# $H_2E$: Hand, Head, Eye a Multimodal Cascade of Natural Inputs


Khushman Patel* Google

Vrushank Phadnis Google

Eric J Gonzalez Google

Hans Gellersen Lancaster and Aarhus University

Ken Pfeuffer Aarhus University

Mar Gonzalez-Franco† Google


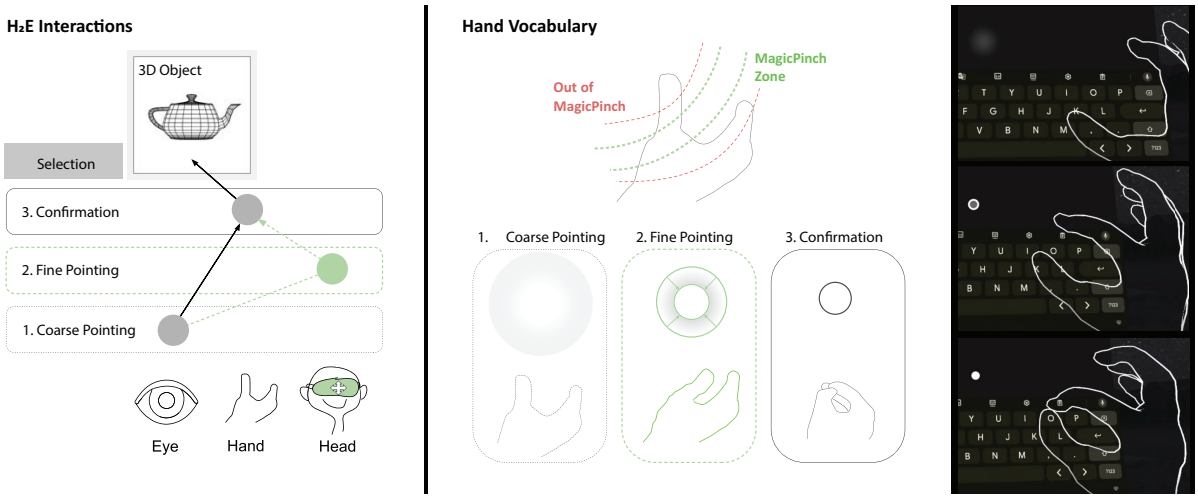


Figure 1: $H_2E$ proposes a cascade of eye and head combined by a hand gesture vocabulary to enable fine grained interactions.


## Abstract

Eye-based interaction techniques for extended reality, such as gaze and pinch, are simple to use however suffer from input precision issues. We present $H_2E$, an integrated fine and coarse-grained pointing framework that cascades Hand, Head, and Eye inputs. We further introduce the MagicPinch gesture as an example of the framework, which allows for smooth transitioning between fine and coarse-grained pointing. When combined together, after users initiate a pinch gesture, a cursor appears midway during the pinch at the position of the gaze, which can be dragged by head pointing if needed before pinch confirmation. This has the advantage that it can add a precision component without changing the semantics of the technique. In this paper, we describe the design of the $H_2E$ framework and implementation of the MagicPinch technique. Furthermore, we present an evaluation of our method in a Fitts-based user study, exploring the speed-accuracy trade-offs against a gaze and pinch interaction baseline.



**Index Terms:** Human-centered computing—User interface toolkits; Human-centered computing—Virtual reality; Human-centered computing—Mixed / augmented reality;


## 1 Introduction

A key component for extended reality (XR) technology is the user interface (UI). Eye-tracking is becoming increasingly relevant for the XR UI as a remote pointer without relying on a handheld controller. It brings in fast target selections coupled with expressive hand gesture manipulation. However, a challenge lies is input precision: eye gaze can be susceptible to precision issues, as of hardware sensing limitations and human physiological constraints [11].

To address this challenge, we leverage design insights from three key works in gaze HCI. First, the cascade of inputs [35]. In here, the role of eye gaze is coarse pointing, and then input is cascaded to the mouse to refine and confirm selections. Cascading inputs between eye motion and hand motion enables the duality of rapidly selecting a target by gaze and click directly, or interacting with the precision mode using gaze and subsequent mouse motion. Second, the concept of eye-head fusion. With considerable work in recent years in fusing eye and head movement together to create gaze-based interaction systems [9, 24, 26], we leverage the principle of coarse and rapid eye pointing complemented by fine-grained head pointing. Third, with using gaze and gestures XR UI, we expand the vocabulary and interaction of a stable gaze interaction technique. The eye-gaze to head-pointing input cascade is enabled by the user's hand gestures, without affecting the basic semantics of the existing gesture set.

