Seeing and Touching the Air: Unraveling Eye-Hand Coordination in Mid-Air Gesture Typing for Mixed Reality

Jinghui Hu
Department of Engineering
University of Cambridge
Cambridge, United Kingdom
jh2265@cam.ac.uk

John J Dudley
Department of Engineering
University of Cambridge
Cambridge, United Kingdom
jjd50@cam.ac.uk

Per Ola Kristensson
Department of Engineering
University of Cambridge
Cambridge, United Kingdom
pok21@cam.ac.uk

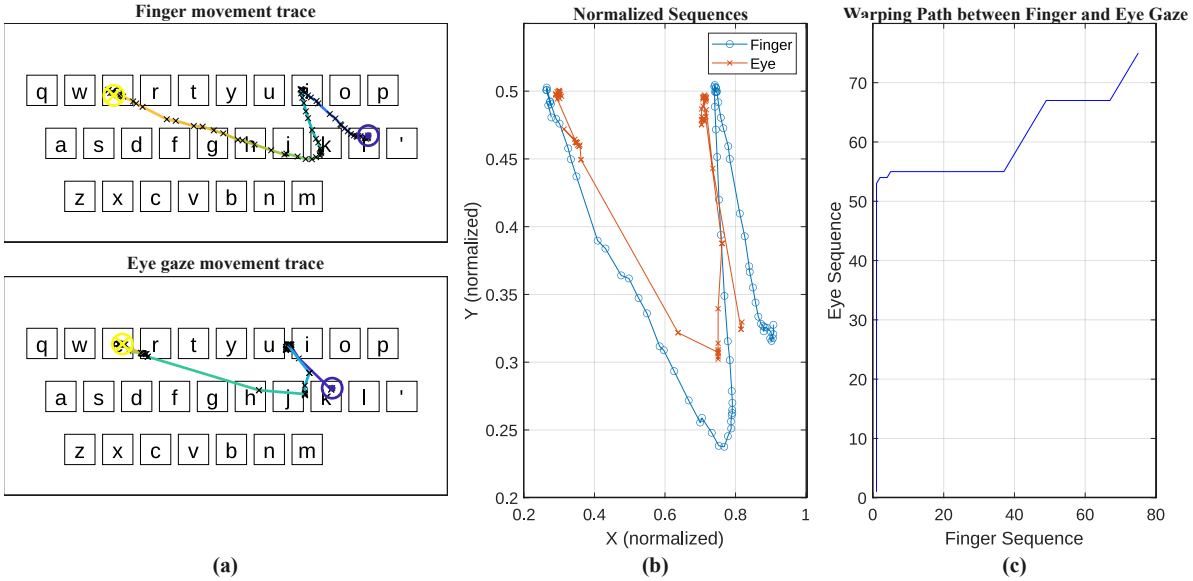

Figure 1: Comparison of eye gaze and finger movement traces during mid-air gesture typing of the word "like." (a) Top: An illustration of the finger movement trace, where × marks sampled points along the gesture, with a color gradient indicating the progression of time. (a) Bottom: A demonstration of the corresponding eye gaze movement trace, where the color gradient demonstrates how the eye gaze anticipates the finger movement. (b) The spatial alignment between eye and finger movements. (c) The dynamic time warping (DTW) path, highlighting the temporal alignment and lead-lag behavior between the eye and finger movements, where a "step" represents a single point in the time-series data. The finger lags the eye by approximately 50 steps, indicating a delay of 50 time-sampled points.

Abstract

Mid-air text entry in mixed reality (MR) headsets has shown promise but remains less efficient than traditional input methods. While research has focused on improving typing performance, the mechanics of mid-air gesture typing, especially eye-hand coordination, are less understood. This paper investigates visuomotor coordination of mid-air gesture keyboards through a user study (n = 16) comparing gesture typing on a tablet and in mid-air. Through an expert task we demonstrate that users were able to achieve a comparable text input performance. Our in-depth analysis of eye-hand coordination reveals significant differences in the eye-hand coordination patterns between gesture typing on a tablet and in-air. The mid-air gesture typing necessitates almost all of the visual attention on the keyboard area and a more consistent synchronization in eye-hand coordination to compensate for the increased motor and cognitive demands without physical boundaries. These insights provide important implications for the design of more efficient text input methods.

CCS Concepts

• **Human-centered computing** → **Mixed / augmented reality**; **Text input**; **Empirical studies in HCI**.

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713743

ACM Reference Format:
Jinghui Hu, John J Dudley, and Per Ola Kristensson. 2025. Seeing and Touching the Air: Unraveling Eye-Hand Coordination in Mid-Air Gesture Typing

for Mixed Reality. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3706598.3713743

1 Introduction

This paper presents new data on how users implicitly coordinate their eye and hand movements during mid-air gesture typing in mixed reality (MR). Most prior understanding of typing behavior stems from research on physical keyboards [9] and touch-based systems [15], which differ significantly from the input demands of MR environments. Unlike physical and touchscreen keyboards, mid-air gesture typing lacks tactile feedback, requiring users to rely on visual cues and auditory feedback to guide their fingers across a virtual keyboard. Despite the increasing use of MR across various applications, the specific visuomotor strategies employed during mid-air typing remain underexplored.

