© 2026 IEEE. This is the author's version of the article that has been published in the adjunct proceedings of IEEE ISMAR conference. The final version of this record is available at: 10.1109/ISMAR-Adjunct73196.2026.00049

Tracing Across Modalities: Investigating Alternative Interaction Techniques for Gesture Typing in Extended Reality

Egor Baydarov*
Ekaterina Trukhan†
Egor Shibanov‡
HSE University

Pavel Manakhov§
Lancaster University

ABSTRACT

Headset-based extended reality (XR) supports a wide variety of tasks, yet fast text entry without physical devices remains a challenge. This work focuses on freehand gesture typing in XR and introduces two novel multimodal text entry techniques: GAZE&PINCH and HEADGAZE&PINCH. Both techniques combine hand gestures for word tracing with either eye- or head-based pointing for initial character selection. The techniques were evaluated in two user studies against established manual text entry methods and a multimodal baseline (GESTURETYPE). Results showed that while manual techniques achieved higher text entry rates, multimodal techniques yielded comparable error rates, with GESTURETYPE receiving favorable user feedback. The findings highlight both the promise and limitations of multimodal gesture typing, allowing us to derive practical recommendations for designing word-gesture text entry methods in XR.

Index Terms: Text entry, gesture typing, extended reality.

1 INTRODUCTION & RELATED WORK

Headset-based extended reality (XR) presents a versatile medium that supports a wide range of activities, from immersive gaming and virtual socializing to productivity-oriented tasks. While XR is increasingly used in contexts that resemble traditional desk work, such as focused tasks that involve sitting and the use of a keyboard [3], typical interaction in XR happens in mid-air. Users commonly rely on handheld controllers or hand gestures performed in front of the body, often without any physical support. Reliance on hand gestures and lack of haptic feedback can make precise input challenging. Despite this, text entry remains a frequent need in XR, for example, when searching for apps, browsing the web, or replying to a text without taking the headset off. In this work, we focus on this common context of entering short phrases in XR, without access to physical input devices such as keyboards or controllers.

Gesture typing — entering text by connecting the letters of each word on a virtual keyboard using a continuous gesture, rather than selecting them individually — has been shown to be tolerant of imprecise mid-air interactions [4], and it demonstrates higher text entry performance compared to selection-based methods in XR [30]. Initially introduced for mobile devices [10, 31], mid-air gesture typing has been extensively studied in 3D contexts with various techniques employing hand gestures [17, 1, 4] or head orientation [30, 12] for word tracing. Recently, word-gesture keyboards have been integrated into XR platforms such as Windows Holographic and Meta Horizon OS. However, both research efforts and commercial implementations have primarily focused on unimodal interaction with such keyboards. For instance, Dudley et al. [4] studied the performance of word-gesture keyboards when words were traced by directly touching virtual keys. Gesture typing has also been investigated in far-field interaction contexts, where words are traced by projecting the user's hand position onto the surface of the keyboard [17], or using hand-based raycasting [4].

Gesture typing techniques that involve multiple input modalities have been largely underexplored. Notable exceptions are the work of Zhao et al. [32] which introduced a technique that uses implicit gaze input to speed up a manually operated cursor towards the fixated key, and Yu et al. [30]'s technique that uses a head cursor for pointing and a button hold to indicate the duration of a word gesture. However, our understanding of gesture typing methods where path drawing incorporates multiple modalities, for example, using gaze to specify the gesture's starting point while hand movements trace the word, remains limited.

To fill in this knowledge gap, we introduce and empirically examine two multimodal techniques for gesture typing:

- GAZE&PINCH (Figure 1a)[1]: the user's gaze selects initial characters. A pinch gesture — touching the tips of the index finger and thumb — transfers control to the hand, which movement is projected onto the keyboard's surface to draw a path. Releasing the pinch returns control to the gaze. This method supports entering single-letter words and interacting with controls like 'Backspace' via a brief pinch. It is inspired by the general-purpose technique introduced by Pfeuffer et al. [22].

- HEADGAZE&PINCH (Figure 1b): resembles GAZE&PINCH, except that it uses a head cursor (also known as a reticle) to select initial characters. The technique is inspired by other existing head gaze-based techniques [19, 18].

To better understand how combining hand with eye or head input shapes gesture typing, we conducted two studies that serve a dual purpose: first, *to evaluate* GAZE&PINCH *and* HEADGAZE&PINCH *against existing text entry techniques*; and second, *to better understand how these established approaches differ from one another*, as they have not previously been examined side by side.

Specifically, we compared the two techniques with: GESTURETYPE (Figure 1c), which relies on head movements for both selecting initial characters and tracing words, with the pinch gesture used to indicate the drawing state [30]; FIXEDORIGINPOINTING (Figure 1d), where selection and drawing — distinguished by pinching — are performed using a ray originating from the user's shoulder and passing through the midpoint of the user's palm[2]; and DIRECT-

*e-mail: eg@byda.io. This and the following two authors contributed equally to this research.
†e-mail: ekaterinatrukhan@icloud.com.
‡e-mail: shibanov.yegor@gmail.com.
§e-mail: p.manakhov@lancaster.ac.uk (corresponding author). The work was done when the author was at HSE University.

[1] In this paper, we name the text entry methods after the general-purpose interaction techniques that they resemble.

[2] It is an existing text entry technique that has been implemented in Meta Horizon OS since v56 (https://mixed-news.com/en/hands-on-with-meta-quest-swipe-keyboard-and-other-v56-upgrades/). The name of the technique is adopted from [9].

