally release the pinch while moving their gaze or hand away to deactivate summoning and disengage the proxy. In contrast, HandToGaze does not necessitate a Gaze+Pinch pre-selection—users can gaze at a target and move their hand into alignment with gaze to trigger summoning. Users then pinch on the near space object to initiate manipulation.

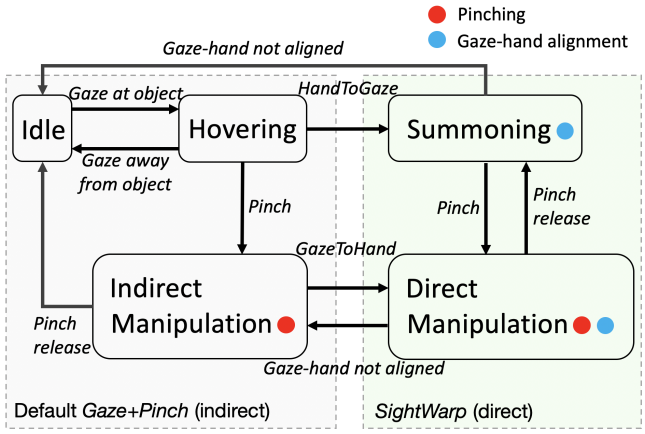

Figure 4: Five-state model illustrating transitions among the three modes: default Gaze+Pinch (left dashed box) and the two SightWarp variants (right dashed box). Colored dots (red and blue) indicate conditions needed to be maintained to remain in a given state. The *Summoning* state denotes that a proxy is summoned but not yet manipulated, while a pinch transitions to the *Direct Manipulation* state for the proxy.

[32, 41, 52]. The contextual summoning preserves relative positioning among objects. It is particularly valuable in the GazeToHand mode, where users can focus entirely on near-space manipulation without needing to visually cross-reference the far-space context.

The **Manipulation** phase is compatible with but not limited to Gaze+Pinch. Once gaze fixates on a target, the user may either pinch to interact indirectly, or transition to direct interaction via summoning: by pinching and then looking at the hand (GazeToHand), or by performing a spatial alignment of their gaze and hand to trigger HandToGaze. Upon gaze-hand alignment, a proxy of the target and its context is summoned into the near space. This summoned proxy persists as long as the gaze and hand are in proximity within the view plane, regardless of whether the pinch is held. This design allows users to move their hand within the proxy to acquire and manipulate different objects in the context.

When users intend to **Release** the summoned proxy, they simply move away their gaze or hand to break their spatial alignment.

For **Trigger** and **Release**, the angular threshold of gaze-hand alignment for summoning is important. A smaller angle (e.g., 5°) enables more precise summoning but requires greater effort to align, which has been successfully used for precise selection techniques [26, 27]. In contrast, a larger angle (e.g., 30 °) offers easier activation but might lead to accidental triggering. In our context, we adopt a generous angular threshold as we consider a different task of summoning a proportion of the far space for users to engage in subsequent interaction, where ease of use outweighs precision. Moreover, we apply an even more relaxed threshold for deactivation, reducing the chance of accidentally breaking the spatial alignment when users move their hand around the summoned context.

Figure 3 illustrates the four phases for both summoning modes of SightWarp. GazeToHand requires users to first select a target via Gaze+Pinch, then triggers summoning by directing their gaze toward their pinch hand. Users then apply direct gestures to move the target in near space, and finally release the pinch while moving their gaze or hand away to deactivate summoning and disengage the proxy. In contrast, HandToGaze does not necessitate a Gaze+Pinch pre-selection, as users can gaze at a target and move their hand into alignment with the gaze to trigger summoning. The user then pinches on a near-space object to initiate manipulation. The release phase is identical between both modes.

3.2 Input State Model

Figure 3 illustrates the steps of using each summoning mode and Figure 4 presents the transitions among five states in a system where Gaze+Pinch, GazeToHand, and HandToGaze modes coexist. Beyond the existing three states of Gaze+Pinch (*Idle*, *Hovering*, *Indirect Manipulation*), the HandToGaze pathway introduces a transition from *Hovering* to *Summoning*, enabling subsequent *Direct Manipulation* with a pinch. In contrast, GazeToHand triggers direct proxy manipulation from the *Indirect Manipulation* state. Overall, the system provides two bidirectional pathways for switching between indirect and direct modes, featuring our design goal of a fluid, always-available transitioning mechanism.

3.3 Design Considerations

Both the far-space original and near-space proxy contexts are visualized as semi-transparent spheres, each centered on the corresponding target object. We choose a spherical shape because it provides uniform coverage of surrounding space in all directions, helping perceive and adjust the scope of both contexts, akin to the design of Poros [41].

Upon selection via Gaze+Pinch, a semi-transparent context sphere appears around a selected object. Any object intersecting this sphere is included in the context to be summoned. A proxy sphere is summoned into the near space upon triggering gaze-hand alignment. To reduce visual clutter, portions of contextual objects extending beyond the near-space sphere bounds are cropped. A larger far-space context encompasses a wider range of spatial references, while a smaller one captures only the immediate surroundings. By default, the initial diameter of the far-space context is set to twice the size of the target's bounding box.

For the near-space proxy, its size influences the Control-Display (CD) ratio. Indirect input techniques such as mouse, hand ray, and Gaze+Pinch commonly employ a CD ratio to map hand movement to object translation to alleviate physical effort and arm fatigue when moving objects over large distances. A common practice for Gaze+Pinch distant manipulation is to use a visual angle-based CD ratio, where the object moves across the same angular distance in the user's view as the hand [25, 56]. This is achieved by matching the translation distance in visual angle between the target in far space and its proxy in near space.

We extend this principle to the HandToGaze mode, which preserves the original viewing angle. Thus, it naturally benefits from the same visual-angle-based CD ratio strategy by scaling the proxy to match the visual size of the original context. In contrast, the GazeToHand mode is not constrained by visual angle consistency, offering greater flexibility to adjust CD gain for optimising manipulation precision or efficiency, depending on the use case. The CD gain can be adjusted by resizing the near-space context. Users can modify the size of both the far and near context spheres via an arc-shaped handle at the upper-left location of each sphere.