Previous studies [9, 17] have shown that typing performance on physical keyboards is largely driven by muscle memory and tactile feedback, which allow users to minimize their reliance on visual attention. However, in mid-air gesture typing, where physical touch is absent, users must continuously allocate visual attention between their fingers and the virtual keyboard, likely increasing cognitive load and reducing performance. There has been limited exploration into the unique challenges that mid-air typing presents for visuomotor coordination.

We investigate these issues by comparing mid-air gesture typing to gesture typing on a tablet, focusing on the coordination between eye and finger movements. Using data collected from the HoloLens 2, we analyze how users allocate visual attention, how finger movements align with gaze, and how typing performance evolves with experience.

Our key research questions are:

RQ1. What are the eye-finger coordination patterns while typing in an MR headset?
RQ2. How synchronized are the movements of the eyes and hands during typing tasks?
RQ3. What are the distinct behaviors of the eyes and fingers separately while typing?
RQ4. How do these eye and hand behaviors relate to typing performance?
RQ5. Do users exhibit common or different eye-finger coordination patterns after practice?
RQ6. How might these findings inform further designs of eye-gaze assisted typing?

The findings of this study highlight the distinct coordination challenges in mid-air typing, such as a greater reliance on visual attention and a more pronounced and consistent lag in finger movements trailing eye movements, as shown in Fig.1. These insights inform actionable design strategies for improving MR text entry interfaces, including predictive models and adaptive feedback mechanisms that reduce the cognitive and motor demands of mid-air input.

In summary, this paper makes the following contributions:

- We provide an in-depth empirical investigation of eye-hand coordination during mid-air gesture typing in mixed reality and understanding of how the absence of tactile feedback in MR impacts typing performance.
- We collect a rich dataset of visuomotor movements during typing in both mid-air and tablet conditions and identify that the mid-air condition exhibits a tighter and more consistent eye-finger coordination characterized by a lagged synchronization.

2 Related Work

2.1 Typing in MR Headsets

Text entry in mixed reality (MR) is an evolving research area, with various input methods being explored to address the specific challenges of AR/VR environments. While physical keyboards can achieve entry rates of over 40 words per minute (WPM) in VR [5, 10, 43], they are impractical for immersive or mobile scenarios, limiting their applicability in MR. Alternative modalities, such as voice-to-text, have been studied but are constrained by ambient noise interference and privacy concerns [1]. Other methods include eye typing[13, 14], head gaze input [46], and the use of controllers [4, 20, 41, 45] are constrained with relatively low entry rates of up to 20 WPM. These challenges have driven increased interest in more adaptable methods like mid-air gesture typing, particularly for MR applications.

Gesture typing, initially popularized by word-gesture keyboards on tablets [21–23, 47], has been adapted for MR environments [8, 38]. The success of gesture typing on touchscreens stems from its intuitive nature and high typing speed, providing a low learning curve for users [23, 48]. In MR settings, gesture typing has been implemented on platforms like the Microsoft HoloLens 2 and Apple Vision Pro, demonstrating its adaptability. Previous studies have explored gesture typing with various interaction techniques in AR/VR. Vulture [32] employed precise hand tracking for mid-air gesture typing on large wall displays, achieving an average rate of 28 WPM. Yu et al. [46] utilized head orientation tracking on an AR headset, enabling gesture typing with projected cursors, reaching an average of 24.7 WPM. Xu et al. [44] compared different projection methods—head, hand, and controller—for gesture typing, finding that controller-based projections resulted in the highest rate of 13.7 WPM. Dudley et al. [8] assessed both direct and indirect mid-air typing, showing that direct gesture typing on a virtual QWERTY keyboard achieved entry rates of 20–30 WPM, with peak rates of 40–45 WPM. This emphasizes the potential for transferring existing typing skills to MR environments by simulating familiar touchscreen typing experiences.

Despite these developments, mid-air gesture typing in MR typically achieves entry rates ranging from 20 to 40 WPM for expert users [4, 32, 45, 46], which remain lower than the 40-50 WPM commonly observed among expert users on smartphones [26]. This discrepancy highlights a challenge in seamlessly transferring typing skills from traditional devices to MR environments.

2.2 Eye-hand Coordination

Eye-hand coordination is a natural and essential part of everyday interactions [18, 25]. Generally, Both the eyes and hands move toward the target at the same time, with the eyes arriving earlier due to faster saccades. This coordination adapts to digital environments

as well, where the eyes focus on or near the interaction site for around 500 ms, similar to real-world interactions with physical objects [3].

When typing on physical keyboard, eye-hand coordination is efficient as the tactile feedback from the physical keyboard permits attending to the keyboard area [9]. Thus, the users focus on the text-entry area, allowing them to quickly detect and correct errors [9, 17]. Experienced typists rely on muscle memory, rarely needing to look at the keyboard [9, 34].

Touchscreen keyboards lack tactile feedback compared to physical keyboards, requiring users to rely heavily on visual attention to guide their fingers, which can slow typing speed [11]. Frequent glances at the text-entry area aid in proofreading but may come at the cost of reduced finger accuracy [33]. Detecting errors late makes them harder to correct, causing users to slow down in an effort to balance speed and accuracy [19, 33]. As a result, many users rely on features like autocorrect to minimize mistakes [33]. Jiang et al. further demonstrated that eye-hand coordination is strongly influenced by the competition for visual attention in the keyboard area [15]. Similarly, Shi et al. used computational models to simulate real-world typing strategies, showing how visual attention and finger movements are managed based on working memory constraints [39]. While these studies have focused on 'tap' touch typing, especially around error correction and inter-word behavior, gesture typing has not been thoroughly investigated. Our research aims to explore eye-hand coordination during gesture typing, with an emphasis on intra-word dynamics.