© 2026 IEEE. This is the author's version of the article that has been published in the adjunct proceedings of IEEE ISMAR conference. The final version of this record is available at: 10.1109/ISMAR-Adjunct73196.2026.00049

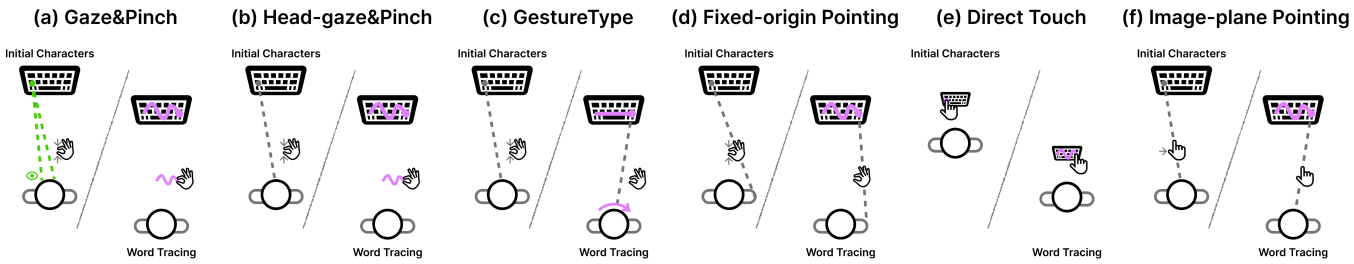


Figure 1: Entering words using gesture typing involves two steps: first, locating the initial characters (images on the left), and second, drawing paths by connecting the letters of each word (images on the right).

TOUCH (Figure 1e), near-field method where words are traced by touching the virtual keyboard with the tip of the index finger [4]. In addition to the three baseline techniques, we compare IMAGEPLANEPOINTING — a far-field method inspired by [23], where pointing is performed by intersecting a cursor, located just above the tip of the index finger of the outstretched hand, with a virtual key within the image plane (Figure 1f). Drawing in this technique is indicated by bringing the thumb closer to the palm. Inspired by [14], we prepared a phrase set of 432 sentences closely representing Russian language to conduct the text entry experiments.

The results of the studies showed that purely manual techniques outperformed multimodal ones in terms of text entry rates while maintaining similar error rates. At the same time, participants' perception of GESTURETYPE was favorable and comparable to that of manual techniques. HEADGAZE&PINCH was perceived as the least usable due to difficulties in alternating between head gaze and hand movements for input. Perception of GAZE&PINCH was more nuanced.

The contributions of this work are threefold: (1) the novel GAZE&PINCH and HEADGAZE&PINCH interaction techniques, which use distinct modalities for selecting initial characters and tracing words, and (2) insights gained through the empirical comparison of six text entry techniques. Additionally, (3) we make the phrase set designed to aid in carrying out text entry experiments conducted in Russian publicly available.

2 COMPARATIVE STUDY OF THE TEXT ENTRY TECHNIQUES

We conducted an experiment using a mixed 6 × 7 factorial design with the following factors:

- *Technique* (between-subjects factor): GAZE&PINCH, HEADGAZE&PINCH, GESTURETYPE, FIXEDORIGINPOINTING, DIRECTTOUCH, and IMAGEPLANEPOINTING.
- *Session* (within-subjects factor): seven consecutive text entry sessions.

2.1 Experimental Task

To evaluate the performance of various text entry methods, participants were randomly assigned one technique and asked to transcribe short phrases from memory one-by-one. Participants were instructed to enter phrases "as quickly and accurately as possible". Corrections were allowed but limited to the most recently entered word, which could be removed using the 'Backspace' key or replaced using predictive text suggestions provided by the keyboard interface (Figure 2). To submit a completed phrase, they pressed the 'Done Typing' button.

Phrase Set. The phrases were carefully selected to suit the experimental context. Sourced from OpenCorpora [21], they adhered to specific constraints: a maximum of 60 characters, no punctuation, numbers, abbreviations, or proper nouns. Linguistically, the set aligned closely with natural Russian usage, demonstrating an average word length of 5.81 characters and exhibiting a near-perfect match with standard letter frequency patterns ($r_s(31) = .99$, $p < .0001$). The full set of 432 phrases can be found in the supplementary materials. Phrases did not repeat during the experiment.

2.2 Study Setup

We used an HTC Vive CE virtual reality (VR) headset with a Deluxe Audio Strap. To enable hand tracking, a Leap Motion controller was attached to the front plate of the headset at a 20-degree downward angle from the horizon line. For eye tracking, we used a Droolon F1 add-on[3] connected directly to the headset. The headset was connected to a PC equipped with a GeForce GTX 1060 GPU and 32 GB of RAM. For this study, we adopted a seated setup using a non-swiveling chair without armrests.

The prototype for this study was developed in Unity. The experiment was conducted in a neutral virtual environment with a dark background for contrast (Figure 2). A virtual keyboard with a standard Russian layout (ЙЦУКЕН) was implemented consistently across all six text entry techniques. To provide reliable gesture typing experience, we developed a subsystem that translated each path drawn on the surface of the virtual keyboard into input for a Gboard keyboard [5] running on Android in an emulator. The subsystem returned both the entered phrase and predictions for the last word, which were then displayed in the Unity prototype (delay was consistently below 10ms). We ensured that all words in our phrase set were included in Gboard's dictionary.

The text entry techniques were implemented according to the designs described in Section 1. Parameters for each technique were empirically tuned by varying them and pilot testing with users to maximize performance:

Keyboard placement For all techniques except DIRECTTOUCH, the keyboard was positioned 2 meters in front of the participant at eye level. In the DIRECTTOUCH technique, the keyboard was placed 35 cm from the user's head at a 40-degree downward angle from the horizon line, mimicking the approach from [4]. In all conditions, the keyboard was oriented to face the user.

Keyboard size The pilot study showed that participants achieved the highest performance with a keyboard size of 30 degrees of visual angle diagonally[4] for GESTURETYPE, FIXEDORIGINPOINTING, IMAGEPLANEPOINTING, and HEADGAZE&PINCH. For GAZE&PINCH, maximum performance was achieved at 50 degrees, and for DIRECTTOUCH, at 45 degrees.