4 User Study

This user study investigates the trade-off between the cost of transitioning from indirect to direct manipulation and the potential performance gains enabled by SightWarp. Existing paradigms for distant object interaction, such as Gaze+Pinch, offer efficient selection but lack the benefits of direct, hands-on manipulation. Our technique, SightWarp, is designed to complement Gaze+Pinch by providing distinct modes for direct gestural interaction. However, this design introduces a necessary transition cost. To evaluate this trade-off, we compare simplified variants of SightWarp's summoning mechanisms against a Gaze+Pinch baseline on a 6DOF docking task. Our research questions (RQs) are as follows:

RQ1: How does user performance with HandToGaze and GazeToHand compare to Gaze+Pinch? For both SightWarp techniques, they are different in the way of triggering (move your gaze, or move your hand), and in the perspective of summoned content (holds perspective, vs. provides a new perspective). What is their effect on the user's performance and experience, and how do they compare against the Gaze+Pinch baseline?

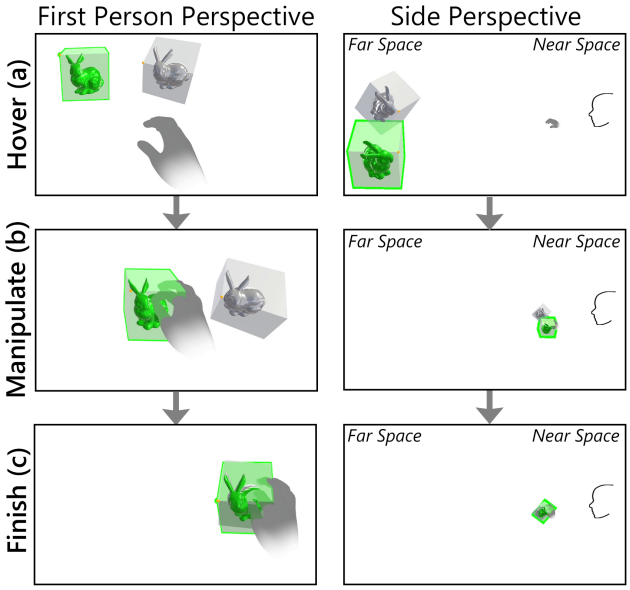

Figure 5: Trial sequence for the two summoning conditions (GazeToHand and HandToGaze). The left column shows the first-person perspective, and the right column shows the side view. (a) The user hovers over the object using gaze. (b) The object and the target are warped to the user's hand as a direct manipulation proxy. (c) The trial is completed once both translational and rotational thresholds are met.

RQ2: How does task complexity affect the user's performance with SightWarp? Given the cost of the context switch, it is unclear which point of task complexity will it become beneficial to use direct gestures in near space, over indirect gestures in far space. We investigate task completion time, manipulation time, clutches and errors across two object sizes and rotation difficulties.

RQ3: How well can users perform the initial summoning? Users need to change their focus distance and change the interaction paradigm from indirect to direct gesture. How do people manage the context shift, and what are potential costs to this context change?

We conducted a within-subject study. We employed a 3×2×2 factorial design of the following independent variables, with condition order counterbalanced across participants:

- Techniques: Gaze+Pinch (Baseline), GazeToHand, HandToGaze
- Rotation Magnitude (between initial/target rotation): 45°, 90°
- Object Size (in visual angle): 7.5°, 12.5°

4.1 Task Design

The task is a 6DOF docking task, where participants manipulate a 3D object to match the position and orientation of a target object [2]. This task is chosen because it is the state-of-the-art evaluation method for 3D manipulation, which SightWarp enables. While SightWarp is also applicable to other tasks, such as selection and multi-step workflows as explored in section 5, evaluation of these scenarios is out of scope.

The manipulable object is a semi-transparent green cube that encloses an opaque Stanford bunny [60], with a size of either 7.5°or 12.5° in visual angle (26.2 cm or 43.8 cm wide at 2 m viewing distance). The target object is identical in shape and size, but rendered in gray and not interactive.

Each trial begins when the object and its target appear, and ends when their positional offset falls below 20% of their visual angle size, and their orientational difference is within 15° in quaternion angle. A three-step trial sequence for the two summoning conditions is illustrated in Figure 5. To avoid accidental completions due to brief or unstable alignments, participants need to maintain these completion conditions for 300 ms. Clutching is allowed.

At the beginning of each trial, the manipulatable cube appears 2 meters in front of the participant, front facing and chest-aligned, as approximated from the HMD position. The target object is positioned at an offset of twice the object's visual size along one of four displacement directions (+X, -X, +Z, -Z), and rotated by either 45° or 90° around one of three randomly selected axis pairs (±X±Y, ±Y±Z, ±X±Z). The Y-axis was omitted from positional displacement to reduce study complexity.

Each block (a unique combination of Technique, Rotation Magnitude, and Object Size) included 12 trials, covering all 4 displacement directions crossed with 3 randomly assigned rotation axes. The trial order within each block is randomized. In total, each participant completed 3 Techniques × 2 Rotation Magnitudes × 2 Object Sizes × 12 combinations = 144 trials.

4.2 Procedure

Participants were first briefed on the study and completed consent and demographics forms. Before each condition, they watch an instructional video demonstrating the respective technique. They then wore the headset and performed fit adjustment and eye-tracking calibration. Participants remained seated in a static, non-swiveling chair throughout the study, allowing only upper-body movement. For each condition, participants began with a hands-on training session in a task-free environment, practicing the technique until they felt comfortable. They then completed four blocks of 12 docking trials, with varying object sizes or rotation magnitudes. Breaks were allowed between blocks, and the next block was initiated upon confirmation with the researcher. Participants were instructed to perform as fast as possible. After each condition, participants completed a post-condition questionnaire and repeated headset fitting and gaze calibration. At the end of the session, they completed a post-study questionnaire.

4.3 Evaluation Metrics

We collected the following metrics to assess task performance, perceived workload, and user experience.