We propose $H_2E$, an interaction framework that cascades the modes of coarse eye-gaze selection, fine head-based refinement, and confirmation of selection. We further introduce the MagicPinch gesture as an implementation of the framework which adds a sub-layer to the hand pinch gesture. Using MagicPinch, at a half-way point in the confirmation pinch gesture, users can pause the pinch motion and optionally refine their selection by precise head movement before completing the pinch to confirm selection. Going one level deeper, we can consider the atomic input states based on Buxton's 3-State Model [2] in Fig. 2. $H_2E$ offers simultaneous dual modes of going from a 'tracking' state to 'dragging': directly from coarse positioning to confirmation, or from coarse to fine positioning, to confirmation.

We will now go through the two selection and confirmation paths offered by $H_2E$ using the state diagram illustrated in Figure 1:

1. Direct coarse pointing: The user looks at the element and


*e-mail: khushman@google.com

†e-mail: margon@google.com

completes the pinch without pausing and without head motion, switching from coarse eye pointing selection to confirming selection via pinch. This modality is similar to using the Gaze and Pinch technique [20]. In Figure 1, this would mean going from state (1) to state (3).

2. Coarse + fine pointing: The user begins from state (1). Upon getting close to their target, and being unable to finish selection with just coarse pointing, they switch to fine-tuning state (2) by performing the half-pinch gesture, and upon completing fine-tuning, they would pinch their hand together to confirm selection as in state (3). An example of a user performing this method of selection can be observed in Figure 3 (a),(b), and (c) in order.

To explore $H_2E$ we ran a user study comparing the technique to the established gaze and pinch interaction technique. In the Fitts' Law based user study protocol, we study the time-accuracy trade-off and perceived task-load for the user. Our results show that regarding usability, no significant differences were revealed. As expected, $H_2E$ led to an increased task completion time due to the additional refinement step. Regarding miss-clicks, no significant differences were revealed. But we can see a clear trend that, even if not significant in the current study, all three factors of accuracy, timeouts and miss-clicks are lower with $H_2E$. This points to a promising method for precision input in XR UIs without controllers.

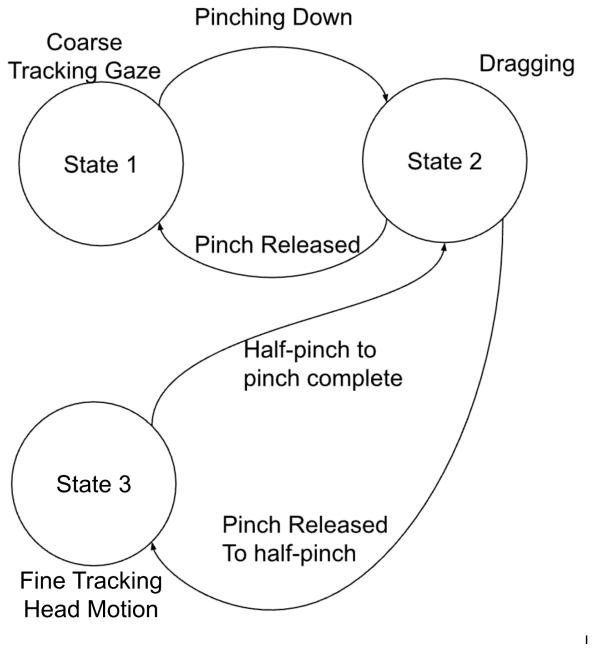


Figure 2: Extended Buxton 3-state model for $H_2E$

## 2 Related Work

Egocentric XR input techniques can be categorised into virtual hand and virtual pointer [23]. Our focus is on virtual pointers for precise interaction over distance, of relevance for the large spaces that XR provides users. Our eyes are considered as convenient, natural, and fast virtual pointer [12, 30]. A body of research has explored interaction techniques where gaze is part of a multimodal fusion with hand-controlled inputs– such as via a handheld controller [3, 21, 34] and bare-hand tracking [15, 16, 20, 22, 31]. A fundamental technique is Gaze+Pinch [20] (e.g., Hololens 2, Apple Vision Pro), where the user looks at a target of interest and executes an indirect pinch to confirm selection. The technique allows to interact over distance with simple hand gestures, and has been shown to improve speed and lower physical effort to hand-gesture-only based raypointing [16, 31]. However, a limitation is the accuracy and precision, making small-target selection difficult. Although eye-trackers have improved greatly in the last decade [5], technical calibration issues and physiological constraints of human eyes demand further research [5, 11, 32].

To address this problem, one line of research investigated a second modality for refinement of the original gaze cursor position. In Zhai's Manual And Gaze Input Cascaded (MAGIC) [35], the eyes represent the coarse pointer and mouse movement refines the cursor position. The cascading of both input modalities allows users to both tackle quick selections of large targets, and precise selection of small targets. Stellmach et al. extended this to large display interaction, where gaze pointing switches to refinement when users perform a touch motion gesture gesture [29]. In XR, Pinpointing [14] proposes a set of precision techniques for gaze pointing with head, hand gesture and controller refinement. For example, one technique is Eye+Head, where holding down a button enters the refinement stage where the head takes over pointing from gaze. In those examples, hand input is necessary for refinement, which conflicts with potential drag & drop actions using hand motion after selection, limiting expressiveness.