In an MR environment, the absence of tactile feedback is even more pronounced. No tactile feedback is provided when typing on mid-air keyboards, which are the default text entry method in major MR headsets like the Apple Vision Pro and HoloLens. Most prior research in MR has focused on explicitly combining eye gaze and hand movements as multimodal interaction techniques [6, 12, 28, 29, 35, 40, 49]. However, how the eye and hand naturally coordinate in an MR environment, especially with the challenges of mid-air typing, remains critical yet underexplored.

2.3 Eye Assisted Typing

As users rely heavily on their visual attention to guide finger movements as discussed in the previous section, various techniques have been proposed to leverage eye gaze for improving typing speed and accuracy. While many studies have focused on how gaze can assist hand movements, the complexity of gaze behavior during typing in MR remains less explored.

Lystbaek et al. [30] explored how gaze could assist freehand gestural text entry. Their study focused on coordinating eye and hand movements across the spatial positions of keys. They introduced a novel gaze-assisted text entry method where users align both gaze and manual pointer at each key, offering an alternative to traditional dwell-time or manual trigger mechanisms.

Gaze Speedup [50] examined an indirect gesture typing method in VR by accelerating the wrist cursor towards the gaze fixation point while keeping the speed perpendicular to this direction constant. A Gaussian smoothing function was applied to gradually decrease the speedup rate as the cursor approached the gaze fixation to minimize overshooting. The study compared three methods: wrist-only, Speedup, and Gaussian-speedup. Both the Speedup and Gaussian-speedup methods reduced hand movement, though no significant differences were found in input speed.

Ren et al. [36] developed a Bayesian gaze model to predict the next intended key on a mid-air virtual keyboard, enlarging the predicted key to improve the typing experience. Their model was trained using gaze data collected from tap typing on the mid-air keyboard. The Eye-Hand method achieved a typing speed of 12.95 WPM, compared to 11.31 WPM for tap typing and 9.28 WPM for a hand-only method. Although they reported significant differences between the methods, there was no pairwise evidence that the Eye-Hand method was faster than the hand-only approach, with only a marginal improvement of 1.64 WPM on average.

Hu et al. [14] introduced SkiMR, which relies solely on eye gaze input for mixed reality headsets, and allows hands-free text entry by decoding gaze trajectories into words. Unlike previous approaches that rely on dwell time for key selection, SkiMR eliminates explicit dwell delays, enabling faster text input. Their study demonstrated that users could achieve an average typing speed of 12 WPM. As a later extension, Hu et al. [13] adapted this typing method for command search in MR.

All these prior works assume that eye gaze either anticipates or remains directly aligned with hand movements throughout typing. However, gaze behavior during typing is more complex. In this study, we investigated the coordination between eye and finger movements during mid-air gesture typing on a HoloLens 2. Our findings challenge the assumption made in these previous studies.

3 Study Method

This study aims to investigate eye-hand coordination patterns during mid-air gesture typing compared to touchscreen typing on a tablet. Understanding these patterns is crucial for improving the usability and efficiency of text input in mixed reality environments, where the absence of tactile feedback presents unique challenges. To this end, we used a within-subjects design with two conditions: Tablet (baseline) and Mid-air, where participants performed typing tasks in both settings. The two conditions allowed us to analyze how the lack of tactile feedback and reliance on visual cues in mid-air typing affect typing performance and coordination.

3.1 Metrics

The following metrics were chosen to assess typing performance, finger movements, eye movements, and eye-finger coordination. Many of these metrics were distilled from prior research on gesture typing [26], touchscreen typing [15], handwriting [24], and the authors' own experience. These metrics are designed to provide comprehensive insights into the unique challenges of mid-air gesture typing, particularly in the absence of tactile feedback, and how it compares to more traditional input methods, such as touchscreen typing.

- Performance
  - Entry Rate (WPM): The typing speed is measured in words per minute (WPM).
  - Error Rate (CER): The Character Error Rate (CER) represents the proportion of incorrect characters entered by the participant during the typing task.