- Trial Completion Time: time from object appearance to trial completion, i.e., meeting the accuracy thresholds.
- Acquisition Time: time from object appearance to the first pinch.
- First Manipulation Duration: duration of the first pinch.
- Clutch Count: number of clutch gestures.
- Failed Gesture Count: number of pinch gestures that had not no effect on the object (i.e., failed grabs).
- Hand Translation: total hand travel distance while pinching.
- Hand Rotation: total hand rotation while pinching.
- After each condition, participants completed the NASA TLX questionnaire [13] to report perceived workload.
- After completing all trials, participants rated their overall experience for each technique and provided written feedback.

4.4 Apparatus and Implementation

The study was implemented in Unity (2022.3.19f1) for the Meta Quest Pro (90 Hz display, 30 Hz eye tracker), using the Meta XR All-in-One SDK (v74.0.1). Hand-tracking data was smoothed using a 1€ Filter [7], and pinch gestures were detected with a relaxed confidence threshold for easier acquisition. During manipulation, a green outline highlights the selected object.

To ensure consistency across all techniques, both summoning and manipulation were initiated with a pinch gesture. In the HandToGaze and GazeToHand conditions, summoning occurs at the moment of pinching, enabling immediate manipulation in near space within the same gesture session. To encourage near-space interaction and avoid visual clutter, far-space objects were removed immediately upon summoning.

For the HandToGaze condition, we set the angular threshold for gaze-hand alignment to 25°, determined through pilot testing to balance ease of triggering, as discussed in subsection 3.1. Piloting also revealed that users sometimes brought their hand too close to their face, causing the summoned object to appear uncomfortably near. To mitigate this, we added a depth constraint: the hand had to be within 0.3–0.5 m from the user to trigger summoning. For exiting, the threshold was slightly relaxed to 30°, and the valid depth range was extended to 0.25–0.65 m. These buffers help prevent unintended exits, such as slipping out of range during mid-pinch.

To ensure comparability across techniques, object movement was computed based on visual angle rather than direct 1:1 hand displacement. Specifically, movement was scaled by the ratio between the distance from the far-space object to the user's eyes and the distance from the hand to the eyes, ensuring that positional manipulation remains consistent in terms of angular displacement across all techniques. For GazeToHand and HandToGaze, summoned objects were scaled down based on this same control-display ratio, preserving their original visual size.

4.5 Participants

12 participants (4 female, 8 male) took part from the local area, primarily university students. Participants ranged in age from 22 to 35 (M = 26.91, SD = 4.09). All were right-handed or ambidextrous; 4 wore glasses and 2 wore contact lenses. On a 5-point scale, participants reported little to moderate experience with VR/AR (M = 2.67 SD = 1.17), 3D hand gestures (M = 2.41 SD = 1.32), and gaze input (M = 2.41 SD = 1.38).

4.6 Results

For task performance data, we applied the Aligned Rank Transform (ART) to address deviations from normality [58]. Next, we performed a repeated measures ANOVA with performance data, and post hoc pairwise comparisons (Holm-Bonferroni corrected). For NASA-TLX and preference ratings, we performed a Friedman's Test and found no significant results. We plot results for measures

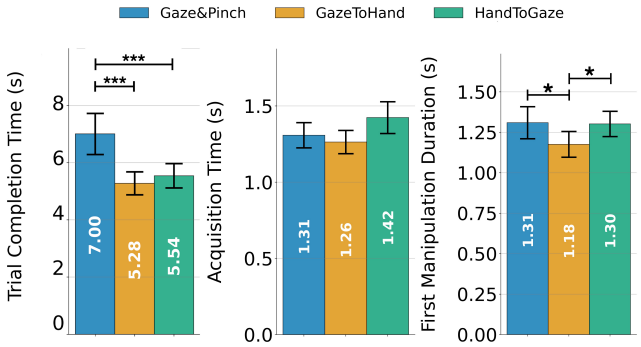

Figure 6: Mean Trial Completion Time, Acquisition Time, and First Manipulation Duration.

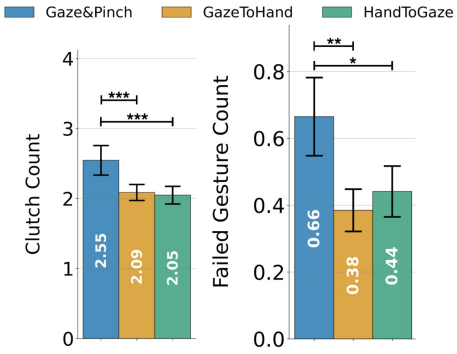

Figure 7: Mean Clutch Count and Failed Gesture Count.

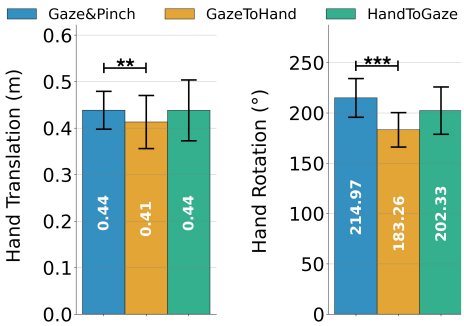

Figure 8: Mean Hand Translation and Hand Rotation.

that yielded significant results regarding Technique in Figure 6-8 while only reporting main effects of Rotation Magnitude and Object Size in text. Statistical significance is shown as * (p < .05), ** (p < .01), and *** (p < .001). Error bars indicate standard deviation.

4.6.1 Trial Completion Time (Figure 6). We found significant effects in Technique ($F_{2,121} = 20.09, p < .001$), Rotation Magnitude ($F_{1,121} = 193.78, p < .001$), Object Size ($F_{1,121} = 8.87, p < .01$), and Rotation×Size ($F_{1,121} = 4.25, p < .05$). Post hoc comparisons revealed that Gaze+Pinch was slower than both GazeToHand ($p < .001$) and HandToGaze ($p < .001$). Performance was also significantly slower for the larger Rotation Magnitude ($p < .001$) and for the smaller Object Size ($p < .001$).

4.6.2 Acquisition Time (Figure 6). While we did not find significant effect, we plot the data in Figure 6.

4.6.3 First Manipulation Duration (Figure 6). We found significant effects in Technique ($F_{2,121} = 4.73, p < .05$) and in Rotation Magnitude ($F_{1,121} = 22.61, p < .001$). Post hoc comparisons revealed that the First Manipulation Duration of GazeToHand was shorter than for Gaze+Pinch ($p < .05$) and HandToGaze ($p < .05$). Additionally, the duration was shorter for the smaller Rotation Magnitude ($p < .001$).