Control-display gains The optimal control-display (CD) gain for HEADGAZE&PINCH was 5, i.e., 1 unit of hand movement

[3] https://www.7invensun.com/droolonf1xqy

[4] Degrees of visual angle are used as a unified measure of size because they are independent of distance.

© 2026 IEEE. This is the author's version of the article that has been published in the adjunct proceedings of IEEE ISMAR conference. The final version of this record is available at: 10.1109/ISMAR-Adjunct73196.2026.00049



(a) (b)

Figure 2: To enter the word "this" ("это") using HEADGAZE&PINCH, a participant (a) targets the first letter with the head cursor (represented by a white dot) and then, following a pinch gesture, (b) draws a path (purple line) connecting the remaining letters of the word by moving their hand.

yielded 5 units of cursor movement. The same gain, scaled to keyboard size, was used for GAZE&PINCH. For FIXEDORIGINPOINTING, CD gain was set by the ray's fixed origin, 0.25 m behind the shoulder and 0.3 m below the head.[5] In GESTURETYPE, ray movement matched head rotation (1:1), and in DIRECTTOUCH, input and output coincided, with the cursor projected from the index fingertip (CD gain = 1).

For all techniques except GAZE&PINCH, the cursor was visualized as a white dot (Figure 2a). During path drawing, the cursor left behind a purple trace that gradually faded (Figure 2b). In addition to visual feedback, audio cues, delivered through external stereo speakers, signaled key transitions as well as the start and end of a word-gesture.

In GAZE&PINCH, which relies on the line of sight for pointing, the white cursor dot was omitted to prevent participants from "chasing" the gaze-attached pointer [15]. To address the issue of late triggering — when the gaze leaves a key before confirmation [25] — we introduced a fixed time offset of 200 ms, chosen empirically.

2.3 Participants

Thirty volunteers participated in the study (9 women and 21 men). Their ages ranged between 18 and 27 years, with an average age of 20.5. All participants were regular users of computers and mobile devices and had normal or corrected-to-normal vision. Three participants identified as left-handed. Each was familiar with the ЙЦУКЕН keyboard layout and could type on it from memory. None have used gesture typing and/or VR on a regular basis.

2.4 Dependent Variables

The Unity prototype we developed logged all interactions with the text entry techniques and, based on that, automatically calculated objective measures, which are detailed below.

Text Entry Rate

Text entry efficiency was measured in words per minute (WPM) using the following equation [13]:

$$Text\ Entry\ Rate = \frac{|P|}{T} \cdot \frac{60}{5} \quad (1)$$

[5]Moving the origin backward improves accuracy but reduces speed, and vice versa.

where $|P|$ represents the length of the phrase entered by the participant (i.e., the number of characters), and T is the entry time in seconds[6].

Error Rates

To gain insight into the accuracy of the text entry techniques, we calculated three measures: **word-level total error rate**, **word-level uncorrected error rate**, and **character-level not corrected error rate**. All dependent variables were derived from an analysis of the input stream, i.e., the step-by-step entry of the transcribed phrase, as described in [26]. For the word-level measures, the unit of analysis was the word; otherwise, the calculations were identical to those used for the character-level measures.

Subjective Evaluation

To assess perceived aspects of interaction, we used the System Usability Scale (SUS) questionnaire [11], translated into Russian, along with a custom questionnaire designed to measure properties specific to the studied text entry techniques, such as **learnability, efficiency, accuracy, comfort of drawing gestures, comfort of pointing, overall comfort, and overall usability**. Please see Appendix B for more detail.

2.5 Procedure

Every experimental session began with a participant signing an informed consent form, followed by a completion of a demographic questionnaire. This was followed by a preparatory phase where they were fitted with a VR headset and given time to explore interactive VR experiences to become comfortable with VR and hand tracking. For those in the IMAGEPLANEPOINTING condition, we conducted an eye dominance test to determine their dominant eye, then aligned the cursor accordingly to reduce parallax-related distortions [29].

Following this, participants proceeded to a training session. This stage introduced the interaction method assigned to them, demonstrating how to use it to operate a virtual keyboard. The training also included a practice round in which participants entered eight phrases.

During the main part of the experiment, each participant completed a total of seven sessions. Each session consisted of eight phrases, followed by a one-minute break before the next began. If the headset was taken off and before each GAZE&PINCH session, eye tracking was calibrated using standard HTC Vive's 5-point calibration procedure. At the conclusion of the final session, participants completed the questionnaires. This was followed by a debriefing discussion aimed at gathering open-ended feedback on their experience. In total, each experimental run lasted approximately 90 minutes. Formal ethics approval was not required under applicable institutional regulations.

2.6 Results

We conducted two-way mixed ANOVAs for each objective dependent variable, with *Session* as a within-subject factor and *Technique* as a between-subject factor ($\alpha = 0.05$). When the sphericity assumption was violated, we applied the Greenhouse-Geisser correction. Holm-corrected pairwise comparisons were performed as needed.

To analyze subjective variables, we employed the Kruskal-Wallis non-parametric test using *Technique* as a between-subject factor ($\alpha = 0.05$).

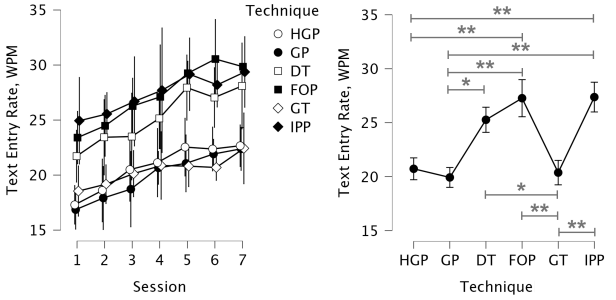
Results of the statistical analysis can be found in Appendix A. Findings are presented below by dependent variable.

[6]Although the average Russian word length is 6 characters, we used 5 as the divisor in Equation 1 to keep our text entry rates comparable to English-based studies.