  - Deletions: This metric quantifies the number of times participants used the delete function to correct mistakes, serving as a measure of error correction behavior.
  - Inter Word Interval (IWI) (ms): The average time interval (in milliseconds) between the completion of one word and the initiation of the next.
- Finger Movements
  - Swipe Length (pixels): The total distance covered by the participant's finger during the gesture typing, measured in pixels.
  - Swipe Time (ms): The time taken to complete each word gesture.
  - Z-axis Movement (m): The Z-axis movement length quantifies how far the finger moves towards or away from the keyboard plane during each word gesture. It is calculated by first transforming the finger tip positions into the virtual keyboard plane. The movement length for a given word is then the sum of the absolute differences in the depth coordinate (along the Z-axis) between each timestep. This metric is not applicable for the Tablet condition.
- Eye Movements
  - Gaze Shifts: The average number of eye shifts from key area to the text or function area. This metric was adopted from a mobile typing study [16] investigating eye-hand coordination.
  - On Keyboard Ratio: The proportion of time spent looking at the key area of the keyboard, calculated by dividing the time spent gazing at the key area by the total duration of the trial.
  - Fixation Numbers: The number of times the user's gaze fixates on a specific point. Fixations were detected using a dispersion-based algorithm [37] tailored to account for the inherent instability in the HoloLens 2's eye tracking. A fixation was defined as a sequence of gaze points where the spatial dispersion in both x and y coordinates remained within 1.5° of visual angle, aligning with the HoloLens 2's reported eye tracking accuracy as documented by Microsoft[1] and prior research [2]. Additionally, the sequence duration had to be longer than a minimum threshold of 100 ms, ensuring that transient gaze points caused by noise or brief saccades were excluded. This method provided reliable fixation detection while compensating for potential variability in the underlying eye tracking data.
  - Fixation Durations (ms): The average time a user's gaze remains fixed on a specific point. Fixation durations were calculated as the time difference between the first and last gaze points in a sequence that met the dispersion and duration criteria, ensuring accurate measurement of sustained gaze behaviors.
  - Saccade Lengths (pixels): The average distance between consecutive fixations.
- Eye-finger Coordination
  - Distance (pixels): The average Euclidean distance between eye gaze position and finger position on the keyboard.
  - Dissimilarity (pixels): We used the Dynamic Time Warping (DTW) distance as a measure of the dissimilarity between finger and eye movement patterns. This metric quantifies the temporal and spatial discrepancy between finger and eye movements. Higher values indicate less coordination between the two modalities.
  - Mean Warping Path Length (WPL): This metric represents the average number of alignment steps (or warping steps) in DTW needed to match the elements of finger positions and eye gaze positions. Each step corresponds to a matched point between the two sequences, with longer paths indicating more temporal variation between the movements. We normalized the WPL by the length of the finger/eye sequences.
  - Signed Deviations: The average signed difference between finger and eye movement timings. Negative values suggest that the eye is trailing the finger, while positive values indicate the opposite.

[1]https://learn.microsoft.com/en-us/windows/mixed-reality/design/eye-tracking

3.2 Apparatus and Implementation

We developed two gesture typing applications with identical physical dimensions (18 × 28 cm), as shown in Fig. 2. Both applications were developed in Unity with one deployed on a Dell XPS 9315 tablet and the other on the HoloLens 2. The physical dimensions of the keyboard were determined based on the largest available space on the tablet. Both the tablet and HoloLens 2 operated at a 60 Hz refresh rate, ensuring smooth interaction. Touch down/up events on the tablet and lift-on/off events on the virtual keyboard were triggered immediately, with imperceptible latency.

In the tablet typing system, touch events on the tablet screen (touch-down and touch-up) defined the input delimiters. We provided visual feedback using a fingertip cursor on the tablet when the finger touched down on the screen. The cursor followed the touch position while the user gestured on the keyboard and disappeared on touch-up events.

In the mid-air typing system, lift-on/off events were determined by the fingertip entering or departing from a fixed threshold distance from the virtual keyboard plane. We provided visual feedback through a fingertip cursor, which changed color and was paired with auditory feedback to signal the finger lift-on and lift-off actions. To minimize unintended exits from the virtual keyboard, we used a 1 cm threshold above the keyboard plane and a 2 cm threshold below the keyboard plane. These thresholds were refined through a pilot study involving three participants who followed the same procedure as the main study. During the pilot, we explored various depth threshold settings. Smaller thresholds often led to accidental lift-offs when gesture typing on the keyboard. Larger thresholds reduced accidental lift-offs but introduced a head and tail portion of the trace when the finger entered or exited from the virtual keyboard plane. The trace with head and tail artifacts can then induce errors when decoding the intended word. Ultimately, all participants in the pilot preferred the larger threshold below the keyboard, leading to the adopted configuration.

To support the investigation of eye behaviors during gesture typing, we restricted visual feedback to a cursor and intentionally did not including any visualization of the recent gesture trace path.

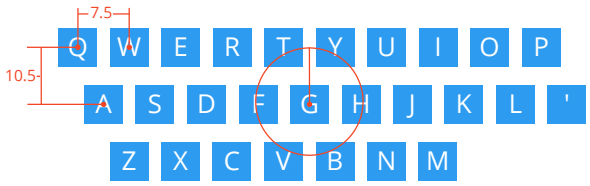

Figure 2: Layout of the virtual keyboard used in the study, showing the height difference of 10.5 units between key centers and a width difference of 7.5 units between adjacent keys. The red circle highlights the spatial range used for calculating the number of fixations later in the height range in Fig. 6.

This design decision was made in order to limit visual distractions and ensure that the primary focus remained on understanding eye-hand coordination during gesture typing.

3.2.1 Gesture Decoder. To interpret input traces in both systems, we integrated a probabilistic word gesture decoder. This model evaluated the statistical likelihood of letter sequences and the spatial proximity of keys, effectively mitigating common gesture input errors such as deletions, insertions, and substitutions. Notably, the decoder could be adapted to both the mid-air virtual keyboard, using finger positions from the integrated hand tracking, as well as the touch positions on the tablet, ensuring seamless compatibility and consistency across the two systems.