4.6.4 Clutch Count (Figure 7). We found significant effects in Technique ($F_{2,121} = 14.10, p < .001$), Rotation Magnitude ($F_{1,121} = 335.05, p < .001$), Object Size ($F_{1,121} = 4.57, p < .05$), and Rotation Magnitude ×Object Size ($F_{1,121} = 6.12, p < .05$). Gaze+Pinch induced more clutches than both GazeToHand ($p < .001$) and HandToGaze ($p < .001$). Larger Rotation Magnitude ($p < .001$) and smaller Object Size ($p < .05$) induced more clutches.

4.6.5 Failed Gesture Count (Figure 7). We found significant effect in Technique ($F_{2,121} = 10.77, p < .001$). Post hoc comparisons revealed that Gaze+Pinch induced more errors than both GazeToHand ($p < .01$) and HandToGaze ($p < .05$).

4.6.6 Hand Translation (Figure 8). We found significant effects in Technique ($F_{2,121} = 4.49, p < .05$), Rotation Magnitude ($F_{1,121} = 213.39, p < .001$), and Object Size ($F_{1,121} = 21.92, p < .001$). We find that GazeToHand induced less hand translation than Gaze+Pinch ($p < .01$). Respectively, the larger Rotation Magnitude ($p < .001$) and Object Size ($p < .001$) induced more hand translation.

4.6.7 Hand Rotation (Figure 8). We found significant effects in Technique ($F_{2,121} = 9.45, p < .001$) and Rotation Magnitude ($F_{1,121} = 466.10, p < .001$). Post hoc comparisons show that GazeToHand induced less hand rotation compared to Gaze+Pinch ($p < .001$). Respectively, the larger Rotation Magnitude ($p < .001$) induced more hand rotation.

4.6.8 User Feedback. Participants generally found Gaze+Pinch natural (2 participants) and offering good control of the object (4), but also fatiguing (3). In contrast, GazeToHand was perceived as less tiring (3) and easier to use (7). Two users favored the hand alignment posture in HandToGaze. However, four participants reported eye strain when refocusing on near objects across both GazeToHand and HandToGaze.

Overall, user feedback focused on perceived control and arm fatigue, with GazeToHand and HandToGaze generally favored over Gaze+Pinch in these aspects, despite an increase in eye fatigue.

4.7 Discussion

Regarding **RQ1**, our results show that both GazeToHand and HandToGaze significantly outperformed the baseline Gaze+Pinch across all task performance measures, including Trial Completion Time, Clutch Count, and Failed Gesture Count. Participants completed the docking task faster, with fewer clutch gestures to re-orient the hand and fewer failed attempts to grab and manipulate the object using GazeToHand and HandToGaze. These findings

suggest that both techniques afford better spatial manipulation by supporting direct manipulation.

Besides the main effects of Technique, we observed main effects of Rotation Magnitude in Trial Completion Time and Clutch Count, where 90° rotations consistently yielded worse performance than 45°. These findings address **RQ2**, indicating that the performance benefits of direct manipulation in near space using HandToGaze and GazeToHand is consistent over Gaze+Pinch.

RQ3 investigated whether adapting to the depth view change between the original object and the summoned proxy would cause temporal overhead for acquisition and overall manipulation. Because we found no significant difference in Acquisition Time among the three technique conditions, this suggests that users were able to handle the initial summoning and context switch with minimal effort. Although HandToGaze involves an additional phase of explicit gaze-hand alignment compared to GazeToHand, we cannot conclude that this overhead led to inferior performance, as there was no significant difference in Trial Completion Time between them, likely due to their distinct eye-hand coordination patterns. Furthermore, our First Manipulation Duration measure showed that the initial manipulation was significantly shorter with GazeToHand than the other two techniques. This suggests that participants might have adopted a strategy with GazeToHand: they would rapidly direct their gaze to their hand after summoning the object, then release the pinch to plan subsequent manipulations. Many participants reported that GazeToHand was "easy and effortless to use", which may specifically refer to the summoning action. In contrast, HandToGaze might offer subtle benefits for planning both the acquisition and initial manipulation, thanks to preserving of the object's visual angle from the user's perspective.

5 Application Scenarios

This section demonstrates the applicability of SightWarp across various XR use cases, including multi-step workflows, by showcasing how its different parameters and design choices can be modulated to facilitate broader near-far interaction paradigms. In principle, SightWarp expands near-far interaction of XR UIs by incorporating the paradigm of direct-indirect interaction for distant objects. Figure 9 illustrates how all four parts can be integrated in the same UI, through mode-switching mechanisms:

Direct & Near By default, direct manipulation is active when a hand intersects with a virtual object.

Indirect & Far Gaze+Pinch is active when interacting with a distant object from a convenient hand position.

Indirect & Near Users can interact with a nearby object if the hand is offset from the gaze-selected object, for occlusion-free [16], flexible-CD-gain [8] and low-effort interaction [56].

Direct & Far When looking at a faraway object and performing GazeToHand or HandToGaze, the user summons object proxies in near space to manipulate them directly.

An example for a generic usage is to utilise HandToGaze as a "zooming" metaphor (Figure 2c–d). Direct gestural interaction such as tapping with 2D UI in XR benefits scenarios when gaze pointing suffers from crowded interfaces, and also affords natural interaction similar to physical touch interfaces. Our approach extends these benefits of direct manipulation to distant UIs. While the interaction

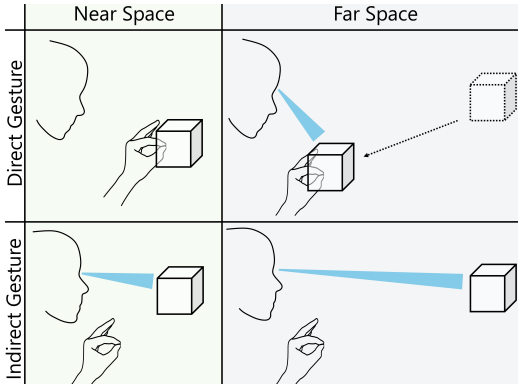

Figure 9: Co-existing modes: Eye-Hand XR UIs support flexible switching of gesture modes, to reap the benefits of direct and indirect inputs for objects across near and far spaces.

of HandToGaze is similar to previous work [26, 27, 57] based on image plane techniques [37], our summoning mechanism brings UI elements closer in front of the user's hand, which naturally resolves the parallax issue identified in the prior studies. Figure 2c–d demonstrates a summoned toggle via a direct tapping gesture.