© 2026 IEEE. This is the author's version of the article that has been published in the adjunct proceedings of IEEE ISMAR conference. The final version of this record is available at: 10.1109/ISMAR-Adjunct73196.2026.00049

2.6.1 Text Entry Rate

The entry rates for the six input techniques across the seven sessions are shown in Figure 3a. A mixed ANOVA revealed significant main effects of both *Technique* ($F_{5,24} = 9.864$, $p < .001$) and *Session* ($F_{6,144} = 21.558$, $p < .001$), with the latter indicating improvement over the course of the study. The interaction between the two factors was not significant ($F_{30,144} = .454$, $p = .993$), meaning that from session to session performance improved at a similar rate across all techniques. Table 1 presents the text entry rates for the first and last sessions, averaged across all participants.



(a) (b)

Figure 3: Mean text entry rates for the six techniques across seven sessions (a) and overall (b). Error bars indicate 95% confidence intervals. Technique abbreviations: HEADGAZE&PINCH (HGP), GAZE&PINCH (GP), DIRECTTOUCH (DT), FIXEDORIGINPOINTING (FOP), GESTURETYPE (GT), and IMAGEPLANEPOINTING (IPP). Statistical significance: * for $p < .05$ and ** for $p < .01$.

The results indicate that the text entry techniques can be divided into two groups — FIXEDORIGINPOINTING, IMAGEPLANEPOINTING, and DIRECTTOUCH in the first group, and HEADGAZE&PINCH, GESTURETYPE, and GAZE&PINCH in the second. Techniques from the first group generally outperformed those from the second, with the exception of the DIRECTTOUCH-HEADGAZE&PINCH pair, as shown by the post-hoc comparison results (see Figure 3b).

Table 1: Entry rate improvements from the first to the last session. Values are reported as mean ± standard deviation.

	Session 1, WPM	Session 7, WPM	Improvement, %
FIXEDORIGINPOINTING	23.40±3.51	29.85±7.11	27.56
IMAGEPLANEPOINTING	24.94±3.71	29.37±3.42	17.76
DIRECTTOUCH	21.72±2.63	28.10±2.96	29.37
HEADGAZE&PINCH	17.30±1.62	22.67±1.82	31.04
GESTURETYPE	18.55±1.76	22.43±5.19	20.92
GAZE&PINCH	16.86±1.27	22.38±1.06	32.74

2.6.2 Error Rates

A mixed ANOVA did not reveal any significant main or interaction effects for any of the three dependent variables that characterize text entry accuracy (see Appendix A for detail). Therefore, we report only the grand means.

The grand mean of the **character-level uncorrected error rate** was 2.14%. The grand means for the **uncorrected and total word-level error rates** were 7.54% and 17.89%, respectively. Both values are, as expected, higher than the character-level error rate, because a single incorrect letter renders an entire word incorrect and there are fewer words than characters in a sentence.

2.6.3 Subjective Evaluation

Neither the analysis of the SUS nor that of our custom questionnaire revealed statistically significant differences among the *Techniques* (see Appendix A for details).

3 Follow-up Study Focused on User Feedback

3.1 Study Design

To gather subjective feedback from participants who experienced all six text entry techniques, we conducted a follow-up study employing a within-subjects design, with *Technique* as a single independent variable. While the structure of this study was similar to that of the main experiment, it differed in that it included only two sessions per technique — training and experimental — with each session consisting of five phrases. The custom questionnaire was administered at the end of the entire experiment, which for within-subject study allowed participants to reflect on the entire experience and compare all techniques directly. Presentation of the techniques was counterbalanced among participants using balanced Latin Square.

We recruited 12 participants (9 male, 3 female), aged 18–24, who had not participated in the first study. All participants had normal or corrected-to-normal vision, were right-handed, and were proficient in the ЙЦУКЕН keyboard layout.

3.2 Results

The results presented below were analyzed using repeated measures ANOVA for objective measures and the Friedman test for subjective ones with *Technique* as a within-subject factor ($\alpha = 0.05$). We used post-hoc tests with Holm correction for pairwise analysis.

3.2.1 Text Entry and Error Rates

RM-ANOVA revealed a significant main effect of *Technique* for **text entry rate** ($F_{5,55} = 13.99$, $p < .001$), but not for **character-level uncorrected error rate** ($F_{5,55} = 1.801$, $p = .128$). The pattern observed for text entry rate was similar to that in Session 1 of the first experiment where IMAGEPLANEPOINTING demonstrated highest performance (Table 1). The results differed, however, in that FIXEDORIGINPOINTING had a lower rate in the follow-up, yielding a significant difference between IMAGEPLANEPOINTING and FIXEDORIGINPOINTING ($p = .02$).

3.2.2 Questionnaire Results

A Friedman test did not reveal a significant main effect of *Technique* on **overall comfort** ($\chi^2(5) = 10.791$, $p = .056$) but it did show a significant main effect of *Technique* on **learnability** ($\chi^2(5) = 24.847$, $p < .001$), **efficiency** ($\chi^2(5) = 22.158$, $p < .001$), **accuracy** ($\chi^2(5) = 19.317$, $p = .002$), **comfort of pointing** ($\chi^2(5) = 13.602$, $p = .018$), **comfort of drawing gestures** ($\chi^2(5) = 13.365$, $p = .020$), and **overall usability** ($\chi^2(5) = 16.635$, $p = .005$). Median values for all techniques and their pairwise comparisons for these dependent variables are shown in Figure 4.