We selected the decoder based on its demonstrated effectiveness in previous research [8], where it enabled high text entry rates on a mid-air QWERTY keyboard in VR with integrated hand tracking. The decision to adopt this specific decoder was motivated by two critical factors. First, it provides strong robustness to imprecise articulation and supports interaction at different scales. This capability is essential for managing the noisy, variable nature of mid-air gestures and allows seamless adaptation to resized keyboards. Second, the decoder operates entirely on-device, eliminating the need to relay data to a separate computer. This contrasts with approaches that relied on off-device processing [7, 20, 31], potentially creating latency and limiting practical deployment.

By running on a processor-constrained device like the HoloLens 2, the decoder provides a realistic assessment of the performance potential of hand-based text input under current hardware constraints. These qualities make it well-suited to the goals of this study, ensuring both systems—mid-air and tablet typing—deliver robust and efficient performance while reflecting practical, real-world usability.

3.2.2 Calibration and Logging. Apart from the keyboard system, the experiment control system was implemented for coordination calibrations and data collection. Participants were required to wear HoloLens 2 in both conditions for capturing eye and hand tracking data. Vuforia[2]'s image target tracking was utilized to align keyboard positions in both conditions and to locate the tablet keyboard during the Tablet condition. The virtual keyboard was only displayed in the Mid-air condition.

[2]https://developer.vuforia.com/library/objects/image-targets

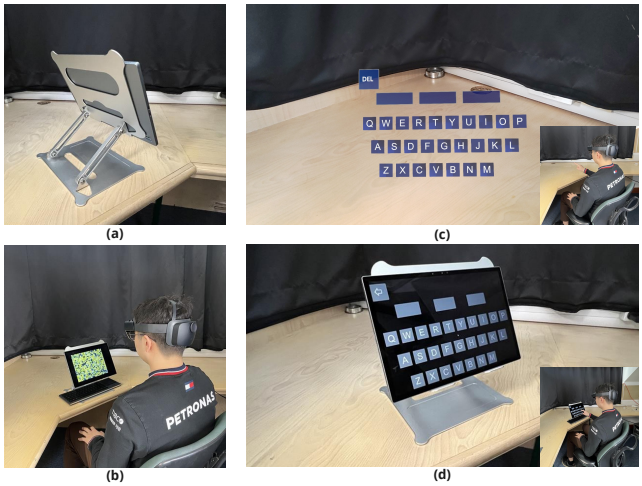

Figure 3: The study setting: (a) Shows the computer stand used for holding the tablet in the study. (b) Shows the keyboard position calibration before each condition. (c) Shows virtual keyboard from the participant's view in Mid-air condition. (d) Shows the tablet keyboard during the study.

After the calibration, the HoloLens 2 tracks both hand and eye movements for both conditions. Additionally, hand joint and head gaze data were recorded, enabling the possibility of future simulations or training applications. The HoloLens 2's hand-tracking system and Extended Eye Tracking API (90 Hz refresh rate) were used to capture finger and eye positions, which were mapped to the keyboard coordinates.

In the Tablet condition, interaction events (e.g., start and stop of each gesture) from the tablet were also logged. Post-processing was performed to synchronize the hand and gaze data between the tablet and Hololens 2 with the interaction events and timestamps. All data from both systems were sampled at the same sample rate, and both systems operated at the same frame rate.

A physical Bluetooth keyboard was employed solely for controlling the study flow. Participants used the SPACE key to display the keyboard and RETURN to submit phrases using the physical keyboard, ensuring smooth transitions during the study.

3.3 Procedure

The conditions were presented to participants in a counterbalanced order. Prior to the study, participants completed a demographic and experience questionnaire. Participants indicated their dominant hand and responded yes or no to the following questions: (i) I have prior experience with head-mounted virtual reality or augmented reality; and (ii) I have prior experience with HoloLens or HoloLens 2. A brief 5-minute training session was conducted, during which participants were familiarized with the HoloLens 2 and gesture typing methods. This was followed by the completion of the HoloLens' in-built eye tracking calibration process.

Participants wore the HoloLens 2 for the entire study to allow for continuous tracking. The study was conducted in a quiet lab environment with participants seated at a desk, as shown in Fig.3.

Regardless of condition order, participants were first instructed to adjust the tablet's position and orientation for comfort. The tablet was held firmly in the set position using an adjustable laptop stand, as shown in Fig.3(a)(d). A calibration image was displayed on the tablet to permit alignment of the virtual keyboard in the HoloLens coordinate system (see Fig.3(b)). The tablet is simply removed for the Mid-Air condition (see Fig.3(c)) and, if the Tablet condition occurred second, the tablet was reinserted and aligned with the virtual keyboard displayed in the HoloLens 2. This procedure allowed for consistent placement and orientation of the keyboard across both conditions.

Each condition started with participants typing five practice phrases to familiarize themselves with the typing modality. Task 1 was split into five blocks, with five phrases per block and a one-minute break between blocks. Task 2 assessed the participants' performance after they had gained proficiency in each condition, involving the repeated typing of the same phrases for two blocks. In total, 25 phrases and 10 phrases were typed in Task 1 and Task 2 respectively. After completing Task 2, participants filled out the raw NASA-TLX questionnaire, completing only the ratings portion. Participants were asked to reflect on their experience of the condition across both tasks (detailed in the following subsection) when filling out the NASA-TLX questionnaire. This approach was intended to provide a holistic evaluation of the perceived workload associated with each typing modality rather than isolating the assessment to a particular task.