A key issue in working in 3D is occlusion, and the need for seeing a 3D model from different perspectives, to which SightWarp provides an elegant solution. For instance, object movement in 3D can be challenging as depth change from the user's forward perspective is difficult to perceive. With GazeToHand, users can quickly switch perspectives to perform this task more efficiently.

We show several application examples for 3D design. We support the interaction with a visual feedback in form of a sphere to indicate what part of the world will be summoned and the boundary of a near-space proxy. Both far and near context spheres have an arc-shaped handle at their upper-left location used to adjust the context size. Since the dominant hand is occupied for object manipulation tasks, the non-dominant hand is used to interact with the scaling handle. To maintain interaction consistency within each space, the far context handle is selected using Gaze+Pinch (indirect gesture) while the near one is grabbed using direct pinch. After selecting a handle, the user moves their non-dominant hand horizontally to change the distance between their two hands, resulting in scaling up or down the corresponding context proportionally.

5.1 Cross-Space Drag-and-Drop (Figure 10)

Both GazeToHand and HandToGaze modes provide a close-up view for precise Drag-and-Drop during the task. A workflow enabled by HandToGaze, for example, is to employ direct manipulation for fine-grained operations and indirect manipulation for rapid long-distance movement–toggled at a glance. For example, after selecting an object, users can either move it in near space by looking at the near-space proxy while dragging, or switch to indirect manipulation by looking off toward the destination. Summoning can be reactivated by moving the pinching hand to the destination area and re-aligning it with gaze for precise placement.

Another use case is shown in Figure 11c, where users can use the non-dominant hand to retrieve a newspaper from the context proxy and place it into their near space. This affordance also allows

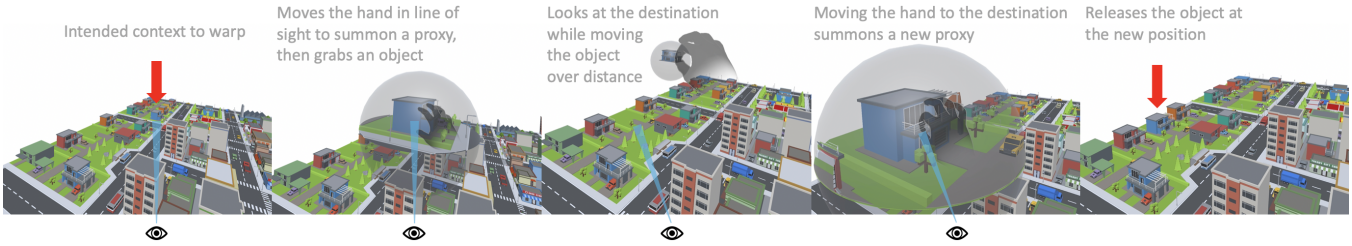

Figure 10: The user performs a HandToGaze drag-and-drop to move the blue building from a starting position to a destination. The eye icon and blue triangle indicate the user's gaze.

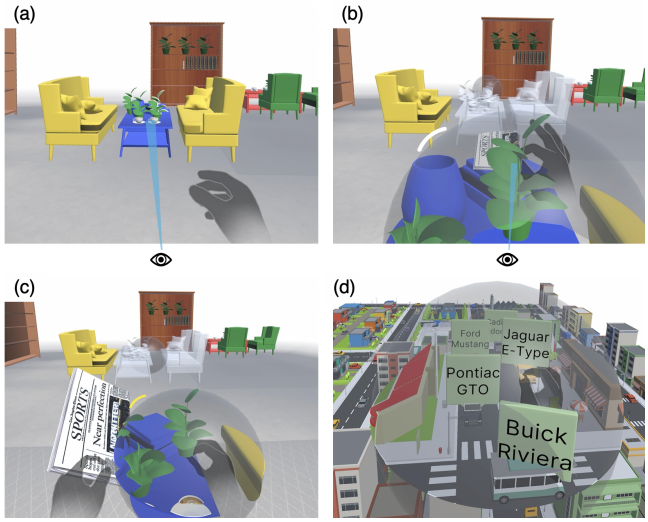

Figure 11: (a)–(b) Using GazeToHand, the user can locate the newspaper that was previously occluded and too small in far space. (c) The user directly acquires the newspaper using the other hand from the summoned near context for further use. (d) Detailed information tags are only rendered when vehicles are summoned into near space. The eye icon and blue triangle indicate the user's gaze.

users to potentially put a new object, using the non-dominant hand, into a context sphere that is summoned into the near space using the dominant hand. In these contexts, the combined use of both hands functions as a distance grab mechanism, providing efficient shortcuts for drag-and-drop operations. This approach could significantly reduce physical effort and speed up the interaction by eliminating the need to traverse between distant spatial contexts.

5.2 Details on Demand (Figure 11d)

SightWarp extends prior WIM techniques in their support for Shneiderman's well-known mantra: "overview first, zoom and filter, then details on demand" [47] through a rapid on-demand creation of WIMs. For example, in an urban design application, summoning a local context can reveal additional details such as annotations or object tags—information that would otherwise cause visual clutter if displayed in the full overview [24]. As shown in Figure 11d, information tags on top of vehicles only appear when the objects are brought into the near context sphere.

5.3 Occluded and Small Objects (Figure 11a–b)

Occluded object selection is a classic challenge for raycasting-based methods especially when users stay in stationary positions and perceive the scene from a single perspective [61]. GazeToHand addresses this challenge by enabling a perspective change as users can pinch their hand at any position in space to summon the selected context to that location, therefore revealing previously occluded objects. Additionally, the scalability of the context allows users to zoom in on the near context and interact with objects which would be too small to select at a distance. Figure 11a–b demonstrates how a newspaper, which is small in size and occluded by a plant, can be easily accessed after summoning the plant's context to near space.