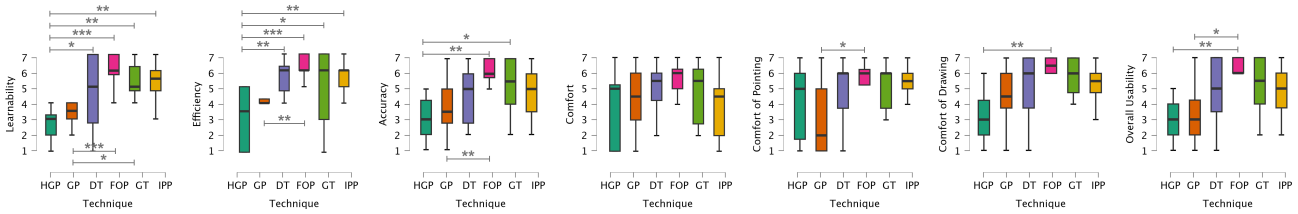
3.2.3 Qualitative Feedback

For all techniques except GESTURETYPE, participants mentioned that they expected their hand to tire from prolonged use. Nine out of twelve participants reported this for IMAGEPLANEPOINTING (P2–3, P5, P7–12), while for other techniques, the number of participants who raised this concern ranged from two to three.

Four participants mentioned that GESTURETYPE could be tiring on the neck with prolonged use (P2, P6, P11–12). At the same time, another group of participants (P1, P7, P9–10) expressed the opinion that gesturing with the head was as easy for them as using the hand.

Selecting keys at the periphery. For GESTURETYPE, two participants observed that it was harder for them to select keys located farther from the keyboard center. They attributed this to being in a hurry and more frequently missing the targets due to the momentum the head gains toward the outer limits of the keyboard (P6, P11). Two participants also mentioned that, due to the keyboard size and reduced eye-tracking accuracy at the periphery in

© 2026 IEEE. This is the author's version of the article that has been published in the adjunct proceedings of IEEE ISMAR conference. The final version of this record is available at: 10.1109/ISMAR-Adjunct73196.2026.00049



(a) (b) (c) (d) (e) (f) (g)

Figure 4: Median learnability (a), efficiency (b), accuracy (c), overall comfort (d), comfort of pointing (e), comfort of drawing word-gestures (f), and overall usability (g) for the six techniques. Technique abbreviations: HEADGAZE&PINCH (HGP), GAZE&PINCH (GP), DIRECTTOUCH (DT), FIXEDORIGINPOINTING (FOP), GESTURETYPE (GT), and IMAGEPLANEPOINTING (IPP). Statistical significance: * for $p < .05$, ** for $p < .01$, and *** for $p < .001$.

the GAZE&PINCH condition, it was difficult to select the first characters located farther from the keyboard center (P1, P8).

Modality switching. Eight participants reported that switching from head to hand input in HEADGAZE&PINCH was confusing (P1–3, P5–6, P8–9, P11). Additionally, during the first sessions, we observed that some participants attempted to move their heads after pinching, before realizing that it was their hand now controlling the cursor[7]. Notably, only four participants mentioned the same issue with GAZE&PINCH (P2, P5, P8, P11). Participant P11 explicitly noted that the switch affected both the selection of the first characters and word-gesture drawing. As a result, they rated the comfort of drawing with HEADGAZE&PINCH lower than with GAZE&PINCH, despite both techniques using the same modality for gesture drawing.

Depth perception. We observed that while using DIRECTTOUCH, all participants had trouble maintaining appropriate depth when drawing word-gestures. Their fingers would move outside the acceptable margins that the keyboard was designed to recognize as touch input (2 cm below and 3 cm above the keyboard surface). Five participants explicitly mentioned this in their feedback (P1, P3, P5–P6, P12).

4 DISCUSSION

Below, we discuss the results obtained throughout the course of the two studies.

Manual Techniques

The techniques that involve only hand gestures, specifically FIXEDORIGINPOINTING and IMAGEPLANEPOINTING, were found to be more efficient according to objective measures (Figure 3b) and were perceived as such by the participants (Figure 4b). The parallax effect did not seem to be an issue for IMAGEPLANEPOINTING. However, the technique was perceived as more tiring, according to participant feedback, which was not unexpected [29]. While the two techniques achieved the same text entry performance throughout multiple sessions of the main experiment, the follow-up study demonstrated lower first-time performance of FIXEDORIGINPOINTING. Since our study involved participants with little to no XR experience, this hints towards lower immediate usability of the technique. While both IMAGEPLANEPOINTING and FIXEDORIGINPOINTING are conceptually simple and were easily understood during training sessions, FIXEDORIGINPOINTING involves raycasting, which may require some getting used to, as opposed to tracing in the image plane. While hand raycasting has been extensively studied in contexts of general target selection, selection-based text entry, and gesture typing [2, 4, 9], its immediate usability is not yet fully understood and requires more research.

[7]This behaviour was observed in the first study too.

Statistically, DIRECTTOUCH, which differs from the two hand-based techniques in both keyboard size (45 vs. 30 degrees) and placement (near-field vs. far-field), did not differ from them performance-wise. This result contradicts the findings of a recent study by Dudley et al. [4], who found that the DIRECTTOUCH technique outperforms FIXEDORIGINPOINTING. This discrepancy can be attributed to two factors: (1) differences in the study setup — namely, the length of the experiment (ours was longer) and the mechanism for detecting touch — and (2) the fact that the researchers used standard implementations of both techniques from the Meta Interaction SDK without tuning their parameters, as we did. In our DIRECTTOUCH implementation, touch detection required users to keep the fingertip within a 5 cm-thick interaction box, which participants occasionally struggled to do during word tracing. Together, these factors may explain why DIRECTTOUCH 's performance fell between that of the manual and multimodal alternatives, while tuning the ray origin position may have improved FIXEDORIGINPOINTING 's performance relative to prior literature.

Multimodal Techniques

The proposed techniques — HEADGAZE&PINCH and GAZE&PINCH — demonstrated inferior performance compared to FIXEDORIGINPOINTING and IMAGEPLANEPOINTING (though comparable to GESTURETYPE) and were rated lower across multiple subjective measures than all other techniques. One factor that may have contributed to the lower performance of these two techniques is the generally lower efficiency of eye and head gaze modalities for pointing, compared to manual input [6]. This could have been compensated by higher performance, compared to GESTURETYPE, during word tracing, which according to our analysis of interaction logs constitutes 62.54% of the phrase entry time (defined as the cumulative time between the start and end of a gesture for all words in a phrase). However, this potential benefit, especially for HEADGAZE&PINCH, was offset by participant-reported confusion caused by alternating between head and hand movements.