Another category of precision techniques enhances precision without engaging the hands in the refinement. HMAGIC [13] uses head pointing for the refinement of a gaze cursor on the desktop screen. WeightedPointer estimates eye-tracking accuracy through an error prediction module, and then implicitly switches from gaze to a fallback modality such as head pointing [27]. A semi-explicit method has been proposed by Sidenmark et al., incorporating head-based refinement by switching between natural and gestural head-eye coordination patterns [25, 26, 28]. In our work, we focus on a more explicit switch by considering the subtleties of a pinch gesture.

Pinching is long considered natural for XR [1] and widely adopted in modern XR headsets. PinchLens [36] builds upon this and proposes to use a "Semi-Pinch" state, where index finger and thumb are close but not in contact yet, to enhance accuracy of pinch gestures. PinchLens extends the family of Bubble cursors [6, 17], by spatially magnifying the area around the pinch center point when starting a pinch gesture. As targets are larger, it becomes easy to select the targets. However, spatial magnification changes the appearance of the scene and can be distracting. In our work, we share the use of a "Semi-Pinch" state to control a cursor which original position is set by gaze.

## 3 $H_2E$ Framework

Our aim is to design a framework that addresses precision input through the multimodal fusion of hand, head, and eyes. We describe our interaction design inspired by the principles proposed in a recent work by Pfeuffer et al. [19].

### 3.1 Division of Labour

There is a clear division of labour between the three modalities of eye gaze, head pointing, and hand pinch: The hand is used for mode switching and confirmation, the eyes travel large distances to coarse point, and the head finetunes if needed. This flexibility allows for an easy mental model: Large motions are achieved naturalistically with the eyes, the fastest modality. If the user is confident in the eye pointing selection, they complete the pinch directly to confirm, passing through the fine-point mode without modification of selection. If not, they pause at the half-pinch gesture to switch modes to precise head-pointing, and after correcting selection, complete the pinch to confirm.

### 3.2 The MagicPinch gesture: Coarse and fine pointing

The coarse/fine pointing duality of $H_2E$ is enabled by what we call the MagicPinch technique (Figure 1). It cascades inputs during selection as the gesture occurs mid-way in the process of confirmation using pinch, i.e. the user will go through the MagicPinch state to be able to confirm selection using hand pinch. This state enables fast mode switching because it introduces no extra steps for hand gestures. This affords to keep the learning curve low, and can integrate with current hand tracking systems with minimal effort. Camera based hand tracking systems in VR were unable to provide stable hand pinch readings, especially with large head movements. To introduce robustness in the MagicPinch gesture, we introduce the banding effect shown in Figure 1. We use time and pinch strength banding to ensure the user stays in the fine-tune state even if we lose tracking for small time periods, i.e. if we lose hand tracking for 150ms, we keep the user in the fine-tune state. To detect the user's intent to pinch, we use the index finger pinch strength value provided by the OVRHand class of the Meta Unity SDK. While Meta's algorithm isn't public, a value of 0 indicates no pinch, and 1 indicates a completed pinch, when the finger touches the thumb. The pinch strength is banded as: if the pinch strength velocity is greater than 0.1/frame for a 60fps Unity application towards a pinch and absolute pinch strength is $> 0.05$ or absolute pinch strength is $> 0.15$, enable fine-tune state. If the pinch strength velocity is greater than 0.1/frame away from a pinch and absolute pinch strength is $< 0.15$ or absolute pinch strength is $< 0.05$, disable fine-tune state. While there is no obvious indicator of how close the user is to the fine-tune state, we instructed users to try out the half-pinch gesture till they observed the strong cursor change visual at the location of the user's eye gaze. With $H_2E$, the user can go for coarse and fine pointing. This flexibility allows for much faster transition and confirmation when the target is well aligned already with the eye gaze, whereas sometimes if there is need for a more fine pointing the users can hold the pinch-state further to enable head pointing selection.

### 3.3 Learning Curve Design

The $H_2E$ framework along with MagicPinch caters to the needs of both novices and experts, one of the principles in the user-centered design community [8]. Novice users prefer a linear workflow that allows them to get the work done while working backwards towards a task. This is easily accomplished when they start using $H_2E$+MagicPinch as described: coarse point with the eyes, half-pinch to switch modes and fine-tune with the head if needed, and confirm by completing hand pinch. But as these users transition to becoming experts, they will want to use the $H_2E$ cascade more efficiently. An expert user might prefer to blend the coarse-pointing, mode-switch, fine-tuning, and confirmation in the course of the same interaction. The layering of Gaze+Pinch, and Gaze+Head+Pinch together in the same $H_2E$ framework, along with the clear division of labour using MagicPinch, provides a straightforward path to mastering the system as they use it.