6 Conclusion

As XR operating systems such as Google's AndroidXR and Apple's visionOS increasingly adopt multimodal input, our research informs the design of future UIs that build on the direct-indirect input paradigms and contribute to seamless and efficient user experiences. In particular, our work shows that using eye-hand coordination to trigger proxy summoning and transition to direct manipulation is easy to use and efficient. While far-space interaction will likely remain dominant in many practical XR scenarios, minimizing the effort required to transition to direct manipulation may make near-field interaction a more viable and attractive option.

Our work presents several limitations and directions for future research. First, further studies are needed to explore how SightWarp can be integrated with other interaction modes in realistic, system-wide use cases. While our study demonstrated performance benefits for 3D object manipulation—a core task in spatial environments—it remains to be seen how SightWarp performs in broader applications, such as selection tasks and multi-step workflows illustrated in section 5. Furthermore, we employed a set of self-tested parameters, e.g., for entering the HandToGaze and GazeToHand modes, which can be further optimised for generic UI as well as specifically-tailored application needs. While our approach assumes that gaze-hand alignment is infrequent during indirect gestures based on prior work and common usage patterns, further empirical validation would strengthen this assumption. SightWarp may also be extended with a wider range of proxy summoning methods. For instance, our eye-hand concept could be combined with prior work that uses only eyes vergence to switch between UI depth layers [17, 63]. Finally, as with other summoning techniques [41, 52], there are open challenges to address regarding spatial conflicts when summoned objects overlap with existing ones, which can be potentially exploited for merging near and far context [41].

Acknowledgments

This work was supported by funding from a Google Research Gift award ('Multimodal and Gaze + Gesture Interactions in XR'), the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant no. 101021229 GEMINI), and the Danish National Research Foundation under the Pioneer Centre for AI in Denmark (DNRF grant P1).

References

[1] Isayas Berhe Adhanom, Paul MacNeilage, and Eelke Folmer. 2023. Eye tracking in virtual reality: a broad review of applications and challenges. *Virtual Reality* 27, 2 (2023), 1481–1505.

[2] Joanna Bergström, Tor-Salve Dalsgaard, Jason Alexander, and Kasper Hornbæk. 2021. How to Evaluate Object Selection and Manipulation in VR? Guidelines from 20 Years of Studies. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 533, 20 pages. doi:10.1145/3411764.3445193

[3] Mark Billinghurst, Adrian Clark, Gun Lee, et al. 2015. A survey of augmented reality. *Foundations and Trends® in Human–Computer Interaction* 8, 2-3 (2015), 73–272.

[4] Idil Bostan, Oğuz Turan Buruk, Mert Canat, Mustafa Ozan Tezcan, Celalettin Yurdakul, Tilbe Göksun, and Oğuzhan Özcan. 2017. Hands as a controller: User preferences for hand specific on-skin gestures. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. 1123–1134.

[5] Doug A Bowman and Larry F Hodges. 1997. An evaluation of techniques for grabbing and manipulating remote objects in immersive virtual environments. In *Proceedings of the 1997 symposium on Interactive 3D graphics*. 35–ff.

[6] Doug A Bowman, Ernst Kruijff, Joseph J LaViola, and Ivan Poupyrev. 2001. An introduction to 3-D user interface design. *Presence* 10, 1 (2001), 96–108.

[7] Géry Casiez, Nicolas Roussel, and Daniel Vogel. 2012. 1 € filter: a simple speed-based low-pass filter for noisy input in interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 2527–2530. doi:10.1145/2207676.2208639

[8] Géry Casiez, Daniel Vogel, Ravin Balakrishnan, and Andy Cockburn. 2008. The impact of control-display gain on user performance in pointing tasks. *Human–computer interaction* 23, 3 (2008), 215–250.

[9] Shujie Deng, Nan Jiang, Jian Chang, Shihui Guo, and Jian J Zhang. 2017. Understanding the impact of multimodal interaction using gaze informed mid-air gesture control in 3D virtual objects manipulation. *International Journal of Human-Computer Studies* 105 (2017), 68–80.

[10] Clifton Forlines, Daniel Vogel, and Ravin Balakrishnan. 2006. HybridPointing: fluid switching between absolute and relative pointing with a direct input device. In *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology* (Montreux, Switzerland) *(UIST '06)*. Association for Computing Machinery, New York, NY, USA, 211–220. doi:10.1145/1166253.1166286

[11] Clifton Forlines, Daniel Wigdor, Chia Shen, and Ravin Balakrishnan. 2007. Direct-touch vs. mouse input for tabletop displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 647–656. doi:10.1145/1240624.1240726

[12] Jenny Gabel, Susanne Schmidt, Oscar Ariza, and Frank Steinicke. 2023. Redirecting rays: Evaluation of assistive raycasting techniques in virtual reality. In *Proceedings of the 29th ACM Symposium on Virtual Reality Software and Technology*. 1–11.

[13] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. doi:10.1016/S0166-4115(08)62386-9

[14] Devamardeep Hayatpur, Seongkook Heo, Haijun Xia, Wolfgang Stuerzlinger, and Daniel Wigdor. 2019. Plane, ray, and point: Enabling precise spatial manipulations with shape constraints. In *Proceedings of the 32nd annual ACM symposium on user interface software and technology*. 1185–1195.

[15] Juan David Hincapié-Ramos, Xiang Guo, Paymahn Moghadasian, and Pourang Irani. 2014. Consumed endurance: a metric to quantify arm fatigue of mid-air interactions. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1063–1072.

[16] Ken Hinckley. 2007. Input technologies and techniques. In *The human-computer interaction handbook*. CRC Press, 187–202.

[17] Teresa Hirzle, Jan Gugenheimer, Florian Geiselhart, Andreas Bulling, and Enrico Rukzio. 2019. A Design Space for Gaze Interaction on Head-mounted Displays. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300855

[18] Yanfei Hu Fleischhauer, Hemant Bhaskar Surale, Florian Alt, and Ken Pfeuffer. 2023. Gaze-based Mode-Switching to Enhance Interaction with Menus on Tablets. In *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications* (Tubingen, Germany) *(ETRA '23)*. Association for Computing Machinery, New York, NY, USA, Article 7, 8 pages. doi:10.1145/3588015.3588409

[19] Richard H Jacoby, Mark Ferneau, and Jim Humphries. 1994. Gestural interaction in a virtual environment. In *Stereoscopic Displays and Virtual Reality Systems*, Vol. 2177. SPIE, 355–364.