3.4 Tasks

We used a transcription task in both conditions, where participants were shown short stimulus phrases and asked to transcribe them as quickly and accurately as possible using the specified typing method. A physical SPACE key was used to activate the keyboard after memorizing the phrase, and a deletion key was provided above the keyboard for error correction.

Phrases were selected from the Enron mobile message dataset's memorable phrases subset [42], filtered to include only phrases with 45 or fewer characters, consisting of four or more words made up of letters (A–Z) and apostrophes. Punctuation was removed, and phrases were presented in lowercase. Phrases containing words not in the decoder's 64,000-word vocabulary were also excluded. Different sets of phrases were used for each condition, but the sets remained consistent across participants.

3.4.1 Task 1: Random Phrases. In Task 1, participants transcribed a total of 30 phrases in each condition, divided into five blocks. To ensure uniform difficulty across blocks, the complexity of the phrases in each block was balanced based on the number of characters and words.

3.4.2 Task 2: Repeated Phrases. Task 2 aimed to measure participants' proficiency in each condition by having them repeatedly transcribe a single phrase ('money wise that is') for 10 times. This task was designed to assess performance after participants had overcome the initial learning curve associated with Mid-air and tablet gesture typing.

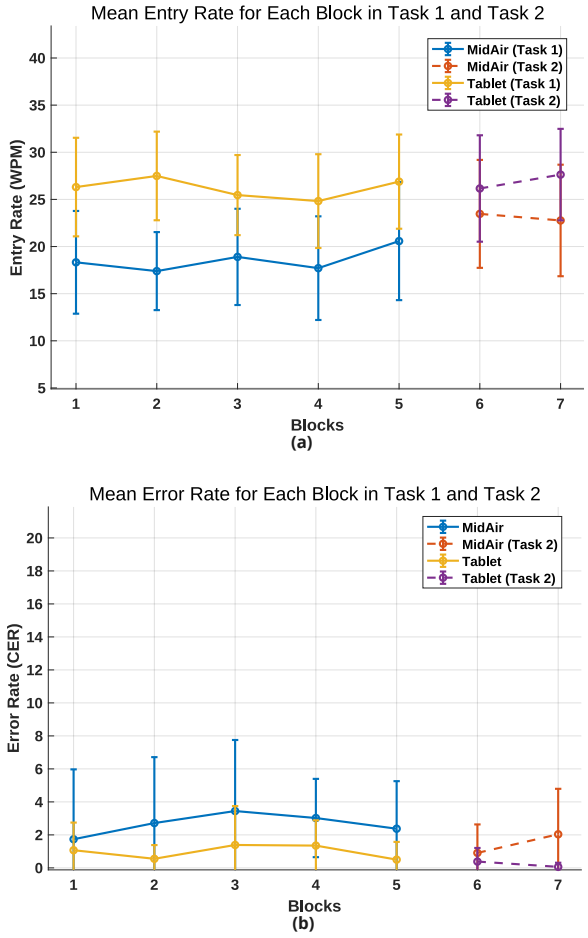

Figure 4: Line graphs showing the performance results of entry rates and error rates across tasks and blocks for Mid-Air and Tablet conditions. All entry rates are calculated in words-per-minute (WPM), and error rates are shown in character error rate (CER). Error bars represent the standard error of the mean (SEM). (a) Mean entry rates (WPM) for Task 1 (Blocks 1–5) and Task 2 (Blocks 6–7) for both conditions. (b) Mean error rates (CER) for Task 1 (Blocks 1–5) and Task 2 (Blocks 6–7) for both conditions.

3.5 Participants

16 participants (6 females, 10 males) aged 20 to 32 completed the study and provide the basis for the results presented in the following section. Among them, 10 had previous VR experience, and 4 were familiar with the HoloLens 2. All participants were native or fluent English speakers. All participants were right-handed and performed the tasks with their dominant hand. Each participant spent approximately one hour typing phrases across the two conditions and was compensated with a voucher for their participation.

Table 1: Metrics Comparison: Mid-air vs. Tablet and Task 1 vs. Task 2. Full definitions and corresponding units in Section 3.1.