Switching modes with MagicPinch allows the $H_2E$ cascade to occur simultaneously with eye gaze pointing. This is important because completing a pinch might takes longer than targeting with the eye: ballistic eye movements can be as small as $\sim 0.2s$ [24], while pinches take longer at $\sim 0.9s$ [18]. This allows experts to transition modes during the pointing flow, not needing to wait and judge if the coarse eye tracker was able to target correctly. The eye pointer converting to a head cursor provides strong feedback that input modalities have transitioned. Eye gaze pointing is implicitly connected with head movement [24, 25], so when the cursor transitions, it follows the same general direction the eyes were following, since the eyes and head move in concert while being offset to each other. This continuation of the same path the eye pointer was tracing introduces a strong cascading effect, where users can operate as if the head pointing is an extension of the same eye movement, and don't need to rely on their judgement to assess if the eye tracker itself worked well. A trained user could start a half-pinch before their eye gaze selection was complete, trusting the head pointer to complete the selection more than coarse eye gaze pointing so they can seamless fine-tune the selection with their head movement right before confirmation. The strong visual indicator at the location of their vision also helps a user know if they are in eye-tracking, or head-pointing mode. This is also supported by the fact that the user goes into a half-pinch before they are able to complete a full pinch. This combination works together seamlessly as it allows each modality to do what it does best.

### 3.4 Feedback

Feedback is useful to let users know which input mode they are in, or might be transitioning into [26]. Similar to the recent findings on methods with fast switching between modes [10], the feedback might not be as important when jumping from coarse pointing to confirmation, but proves useful when needing to switch to the explicit fine-pointing state. The need for feedback is slightly alleviated due to the user explicitly switching modes when necessary, and the fast switching gesture of MagicPinch which occurs half-way during confirmation. Nevertheless, in our approach we provide direct feedback on this mode switch with a smart cursor state (Figure 1). Cursor states depend on the hand gesture of the user, whether they are doing coarse eye input, or enabling head fine pointing, or directly confirming with the pinch. During coarse eye input, we aim to reduce the Midas Touch problem, and use a hazy cursor to indicate to the user where the eye tracking system is detecting their input. Upon switching to the fine-tune state, the cursor switches to an opaque donut, a strong visual indicator that appears exactly where the user is gazing, making it hard to miss. Upon confirmation, the donut collapses into an opaque circle, indicating that the user is confirming selection at that location.

Apart from the distinct cursor states, we also provide constant feedback to the user about where they are in the confirmation process by continuously manipulating both, the hazy coarse eye pointer, and the fine-tuning head pointing donut. The higher the pinch strength readings of the user, the smaller the coarse eye pointer and the fine-tuning donut get towards their center. This mechanism reinforces to the user where they are confirming, and at what level of confirmation of selection they are. Continuous feedback to the user when they make pinch-based hand motion through their selection pointer/cursor might also help overcome the noisiness in modern camera based VR hand tracking, as the user should be able to detect and correct incorrect hand tracking readings by repositioning their hands.

### 3.5 Head-Eye Gaze Input cascading

To make the eye-head transition seamless, the eye gaze pointer converts into a head-movement controlled donut cursor at the location of the Eye gaze pointer on entering the refinement state. Then, users control via head - the control-display gain applied to the head pointer is: 1.3 * distance between the center of the user's eyes and the donut cursor. The feedback serves to reinforce the different modes of input while still appearing to be one consistent style. When making large eye movements and switching to refinement mode, the cursor will smoothly follow the same path as your eyes, creating a seamless and continuous input.

## 4 Evaluation

To evaluate $H_2E$, we used a within subjects experiment ran based on Fitts' law. We compared the following conditions in a randomised order. First, $H_2E$: Participants used their gaze for either coarse and fine selection and had the option to enable head tracking for fine selection. Selection confirmation was done using two finger pinching gesture. Second, *Gaze+Pinch*: Gaze was used for both

coarse and fine selection and selection confirmation was done using a two finger pinching gesture.

### 4.1 Study Design

We use a ISO 9241-9 based Fitts' style study protocol showing targets in circular pattern on a 2D plane one meter away. Each target plane consisted of seven targets equally spaced apart. We chose three target sizes (31.25mm, 62.5mm and 125mm) and two target distances (700mm and 500mm). This resulted in six index of difficulty (ID). We selected our target size and distances to provide an equal spread in ID.

In the beginning of each experiment, participants went through an eye tracking calibration. Next they were shown a demonstration of the input system and then provided with a warm-up round to try using each input (warm-up trial data was discarded). Participants spent an hour in the experiment and were sedentary throughout the study.