The challenge of switching between head and hand input can be explained by motor planning remapping. According to the CASE model [20], GESTURETYPE is a "concurrent" technique: the head is responsible for both pointing and tracing, while a hand gesture (pinch) marks the beginning and end of the word gesture, leading to no such errors. In contrast, HEADGAZE&PINCH is an "alternate" technique in which pointing is performed by the head and tracing by the hand. This alternation could have introduced two sources of difficulty. First, head pointing relies on an integration of visual and proprioceptive feedback, whereas hand tracing projected onto a surface with a CD gain is largely dominated by visual feedback [17]. Switching between them likely requires remapping the spatial goal into distinct motor plans. Second, head pointing is primarily a rotational movement, while hand tracing is primarily translational. Alternating means engaging different motor control

© 2026 IEEE. This is the author's version of the article that has been published in the adjunct proceedings of IEEE ISMAR conference. The final version of this record is available at: 10.1109/ISMAR-Adjunct73196.2026.00049

systems with different dynamics, increasing the likelihood of interference or confusion during switching [7].

GAZE&PINCH was perceived by most participants as a better alternative. It suffered significantly less from switching from gaze to hand input, most probably due to the closer integration between these modalities [22]. In terms of overall usability, it differed unfavorably only from FIXEDORIGINPOINTING. However, the results for comfort of pointing revealed that participants had trouble selecting initial characters using gaze. While we took measures to improve gaze pointing accuracy (i.e., a larger keyboard size and a fixed offset to account for late-trigger errors), it still was not on par with head gaze or manual pointing. Gaze filtering [27, 16] could have helped address this issue. As alternating between gaze and hand gesturing was perceived as more natural, we can speculate that, with higher gaze pointing accuracy, GAZE&PINCH could demonstrate better results. However, it is unlikely that it would reach the performance level of the manual techniques.

GESTURETYPE demonstrated lower performance compared to the manual techniques (except for DIRECTTOUCH), which is not surprising given the biomechanics of head pointing [24, 8]. The higher dispersion in the subjective 'Efficiency' measure observed in the follow-up study (Figure 4b) suggests that some participants noticed the performance difference. However, it did not affect their overall positive perception of the technique (Figure 4g). GESTURETYPE is the only technique out of the six that does not require movement of the hand for typing. Pinch gesture can be sensed even outside of the working volume of the headset's cameras using, for instance, electromyography-based bracelets [28], allowing to enter text without raising hands. This, combined with our findings, makes GESTURETYPE a promising alternative to manual techniques in scenarios where hand movement is constrained. However, further research is needed to evaluate the techniques under such conditions.

General Text Entry Performance

It is noteworthy that compared to prior literature that explored some of the techniques, our study revealed their performance to be higher. For instance, GESTURETYPE demonstrated an average text entry rate of 19.04 WPM in the sixth session of an experiment with a similar setup [30], which is considerably lower than the 22.43 WPM observed in our study. However, expert participants in that study achieved a performance of 24.73 WPM in the eighth session, entering the same phrases repeatedly [30]. In [4], DIRECTTOUCH showed a performance of only 23.4 WPM (block 4–8 average) compared to our 28.1 WPM. In the same study, FIXEDORIGINPOINTING demonstrated a performance of 11.9 WPM in the first session, as opposed to 21.72 WPM in our study. One possible explanation is that we used an existing advanced word-gesture decoder — the Gboard keyboard [5] — instead of implementing the decoder ourselves, as was done in prior studies. Specifically for FIXEDORIGINPOINTING, differences in ray implementation and parameter tuning could have played a significant role in improving efficiency.

Design Recommendations

We summarize design recommendations informed by the study's findings as follows:

- Tuning technique parameters matters, as demonstrated by FIXEDORIGINPOINTING that delivers high text entry performance comparable to DIRECTTOUCH.
- Although conceptually simple, the immediate usability of FIXEDORIGINPOINTING must be taken into account when designing general-purpose spatial interfaces and text entry techniques.
- Switching between head and hand movements for selecting initial characters and word tracing, as in HEADGAZE&PINCH, should be avoided, as this modality alternation leads to confusion, especially when only starting to use the technique.
- Head movement is a well-perceived alternative input modality for gesture typing when hand movement is limited, as demonstrated by GESTURETYPE.

Limitations

This study has several limitations that should be acknowledged. First, while we identified clear patterns in text entry and error rates, other dependent variables in the first experiment, such as subjective measures, may have benefited from a larger participant sample. Second, because all seven sessions of the first experiment were conducted consecutively on the same day, the findings primarily reflect participants' initial experiences with the text entry techniques. Additional research is needed to explore long-term performance and adaptation. Finally, issues with word tracing in DIRECTTOUCH may have influenced the results. Although trials affected by these issues were discarded and repeated to ensure they did not impact objective data, they could still have led participants to adopt more cautious tracing strategies and could have affected subjective evaluations, despite our instructions to disregard technical problems when rating the techniques.

5 Conclusion

This research introduced GAZE&PINCH and HEADGAZE&PINCH, novel multimodal techniques for gesture typing in XR, and empirically compared them with four representative methods. Our central finding is a cautionary note for the design of future text entry interfaces: while purely manual techniques demonstrated the highest efficiency (up to 29.85 WPM), requiring users to alternate modalities (head and hand) for a single continuous gesture is detrimental to usability, leading to confusion and poor user perception. Conversely, the success of the head-based GESTURETYPE technique confirms that non-manual, controller-free gesture input is a well-perceived and viable alternative when hand use is impractical or constrained.