	Task 1 (Novice)			Task 2 (Expert)			Mid-air	Tablet
Performance	**Mid-air**	**Tablet**	$p.$	**Mid-air**	**Tablet**	$p.$	$p.$	$p.$
Entry Rate (WPM)	18.58(4.42)	26.19(3.73)	**0.005****	23.11(5.59)	26.89(4.73)	0.10	0.02	0.33
Error Rate (CER)	2.66(2.10)	0.97(0.76)	0.14	1.39(1.61)	0.22(0.49)	0.18	0.36	0.13
Deletions	1.50(0.74)	0.72(0.54)	**0.002***	0.83(0.84)	0.48(0.54)	2.81	0.24	0.46
IWI	2221(685)	1487(534)	**0.009***	1261(583)	852(297)	0.12	**0.003***	**0.002***
Finger Movements								
Swipe Length	89(4)	81(11)	0.36	102(6)	90(13)	0.038	**0.005***	0.02
Swipe Time	1549(320)	902(138)	**0.005****	1457(343)	861(129)	**0.005***	1.17	0.59
Z-Axis Movement	0.136(0.02)	-	-	0.142(0.026)	-	-	0.41	-
Eye Movements								
Gaze Shifts	3.23(1.24)	2.97(1.35)	0.37	1.95(1.08)	2.09(1.19)	0.91	**0.005***	0.053
On Keyboard Ratio	0.95(0.03)	0.79(0.15)	**0.007***	0.95(0.025)	0.82(0.13)	**0.005***	0.41	0.45
Fixation Numbers	23.57(4.07)	17.06(2.80)	**0.005****	16.21(3.45)	12.09(2.24)	**0.009***	**0.005***	**0.005***
Fixation Durations	353(40)	276(50)	**0.019***	340(46)	260(26)	**0.005****	0.12	0.36
Saccade Lengths	17.86(2.00)	17.39(3.17)	0.68	20.22(2.71)	19.24(2.85)	0.42	0.02	0.67
Eye-finger Coordination								
Distance	15.06(2.51)	16.49(3.043)	0.16	14.05(3.57)	15.03(3.00)	0.41	0.35	0.18
Dissimilarity	9.34(1.64)	13.14(3.63)	**0*****	8.29(2.02)	11.78(3.2)	**0*****	0.12	0.27
Mean WPL	1.022(0.004)	1.034(0.006)	**0*****	1.020(0.006)	1.032(0.007)	**0*****	0.42	0.50
Signed Deviations	-18.88(4.17)	-9.43(3.71)	**0*****	-15.81(5.35)	-8.50(4.10)	**0.003***	0.08	0.51

4 Results

4.1 Performance

Our results reveal clear differences in performance between mid-air and tablet typing, particularly in the novice condition (Task 1). Participants typed significantly slower in Mid-air (M = 18.58, SD = 4.42) compared to the Tablet (M = 26.19, SD = 3.73), $p < 0.05$. This finding highlights the steeper learning curve associated with Mid-air typing, where the absence of tactile feedback and physical stability makes it difficult for users to achieve the same level of proficiency as on the tablet. Despite the increased difficulty, error rates (CER) were not significantly different between conditions. However, the backspace key was used significantly more frequently ($p < 0.05$) by participants in the Mid-air condition (M = 1.50, SD = 0.74) than on the Tablet (M = 0.72, SD = 0.54), indicating a greater need for correction.

Additionally, participants took significantly longer to resume typing between words in Mid-air, as shown by the higher inter-word interval (M = 2221 ms, SD = 685 ms) compared to the Tablet (M = 1487 ms, SD = 534 ms), $p < 0.05$. This reinforces the notion that Mid-air typing demands greater motor effort, likely due to the increased reliance on visual attention and proprioception.

Interestingly, in Task 2 (expert condition), no significant differences in text entry rates or error rates were observed between Mid-air (M = 23.11, SD = 5.59) and Tablet typing (M = 26.89, SD = 4.73), $p = 0.18$. The inter-word interval also decreased in Mid-air typing (M = 1261 ms), suggesting that participants became more efficient with practice, though no significant differences were found compared to Tablet typing (M = 852, $p = 0.12$). This finding indicates that with sufficient experience, participants adapted to the Mid-air interface, achieving performance comparable to the Tablet condition. The decreased frequency of deletions from Task 1 to Task 2 in Mid-air (M = 1.50 to M = 0.83) likely resulted from increased familiarity which, in turn, improves accuracy and reduces the need for corrections over time.

4.2 Finger Movements

The analysis of finger movements during swiping gestures revealed different motor patterns between conditions. During Task 1, participants took significantly longer to perform swipe gestures in the Mid-air condition (M = 1549 ms, SD = 320 ms) compared to the Tablet condition (M = 902 ms, SD = 138 ms), $p < 0.05$, indicating that participants took more time to complete swipes when typing mid-air. However, there was no significant difference in Swipe Length between Mid-air (M = 89, SD = 4) and Tablet typing (M = 81, SD = 11), $p = 0.36$. This suggests that while participants maintained consistent gesture ranges across both conditions, they performed these gestures more cautiously in the Mid-air setting. The lack of physical boundaries in mid-air typing likely led participants to slow down and execute the gestures with greater control to ensure accuracy, compensating for the absence of tactile feedback provided by the tablet surface.

In Task 2, Swipe Time remained significantly longer in Mid-air (M = 1457 ms, SD = 343 ms) compared to Tablet typing (M = 861 ms, SD = 129 ms), $p < 0.05$, indicating that even with practice, mid-air gestures still take longer to execute. Interestingly, Swipe Length increased significantly in Mid-air typing during Task 2 (M = 102, SD = 6). This suggests that as participants became more familiar with the mid-air interface, they developed more exaggerated gestures,

perhaps as a result of growing confidence in their ability to interact with the system. With practice, participants may have adapted to the virtual keyboard's boundaries and began gesturing more freely, making use of the available space to enhance control and accuracy.

Z-axis movement did not show significant differences across tasks or conditions, indicating that the vertical component of the swipes remained consistent in Mid-air typing across both tasks, regardless of experience. As the onset and offset of the swiping gesture were controlled by a depth threshold in the system, it is likely that participants adapted their motor control strategies to operate within this constraint. Most of the motor control adjustments in Mid-air typing may have been focused on limiting finger movements in the depth direction to avoid crossing the threshold unintentionally, thereby ensuring more accurate gesture input.