To avoid learning effects a randomized sequence of target planes varying in target size and distance was presented. Each experiment included two runs of seven targets for a given target size and target distance. If a target could not be selected for more than 10 seconds, that attempt was considered timed-out and logged accordingly. Each target was shown initially shown in gray. The trail target was shown in green and upon selection the green target moved to its next location. Miss click was denoted in red.

We used a Meta Quest Pro as our prototyping vehicle. As per Wei et al. [33], it has an average accuracy of 1.652° with a precision of 0.699° (standard deviation). Device calibration was done using the in-built eye calibration process at every don-on and don-off of the headset.

We primarily used logging based data to derive our input performance metrics. In addition, self-reported NASA-TLX responses were used to compute physical and mental workload [7].

### 4.2 Participants

Ten participants took part in our study. All participants were part of [Anon.] and were recruited according to [Anon.] internal ethics protocol and appropriate consent process was followed. No user wore glasses. Participant demographics were spread across age and job roles. Six participants identified as female, 4 as male. Nine were right handed and most participants (9/10) had little to no virtual reality (VR) experience.

## 5 Results

Perhaps the main result in favour of $H_2E$+MagicPinch, is its increased accuracy when compared to Gaze+Pinch. When we look at the accuracy distribution for $H_2E$ (M=0.53, SD=1.06) (Fig. 4) and Gaze+Pinch (M=0.71, SD=1.84) for the different target locations and target sizes of Fitts, we find that while Gaze+Pinch fails more in the outer periphery of user's field of view particularly when targets are set towards the upper half of the horizon. This effect is less noticeable with $H_2E$ and accuracy is more consistent in all directions.

Timeouts were higher in the Gaze+Pinch condition (M=4.7, SD=6.68) as compared to $H_2E$ (M=1.8, SD=2.49). Given this difference, we show median time to selection in the subsequent sections instead of average times. We found that $H_2E$ (M=0.53, SD=1.06) produced fewer miss-clicks than Gaze+Pinch (M=0.71, SD=1.84), but $H_2E$ was slower than Gaze+Pinch. However, only the speed difference was found to be statistically significant on a one-tailed t-test $t(9) = 2.325$, $p = 0.03$. The median time to select compared to the ID can be seen in Fig. 5. The results from the NASA TLX show no significant differences for individual questions and overall task-load for $H_2E$ (M=3.92, SD=1.09) and Gaze+Pinch (M=3.55, SD=0.82).

## 6 Discussion and Conclusions

In the current work we demonstrated the working principles of the $H_2E$ framework, implemented using the MagicPinch gesture, and tested it in a user study of 10 participants using a target acquisition Fitts-law based test. When comparing $H_2E$+MagicPinch to regular Gaze+Pinch, we found $H_2E$ seemed to lesser input failures (timeouts and miss clicks) and improved performance in the periphery of the users field of view, where eye tracking is normally harder (above the horizon line) as noted in Fig. 4.

The findings may depend on the quality of the eye tracker for the particular users and devices. And as eye trackers improve we might not need as much the use of $H_2E$. In our pool of participants had varied levels of eye tracking performance. Our study gave first insights into the time-accuracy trade-off and a natural next step would be to conduct a larger sample experiment for a deeper investigation. One caveat to the $H_2E$ technique is that it was significantly slower (median time to select) as compared to Gaze+Pinch. We suspect this could be due to the additional head refinement step slowing down the user when they aren't completely confident in their selection, preferring to correct their selection instead of attempting to press as with Gaze+Pinch. Further investigation with a larger sample size would shed light on the reason for this. However since with the Magic Pinch implementation, it is up to the user to enable the head cascading or not, we predict that users would in general not need to use it and only enable this accurate fine pointing when needing higher precision of selection. Typical eye trackers only provide 0.5°accuracy under optimal conditions [4], and as we move towards wider adoption of XR eye tracking, $H_2E$ could enable extremely precise eye-tracking applications with high accuracy, e.g. 3D modelling.

Overall, we believe the $H_2E$ input cascading framework can be implemented easily to any XR device with hand and eye tracking, as all of them incorporate head tracking, which would augment the accuracy of input. The MagicPinch gesture additionally minimally expands the vocabulary of interactions from Gaze+Pinch, a valuable step for future research to enable fine pointing with natural inputs. The initial results indicate that our technique of cascading head-refinement with coarse gaze-pointing seems to improve acquisition of smaller targets, which could mean that it would particularly enable users with poor eye-tracking quality, interaction tasks with many small targets, use cases with particular accessibility or accuracy needs, and for potential Gaze+Pinch UIs that have not been designed by default for Gaze+Pinch interaction.

## References


[1] D. Bowman, C. Wingrave, J. Campbell, and V. Ly. Using pinch gloves (tm) for both natural and abstract interaction techniques in virtual environments. 2001.