[20] Jacek Jankowski and Martin Hachet. 2013. A survey of interaction techniques for interactive 3D environments. In *Eurographics 2013-STAR*.

[21] Jaejoon Jeong, Soo-Hyung Kim, Hyung-Jeong Yang, Gun A Lee, and Seungwon Kim. 2023. GazeHand: A Gaze-Driven Virtual Hand Interface. *IEEE Access* 11 (2023), 133703–133716.

[22] Manu Kumar, Jeff Klingner, Rohan Puranik, Terry Winograd, and Andreas Paepcke. 2008. Improving the accuracy of gaze input for interaction. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications* (Savannah, Georgia) *(ETRA '08)*. Association for Computing Machinery, New York, NY, USA, 65–68. doi:10.1145/1344471.1344488

[23] Mikko Kytö, Barrett Ens, Thammathip Piumsomboon, Gun A Lee, and Mark Billinghurst. 2018. Pinpointing: Precise head-and eye-based target selection for augmented reality. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–14.

[24] David Lindlbauer, Anna Maria Feit, and Otmar Hilliges. 2019. Context-Aware Online Adaptation of Mixed Reality Interfaces. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 147–160. doi:10.1145/3332165.3347945

[25] Mathias N Lystbæk, Thorbjørn Mikkelsen, Roland Krisztandl, Eric J Gonzalez, Mar Gonzalez-Franco, Hans Gellersen, and Ken Pfeuffer. 2024. Hands-on, Hands-off: Gaze-Assisted Bimanual 3D Interaction. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–12.

[26] Mathias N Lystbæk, Ken Pfeuffer, Jens Emil Sloth Grønbæk, and Hans Gellersen. 2022. Exploring gaze for assisting freehand selection-based text entry in ar. *Proceedings of the ACM on Human-Computer Interaction* 6, ETRA (2022), 1–16.

[27] Mathias N. Lystbæk, Peter Rosenberg, Ken Pfeuffer, Jens Emil Grønbæk, and Hans Gellersen. 2022. 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* 6, ETRA (May 2022), 1–18. doi:10.1145/3530886

[28] Diako Mardanbegi, Benedikt Mayer, Ken Pfeuffer, Shahram Jalaliniya, Hans Gellersen, and Alexander Perzl. 2019. EyeSeeThrough: Unifying Tool Selection and Application in Virtual Environments. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 474–483. doi:10.1109/VR.2019.8797988

[29] David C. McCallum and Pourang Irani. 2009. ARC-Pad: absolute+relative cursor positioning for large displays with a mobile touchscreen. In *Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology* (Victoria, BC, Canada) *(UIST '09)*. Association for Computing Machinery, New York, NY, USA, 153–156. doi:10.1145/1622176.1622205

[30] Daniel Mendes, Fabio Marco Caputo, Andrea Giachetti, Alfredo Ferreira, and Joaquim Jorge. 2019. A survey on 3d virtual object manipulation: From the desktop to immersive virtual environments. In *Computer graphics forum*, Vol. 38. Wiley Online Library, 21–45.

[31] Mark R Mine. 1995. Virtual environment interaction techniques. *UNC Chapel Hill CS Dept* (1995).

[32] Mark R Mine, Frederick P Brooks Jr, and Carlo H Sequin. 1997. Moving objects in space: exploiting proprioception in virtual-environment interaction. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*. 19–26.

[33] Aunnoy K Mutasim, Anil Ufuk Batmaz, and Wolfgang Stuerzlinger. 2021. Pinch, click, or dwell: Comparing different selection techniques for eye-gaze-based pointing in virtual reality. In *Acm symposium on eye tracking research and applications*. 1–7.

[34] Siyou Pei, David Kim, Alex Olwal, Yang Zhang, and Ruofei Du. 2024. UI Mobility Control in XR: Switching UI Positionings between Static, Dynamic, and Self Entities. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 611, 12 pages. doi:10.1145/3613904.3642220

[35] Ken Pfeuffer, Jason Alexander, Ming Ki Chong, Yanxia Zhang, and Hans Gellersen. 2015. Gaze-Shifting: Direct-Indirect Input with Pen and Touch Modulated by Gaze. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (Charlotte, NC, USA) *(UIST '15)*. Association for Computing Machinery, New York, NY, USA, 373–383. doi:10.1145/2807442.2807460

[36] Ken Pfeuffer, Benedikt Mayer, Diako Mardanbegi, and Hans Gellersen. 2017. Gaze + Pinch Interaction in Virtual Reality. In *Proceedings of the 5th Symposium on Spatial User Interaction* (Brighton, United Kingdom) *(SUI '17)*. Association for Computing Machinery, New York, NY, USA, 99–108. doi:10.1145/3131277.3132180

[37] Jeffrey S Pierce, Andrew S Forsberg, Matthew J Conway, Seung Hong, Robert C Zeleznik, and Mark R Mine. 1997. Image plane interaction techniques in 3D immersive environments. In *Proceedings of the 1997 symposium on Interactive 3D graphics*. 39–ff.

[38] Jeffrey S. Pierce, Brian C. Stearns, and Randy Pausch. 1999. Voodoo dolls: seamless interaction at multiple scales in virtual environments. In *Proceedings of the 1999 symposium on Interactive 3D graphics*. ACM, Atlanta Georgia USA, 141–145. doi:10.1145/300523.300540

[39] Thammathip Piumsomboon, Gun Lee, Robert W Lindeman, and Mark Billinghurst. 2017. Exploring natural eye-gaze-based interaction for immersive virtual reality. In *2017 IEEE symposium on 3D user interfaces (3DUI)*. IEEE, 36–39.