AI Tool Disclosure

Generative AI models, specifically a locally installed Gemma3, were used to assist with proofreading the text of this paper. The authors reviewed and retain full responsibility for all content. No generative AI is listed as an author.

Acknowledgments

The authors wish to thank Dmitry Amelin for his help with the study. The last author would like to thank Stanislava Manakhova for her support during this work.

References

[1] G. Benoit, G. M. Poor, and A. Jude. Bimanual Word Gesture Keyboards for Mid-air Gestures. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, pp. 1500–1507. Association for Computing Machinery, New York, NY, USA, May 2017. doi: 10.1145/3027063.3053137 1

[2] A. Bhatia, M. H. Mughrabi, D. Abdlkarim, M. Di Luca, M. Gonzalez-Franco, K. Ahuja, and H. Seifi. Text Entry for XR Trove (TEXT): Collecting and Analyzing Techniques for Text Input in XR. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pp. 1–18. Association for Computing Machinery, New York, NY, USA, Apr. 2025. doi: 10.1145/3706598.3713382 5

© 2026 IEEE. This is the author's version of the article that has been published in the adjunct proceedings of IEEE ISMAR conference. The final version of this record is available at: 10.1109/ISMAR-Adjunct73196.2026.00049

[3] Y. F. Cheng, A. Carden, H. Cho, C. G. Fidalgo, J. Wieland, and D. Lindlbauer. Augmented Reality In-the-Wild: Usage Patterns and Experiences of Working with AR Laptops in Real-World Settings, Feb. 2025. doi: 10.48550/arXiv.2502.14241 1

[4] J. J. Dudley, J. Zheng, A. Gupta, H. Benko, M. Longest, R. Wang, and P. O. Kristensson. Evaluating the Performance of Hand-Based Probabilistic Text Input Methods on a Mid-Air Virtual Qwerty Keyboard. *IEEE Transactions on Visualization and Computer Graphics*, 29(11):4567–4577, Nov. 2023. doi: 10.1109/TVCG.2023.3320238 1, 2, 5, 6

[5] Google. Gboard - the Google Keyboard - Apps on Google Play. https://play.google.com/store/apps/details?id=com.google.android.inputmethod.latin&hl=en. 2, 6

[6] J. P. Hansen, V. Rajanna, I. S. MacKenzie, and P. Bækgaard. A Fitts' law study of click and dwell interaction by gaze, head and mouse with a head-mounted display. In *Proceedings of the Workshop on Communication by Gaze Interaction*, pp. 1–5. ACM, Warsaw Poland, June 2018. doi: 10.1145/3206343.3206344 5

[7] D. Henriques and J. D. Crawford. Role of Eye, Head, and Shoulder Geometry in the Planning of Accurate Arm Movements. *Journal of Neurophysiology*, 87(4):1677–1685, Apr. 2002. doi: 10.1152/jn.00509.2001 6

[8] R. J. Jagacinski and D. L. and Monk. Fitts' Law in Two Dimensions with Hand and Head Movements Movements. *Journal of Motor Behavior*, 17(1):77–95, Mar. 1985. doi: 10.1080/00222895.1985.10735338 6

[9] R. Jota, M. Nacenta, J. Jorge, S. Carpendale, and S. Greenberg. A Comparison of Ray Pointing Techniques for Very Large Displays. Technical Report 2009-942-21, University of Calgary, Calgary, AB. Canada, Sept. 2009. 1, 5

[10] P.-O. Kristensson. Design and evaluation of a shorthand aided soft keyboard. Master's thesis, Linköping University, Sweden, 2002. 1

[11] J. R. Lewis and J. Sauro. The Factor Structure of the System Usability Scale. In M. Kurosu, ed., *Human Centered Design*, pp. 94–103. Springer, Berlin, Heidelberg, 2009. doi: 10.1007/978-3-642-02806-9_12 3

[12] X. Lu, D. Yu, H.-N. Liang, and J. Goncalves. iText: Hands-free Text Entry on an Imaginary Keyboard for Augmented Reality Systems. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, UIST '21, pp. 815–825. Association for Computing Machinery, New York, NY, USA, Oct. 2021. doi: 10.1145/3472749.3474788 1

[13] I. S. MacKenzie and R. W. and Soukoreff. Text Entry for Mobile Computing: Models and Methods,Theory and Practice. *Human–Computer Interaction*, 17(2-3):147–198, Sept. 2002. doi: 10.1080/07370024.2002.9667313 3

[14] I. S. MacKenzie and R. W. Soukoreff. Phrase sets for evaluating text entry techniques. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '03, pp. 754–755. Association for Computing Machinery, New York, NY, USA, Apr. 2003. doi: 10.1145/765891.765971 2

[15] P. Majaranta and K.-J. Räihä. Twenty years of eye typing: Systems and design issues. In *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications*, ETRA '02, pp. 15–22. Association for Computing Machinery, New York, NY, USA, Mar. 2002. doi: 10.1145/507072.507076 3

[16] P. Manakhov, L. Sidenmark, K. Pfeuffer, and H. Gellersen. Filtering on the Go: Effect of Filters on Gaze Pointing Accuracy During Physical Locomotion in Extended Reality. *IEEE Transactions on Visualization and Computer Graphics*, 30(11):7234–7244, Nov. 2024. doi: 10.1109/TVCG.2024.3456153 6

[17] A. Markussen, M. R. Jakobsen, and K. Hornbæk. Vulture: A mid-air word-gesture keyboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pp. 1073–1082. Association for Computing Machinery, New York, NY, USA, Apr. 2014. doi: 10.1145/2556288.2556964 1, 5

[18] Microsoft. Head-gaze and commit - Mixed Reality Design Guidelines. https://learn.microsoft.com/en-us/windows/mixed-reality/design/gaze-and-commit-head. 1

[19] M. R. Mine. Virtual Environment Interaction Techniques. Technical Report, University of North Carolina at Chapel Hill, USA, Apr. 1995. 1