[2] W. Buxton et al. A three-state model of graphical input. In *Human-computer interaction-INTERACT*, vol. 90, pp. 449–456, 1990.

[3] D. L. Chen, M. Giordano, H. Benko, T. Grossman, and S. Santosa. Gazeraycursor: Facilitating virtual reality target selection by blending gaze and controller raycasting. In *Proceedings of the 29th ACM Symposium on Virtual Reality Software and Technology*, VRST '23. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3611659.3615693

[4] A. M. Feit, S. Williams, A. Toledo, A. Paradiso, H. Kulkarni, S. Kane, and M. R. Morris. Toward everyday gaze input: Accuracy and precision of eye tracking and implications for design. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, p. 1118–1130. Association for Computing Machinery, New York, NY, USA, 2017. doi: 10.1145/3025453.3025599

[5] A. S. Fernandes, T. S. Murdison, and M. J. Proulx. Leveling the playing field: A comparative reevaluation of unmodified eye tracking as an input and interaction modality for vr. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2269–2279, 2023.

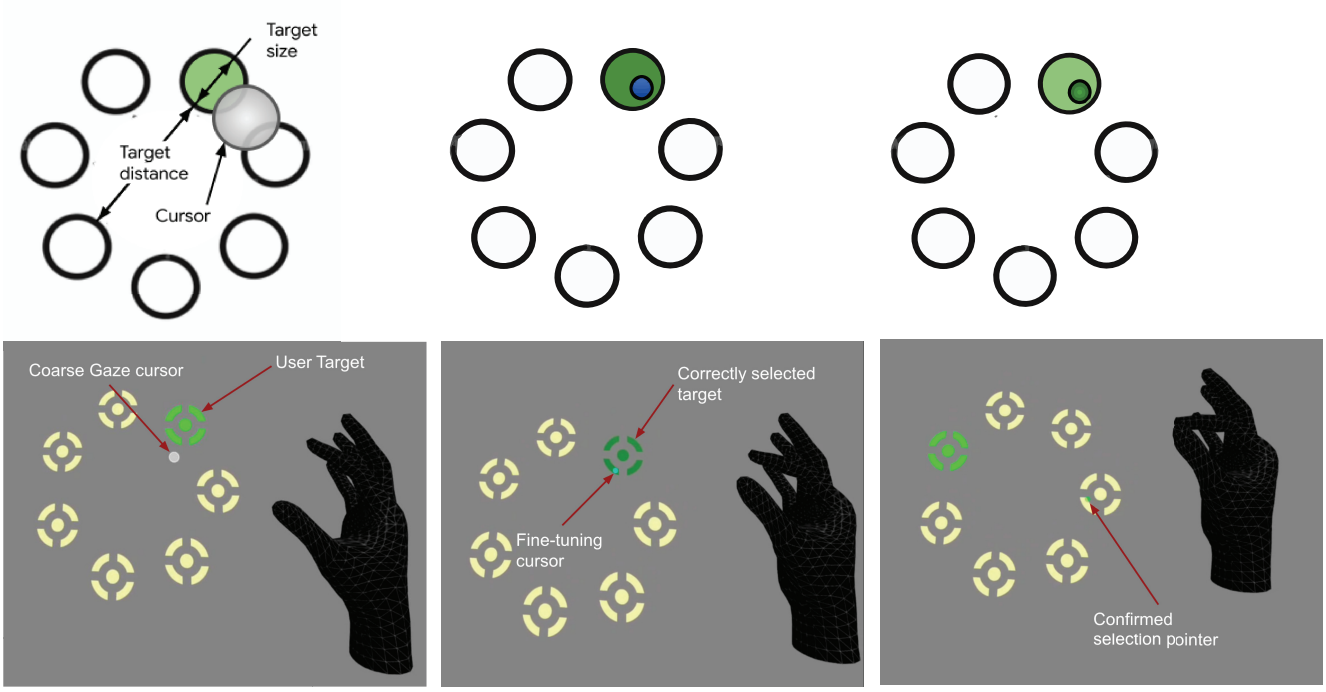


Figure 3: Fitts' task implementation of $H_2E$ from left to right: (a) Describing the Fitts' task setup with coarse pointing (b) User refines with half-pinch and with head movement (c) User confirming selection with full-pinch

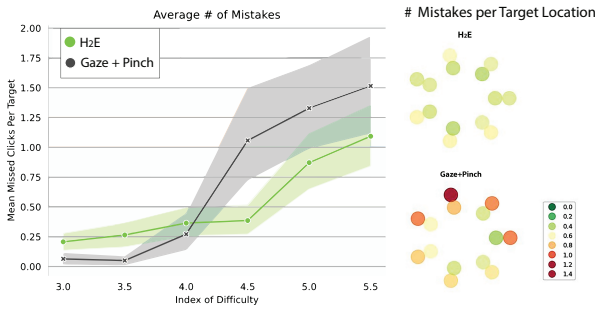


Figure 4: Accuracy distribution. Average number of miss clicks per index of difficulty for $H_2E$ and Gaze+Pinch.