[40] Alexander Plopski, Teresa Hirzle, Nahal Norouzi, Long Qian, Gerd Bruder, and Tobias Langlotz. 2022. The Eye in Extended Reality: A Survey on Gaze Interaction and Eye Tracking in Head-worn Extended Reality. *ACM Comput. Surv.* 55, 3, Article 53 (March 2022), 39 pages. doi:10.1145/3491207

[41] Henning Pohl, Klemen Lilija, Jess McIntosh, and Kasper Hornbæk. 2021. Poros: Configurable Proxies for Distant Interactions in VR. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3411764.3445685

[42] Ivan Poupyrev, Mark Billinghurst, Suzanne Weghorst, and Tadao Ichikawa. 1996. The go-go interaction technique: non-linear mapping for direct manipulation in VR. In *Proceedings of the 9th annual ACM symposium on User interface software and technology*. 79–80.

[43] Ivan Poupyrev, Tadao Ichikawa, Suzanne Weghorst, and Mark Billinghurst. 1998. Egocentric object manipulation in virtual environments: empirical evaluation of interaction techniques. In *Computer graphics forum*, Vol. 17. Wiley Online Library, 41–52.

[44] Markus Sasalovici, Albin Zeqiri, Robin Connor Schramm, Oscar Javier Ariza Nunez, Pascal Jansen, Jann Philipp Freiwald, Mark Colley, Christian Winkler, and Enrico Rukzio. 2025. Bumpy Ride? Understanding the Effects of External Forces on Spatial Interactions in Moving Vehicles. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.48550/ARXIV.2502.16656

[45] Robin Schweigert, Valentin Schwind, and Sven Mayer. 2019. EyePointing: A Gaze-Based Selection Technique. In *Proceedings of Mensch Und Computer 2019* (Hamburg, Germany) *(MuC'19)*. Association for Computing Machinery, New York, NY, USA, 719–723. doi:10.1145/3340764.3344897

[46] Rongkai Shi, Yushi Wei, Xueying Qin, Pan Hui, and Hai-Ning Liang. 2023. Exploring gaze-assisted and hand-based region selection in augmented reality. *Proceedings of the ACM on Human-Computer Interaction* 7, ETRA (2023), 1–19.

[47] B. Shneiderman. 1996. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*. 336–343. doi:10.1109/VL.1996.545307

[48] Shaishav Siddhpuria, Keiko Katsuragawa, James R Wallace, and Edward Lank. 2017. Exploring at-your-side gestural interaction for ubiquitous environments. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. 1111–1122.

[49] Ludwig Sidenmark and Hans Gellersen. 2019. Eye&head: Synergetic eye and head movement for gaze pointing and selection. In *Proceedings of the 32nd annual ACM symposium on user interface software and technology*. 1161–1174.

[50] Frank Steinicke, Timo Ropinski, and Klaus Hinrichs. 2006. Object selection in virtual environments using an improved virtual pointer metaphor. In *Computer Vision and Graphics: International Conference, ICCVG 2004, Warsaw, Poland, September 2004, Proceedings*. Springer, 320–326.

[51] Sophie Stellmach and Raimund Dachselt. 2012. Look & touch: gaze-supported target acquisition. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 2981–2990. doi:10.1145/2207676.2208709

[52] Richard Stoakley, Matthew J. Conway, and Randy Pausch. 1995. Virtual reality on a WIM: interactive worlds in miniature. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '95)*. ACM Press/Addison-Wesley Publishing Co., USA, 265–272. doi:10.1145/223904.223938

[53] Vildan Tanriverdi and Robert JK Jacob. 2000. Interacting with eye movements in virtual environments. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 265–272.

[54] Juan Pablo Wachs, Mathias Kölsch, Helman Stern, and Yael Edan. 2011. Vision-based hand-gesture applications. *Commun. ACM* 54, 2 (2011), 60–71.

[55] Uta Wagner, Matthias Albrecht, Andreas Asferg Jacobsen, Haopeng Wang, Hans Gellersen, and Ken Pfeuffer. 2024. Gaze, Wall, and Racket: Combining Gaze and Hand-Controlled Plane for 3D Selection in Virtual Reality. *Proceedings of the ACM on Human-Computer Interaction* 8, ISS (2024), 189–213.

[56] Uta Wagner, Andreas Asferg Jacobsen, Tiare Feuchtner, Hans Gellersen, and Ken Pfeuffer. 2024. Eye-Hand Movement of Objects in Near Space Extended Reality. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 84, 13 pages. doi:10.1145/3654777.3676446

[57] Uta Wagner, Mathias N Lystbæk, Pavel Manakhov, Jens Emil Sloth Grønbæk, Ken Pfeuffer, and Hans Gellersen. 2023. A fitts' law study of gaze-hand alignment for selection in 3d user interfaces. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[58] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. ACM, New York, NY, USA, 143–146. doi:10.1145/1978942.1978963

[59] Haijun Xia, Sebastian Herscher, Ken Perlin, and Daniel Wigdor. 2018. Spacetime: Enabling Fluid Individual and Collaborative Editing in Virtual Reality. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (Berlin, Germany) *(UIST '18)*. Association for Computing Machinery, New York, NY, USA, 853–866. doi:10.1145/3242587.3242597

[60] Difeng Yu, Xueshi Lu, Rongkai Shi, Hai-Ning Liang, Tilman Dingler, Eduardo Velloso, and Jorge Goncalves. 2021. Gaze-supported 3d object manipulation in virtual reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[61] Difeng Yu, Qiushi Zhou, Joshua Newn, Tilman Dingler, Eduardo Velloso, and Jorge Goncalves. 2020. Fully-Occluded Target Selection in Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics* 26, 12 (2020), 3402–3413. doi:10.1109/TVCG.2020.3023606

[62] Difeng Yu, Qiushi Zhou, Benjamin Tag, Tilman Dingler, Eduardo Velloso, and Jorge Goncalves. 2020. Engaging Participants during Selection Studies in Virtual Reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 500–509. doi:10.1109/VR46266.2020.00071

[63] Chenyang Zhang, Tiansu Chen, Eric Shaffer, and Elahe Soltanaghai. 2024. FocusFlow: 3D Gaze-Depth Interaction in Virtual Reality Leveraging Active Visual Depth Manipulation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 372, 18 pages. doi:10.1145/3613904.3642589

[64] Qiushi Zhou, Brandon Victor Syiem, Beier Li, Jorge Goncalves, and Eduardo Velloso. 2024. Reflected Reality: Augmented Reality through the Mirror. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 4, Article 202 (Jan. 2024), 28 pages. doi:10.1145/3631431