[20] L. Nigay and J. Coutaz. A design space for multimodal systems: Concurrent processing and data fusion. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, CHI '93, pp. 172–178. Association for Computing Machinery, New York, NY, USA, May 1993. doi: 10.1145/169059.169143 5

[21] OpenCorpora. Open corpus for Russian language. https://opencorpora.org/. 2

[22] 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, pp. 99–108. Association for Computing Machinery, New York, NY, USA, 2017. doi: 10.1145/3131277.3132180 1, 6

[23] J. S. Pierce, A. S. Forsberg, M. J. Conway, S. Hong, R. C. Zeleznik, and M. R. Mine. Image plane interaction techniques in 3D immersive environments. In *Proceedings of the 1997 Symposium on Interactive 3D Graphics*, I3D '97, pp. 39–ff. Association for Computing Machinery, New York, NY, USA, Apr. 1997. doi: 10.1145/253284.253303 2

[24] R. G. Radwin, G. C. Vanderheiden, and M.-L. Lin. A Method for Evaluating Head-Controlled Computer Input Devices Using Fitts' Law. *Human Factors*, 32(4):423–438, Aug. 1990. doi: 10.1177/001872089003200405 6

[25] D. D. Salvucci and J. R. Anderson. Intelligent gaze-added interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '00, pp. 273–280. Association for Computing Machinery, New York, NY, USA, Apr. 2000. doi: 10.1145/332040.332444 3

[26] R. W. Soukoreff and I. S. MacKenzie. Metrics for text entry research: An evaluation of MSD and KSPC, and a new unified error metric. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pp. 113–120. Association for Computing Machinery, New York, NY, USA, Apr. 2003. doi: 10.1145/642611.642632 3

[27] O. Špakov. Comparison of eye movement filters used in HCI. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '12, pp. 281–284. Association for Computing Machinery, New York, NY, USA, Mar. 2012. doi: 10.1145/2168556.2168616 6

[28] D. Tan, D. Morris, S. Saponas, and R. Balakrishnan. Recognizing gestures from forearm EMG signals, May 2013. 6

[29] 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, pp. 1–15. Association for Computing Machinery, New York, NY, USA, Apr. 2023. doi: 10.1145/3544548.3581423 3, 5

[30] C. Yu, Y. Gu, Z. Yang, X. Yi, H. Luo, and Y. Shi. Tap, Dwell or Gesture? Exploring Head-Based Text Entry Techniques for HMDs. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pp. 4479–4488. Association for Computing Machinery, New York, NY, USA, May 2017. doi: 10.1145/3025453.3025964 1, 6

[31] S. Zhai and P.-O. Kristensson. Shorthand writing on stylus keyboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pp. 97–104. Association for Computing Machinery, New York, NY, USA, Apr. 2003. doi: 10.1145/642611.642630 1

[32] M. Zhao, A. M. Pierce, R. Tan, T. Zhang, T. Wang, T. R. Jonker, H. Benko, and A. Gupta. Gaze Speedup: Eye Gaze Assisted Gesture Typing in Virtual Reality. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, pp. 595–606. Association for Computing Machinery, New York, NY, USA, Mar. 2023. doi: 10.1145/3581641.3584072 1

© 2026 IEEE. This is the author's version of the article that has been published in the adjunct proceedings of IEEE ISMAR conference. The final version of this record is available at: 10.1109/ISMAR-Adjunct73196.2026.00049

A Results of the Statistical Analysis for the Main Study

Table 2: Statistical analysis of the effects for all variables of the main experiment. The effect names are abbreviated as follows: *Technique* (T), *Session* (S), *Technique*×*Session* (T×S).

Objective Measures

Variable	ANOVA		
	Effect	F-value	p
Entry Rate			
Text entry rate	T	F(5, 24)=9.864	<.001
	S	F(6, 144)=21.558	<.001
	T×S	F(30, 144)=0.454	.993
Error Rates			
Character-level uncorrected error rate	T	F(5, 24)=1.039	.418
	S	F(6, 144)=0.841	.540
	T×S	F(30, 144)=0.744	.828
Word-level uncorrected error rate	T	F(5, 24)=0.871	.515
	S	F(4.015, 96.361)=0.501	.736
	T×S	F(20.075, 96.361)=0.655	.860
Word-level total error rate	T	F(5, 24)=1.488	.231
	S	F(6, 144)=0.344	.912
	T×S	F(30, 144)=0.805	.752

Subjective Measures

Variable	Kruskal-Wallis	
	H-value	p
System Usability Scale		
SUS	$\chi^2(5)$=5.417	.367
Custom Questionnaire		
Learnability	$\chi^2(5)$=8.853	.115
Efficiency	$\chi^2(5)$=2.540	.771
Accuracy	$\chi^2(5)$=3.245	.662
Comfort	$\chi^2(5)$=6.865	.231
Comfort of drawing	$\chi^2(5)$=3.882	.567
Comfort of pointing	$\chi^2(5)$=4.862	.433
Overall usability	$\chi^2(5)$=10.102	.072

B Custom Questionnaire

To gain insight into the perceived usability of the text entry techniques, we developed a custom questionnaire. The statements translated from Russian are as follows:

1. **Learnability**: It was easy for me to learn to enter text using this input method.
2. **Efficiency**: I entered text slowly using this input method.
3. **Accuracy**: My actions always led to entering the desired word.
4. **Overall comfort**: It was physically difficult for me to enter text using this input method.
5. **Comfort of drawing gestures**: It was easy for me to draw gestures.
6. **Comfort of pointing**: It was difficult for me to point to the first characters of words and to suggestions.
7. **Overall usability**: Overall, it was convenient for me to enter text using this input method.

Note that all even-numbered statements are positive, while all odd-numbered statements are negative. This design choice was intended to encourage participants to reflect more carefully on their responses. Each statement was rated on a 7-point Likert scale.