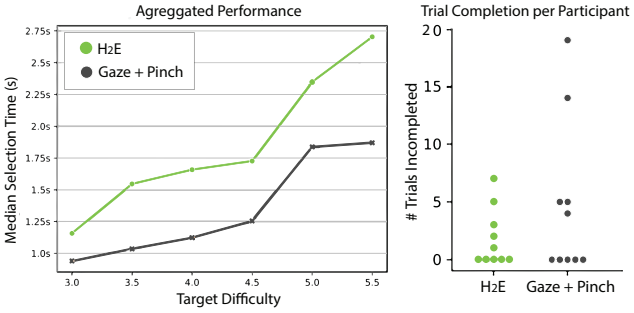


Figure 5: (Left) Median time for selections with the two technique conditions, (Right) Scatterplot for timeouts during trials.

[6] T. Grossman and R. Balakrishnan. The bubble cursor: enhancing target acquisition by dynamic resizing of the cursor's activation area. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, p. 281–290. Association for Computing Machinery, New York, NY, USA, 2005. doi: 10.1145/1054972.1055012

[7] S. G. Hart. Nasa-task load index (nasa-tlx); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(9):904–908, 2006. doi: 10.1177/154193120605000909

[8] D. A. Hooper. Are we designing products for experts or just really good beginners? *Procedia Manufacturing*, 3:166–172, 2015. 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015. doi: 10.1016/j.promfg.2015.07.122

[9] B. J. Hou, J. Newn, L. Sidenmark, A. Ahmad Khan, P. Bækgaard, and H. Gellersen. Classifying head movements to separate head-gaze and head gestures as distinct modes of input. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3544548.3581201

[10] B. J. Hou, J. Newn, L. Sidenmark, A. A. Khan, and H. Gellersen. Gazeswitch: Automatic eye-head mode switching for optimised hands-free pointing. *Proc. ACM Hum.-Comput. Interact.*, 8(ETRA), May 2024. doi: 10.1145/3655601

[11] R. Jacob and S. Stellmach. What you look at is what you get: gaze-based user interfaces. *Interactions*, 23(5):62–65, aug 2016. doi: 10.1145/2978577

[12] R. J. Jacob. The use of eye movements in human-computer interaction techniques: what you look at is what you get. *ACM Transactions on Information Systems (TOIS)*, 9(2):152–169, 1991.

[13] A. Kurauchi, W. Feng, C. Morimoto, and M. Betke. Hmagic: head movement and gaze input cascaded pointing. In *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '15. Association for Computing Machinery, New York, NY, USA, 2015. doi: 10.1145/2769493.2769550

[14] M. Kytö, B. Ens, T. Piumsomboon, G. A. Lee, and M. Billinghurst. Pinpointing: Precise head- and eye-based target selection for augmented reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, p. 1–14. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3173574.

3173655
[15] M. N. Lystbæk, K. Pfeuffer, J. E. S. Grønbæk, and H. Gellersen. Exploring gaze for assisting freehand selection-based text entry in ar. *Proc. ACM Hum.-Comput. Interact.*, 6(ETRA), may 2022. doi: 10.1145/3530882
[16] M. N. Lystbæk, P. Rosenberg, K. Pfeuffer, J. E. Grønbæk, and H. Gellersen. Gaze-hand alignment: Combining eye gaze and mid-air pointing for interacting with menus in augmented reality. *Proc. ACM Hum.-Comput. Interact.*, 6(ETRA), may 2022. doi: 10.1145/3530886
[17] M. E. Mott and J. O. Wobbrock. Beating the bubble: using kinematic triggering in the bubble lens for acquiring small, dense targets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, p. 733–742. Association for Computing Machinery, New York, NY, USA, 2014. doi: 10.1145/2556288.2557410
[18] A. K. Mutasim, A. U. Batmaz, and W. Stuerzlinger. 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*, ETRA '21 Short Papers. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3448018.3457998
[19] K. Pfeuffer. Design Principles for Gaze + Pinch Interaction Design. `https://medium.com/@ken.pfeuffer/a95e251169ae#8609`, 2024. [Online; accessed 17-January-2024].
[20] K. Pfeuffer, B. Mayer, D. Mardanbegi, and H. Gellersen. Gaze + pinch interaction in virtual reality. In *Proceedings of the 5th Symposium on Spatial User Interaction*, SUI '17, p. 99–108. Association for Computing Machinery, New York, NY, USA, 2017. doi: 10.1145/3131277.3132180
[21] K. Pfeuffer, L. Mecke, S. Delgado Rodriguez, M. Hassib, H. Maier, and F. Alt. Empirical evaluation of gaze-enhanced menus in virtual reality. In *26th ACM Symposium on Virtual Reality Software and Technology*, VRST '20. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3385956.3418962
[22] K. Pfeuffer, J. Obernolte, F. Dietz, V. Mäkelä, L. Sidenmark, P. Manakhov, M. Pakanen, and F. Alt. Palmgazer: Unimanual eye-hand menus in augmented reality. In *Proceedings of the 2023 ACM Symposium on Spatial User Interaction*, SUI '23. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3607822.3614523
[23] I. Poupyrev, T. Ichikawa, S. Weghorst, and M. Billinghurst. Egocentric object manipulation in virtual environments: empirical evaluation of interaction techniques. In *Computer graphics forum*, vol. 17, pp. 41–52. Wiley Online Library, 1998.
[24] L. Sidenmark and H. Gellersen. Eye, head and torso coordination during gaze shifts in virtual reality. *ACM Trans. Comput.-Hum. Interact.*, 27(1), dec 2019. doi: 10.1145/3361218
[25] L. Sidenmark and H. Gellersen. 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*, UIST '19, p. 1161–1174. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3332165.3347921
[26] L. Sidenmark, D. Mardanbegi, A. R. Gomez, C. Clarke, and H. Gellersen. Bimodalgaze: Seamlessly refined pointing with gaze and filtered gestural head movement. In *ACM Symposium on Eye Tracking Research and Applications*, ETRA '20 Full Papers. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3379155.3391312
[27] L. Sidenmark, M. Parent, C.-H. Wu, J. Chan, M. Glueck, D. Wigdor, T. Grossman, and M. Giordano. Weighted pointer: Error-aware gaze-based interaction through fallback modalities. *IEEE Transactions on Visualization and Computer Graphics*, 28(11):3585–3595, 2022. doi: 10.1109/TVCG.2022.3203096
[28] L. Sidenmark, D. Potts, B. Bapisch, and H. Gellersen. Radi-eye: Hands-free radial interfaces for 3d interaction using gaze-activated head-crossing. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3411764.3445697
[29] S. Stellmach and R. Dachselt. Look & touch: Gaze-supported target acquisition. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pp. 2981–2990. ACM, New York, NY, USA, 2012. doi: 10.1145/2207676.2208709
[30] V. Tanriverdi and R. J. K. Jacob. Interacting with eye movements in virtual environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '00, p. 265–272. Association for Computing Machinery, New York, NY, USA, 2000. doi: 10.1145/332040.332443
[31] U. Wagner, M. N. Lystbæk, P. Manakhov, J. E. S. Grønbæk, K. Pfeuffer, and H. Gellersen. 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*, CHI '23. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3544548.3581423
[32] C. Ware and H. H. Mikaelian. An evaluation of an eye tracker as a device for computer input2. In *Proceedings of the SIGCHI/GI Conference on Human Factors in Computing Systems and Graphics Interface*, CHI '87, p. 183–188. Association for Computing Machinery, New York, NY, USA, 1986. doi: 10.1145/29933.275627
[33] S. Wei, D. Bloemers, and A. Rovira. A preliminary study of the eye tracker in the meta quest pro. In *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences*, IMX '23, p. 216–221. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3573381.3596467
[34] D. Yu, X. Lu, R. Shi, H.-N. Liang, T. Dingler, E. Velloso, and J. Goncalves. Gaze-supported 3d object manipulation in virtual reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3411764.3445343
[35] S. Zhai, C. Morimoto, and S. Ihde. Manual and gaze input cascaded (magic) pointing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, p. 246–253. Association for Computing Machinery, New York, NY, USA, 1999. doi: 10.1145/302979.303053
[36] F. Zhu, L. Sidenmark, M. Sousa, and T. Grossman. Pinchlens: Applying spatial magnification and adaptive control-display gain for precise selection in virtual reality. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 1221–1230. IEEE, 2023.

## A Appendix

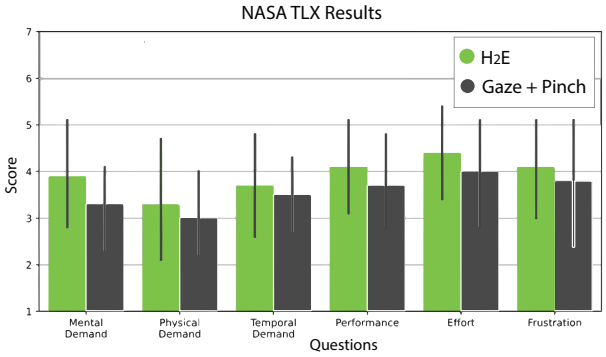


Figure 6: NASA TLX ratings for $H_2E$ and Gaze+Pinch.