4.3 Eye Movements

Our eye movement analysis demonstrates the additional visual demands associated with Mid-air typing. In Task 1, participants made significantly more fixations in the Mid-air condition (M = 23.57, SD = 4.07) than in the Tablet condition (M = 17.06, SD = 2.80), $p < 0.005$, and spent more time fixating on the virtual keyboard (On Keyboard Ratio: M = 0.95, SD = 0.03) compared to the Tablet (M = 0.79, SD = 0.15), $p < 0.01$. This suggests that the lack of tactile feedback in Mid-air typing forces users to rely heavily on visual feedback to guide their input, making the process more visually demanding. Additionally, Fixation Durations were significantly longer in Mid-air typing (M = 353 ms, SD = 40 ms) compared to the Tablet (M = 276 ms, SD = 50 ms), $p < 0.05$, further supporting the hypothesis that Mid-air typing requires sustained visual attention.

We also found a strong positive correlation between the number of fixations and the number of characters in Task 1 for both Mid-air ($r = 0.77$, $p < 0.0001$) and Tablet typing ($r = 0.81$, $p < 0.0001$), with no significant difference between the two conditions ($Z = 0.27$, $p = 0.7856$), as shown in Fig.5. This indicates that both input modalities require more visual attention as text complexity (number of characters) increases, but mid-air typing demands higher levels of visual engagement overall.

We further analyzed the spatial distributions of these fixations. As shown in Fig. 6, we calculated the number of fixations within each key that fell into the spatial range corresponding to the vertical offset between keyboard rows (highlighted by the red circle in the Fig.2). In our keyboard design, the vertical offset between rows (10.5 units) was greater than the horizontal offset between keys (7.5 units). We used the larger vertical offset as the radius of this spatial range in order to account for peripheral fixations that may not directly align with key centers. This approach aimed to accommodate the natural imprecision in gaze fixation points relative to key locations. Setting the vertical offset as the radius for the spatial range represents a larger threshold for fixation analysis. If the Tablet condition still shows significantly fewer fixations within this range, it suggests that participants allocated less visual attention to the critical area while typing on the tablet.

A significantly lower number of fixations were observed within this range for the Tablet condition compared to the Mid-air condition for both Task 1 ($p < 0.05$, Cohen's $d = 1.28$) and Task 2 ($p < 0.05$, Cohen's $d = 1.67$). This indicates that the participants directed their eye gaze significantly less frequently to the critical area during typing on tablet.

In Task 2, while the number of fixations reduced for both conditions, Mid-air typing still required more fixations (M = 16.21, SD

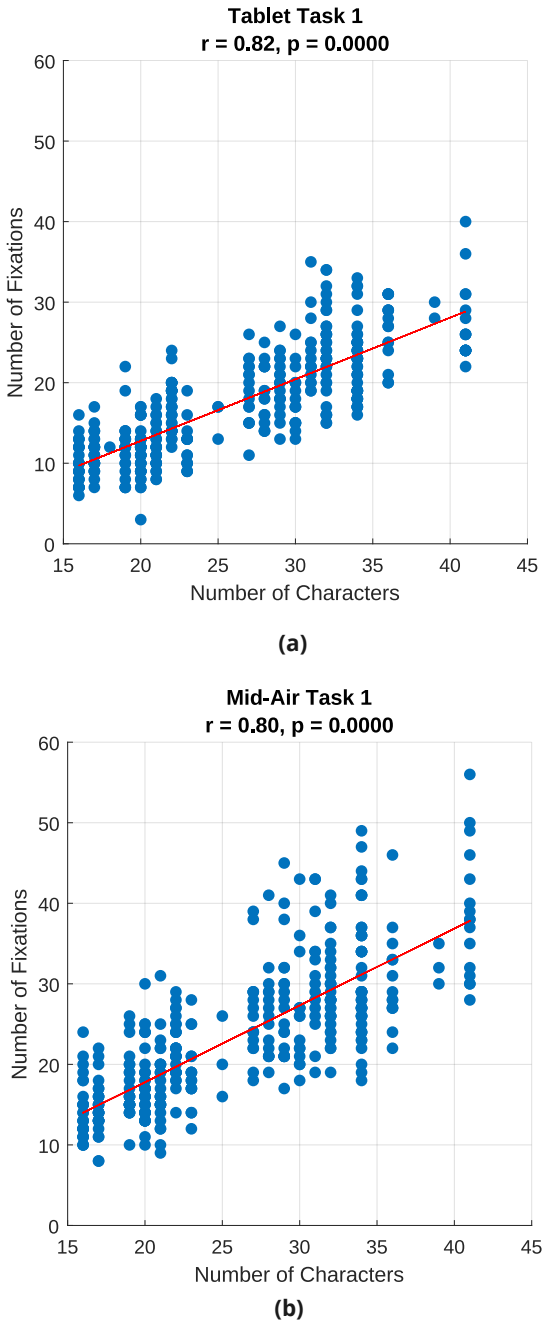

Figure 5: Scatter plots showing the correlation between the Number of Characters in a phrase (X-axis) and the Number of Fixations (Y-axis) for the Tablet and Mid-air conditions. The red lines represent the linear trendlines fitted to the data.

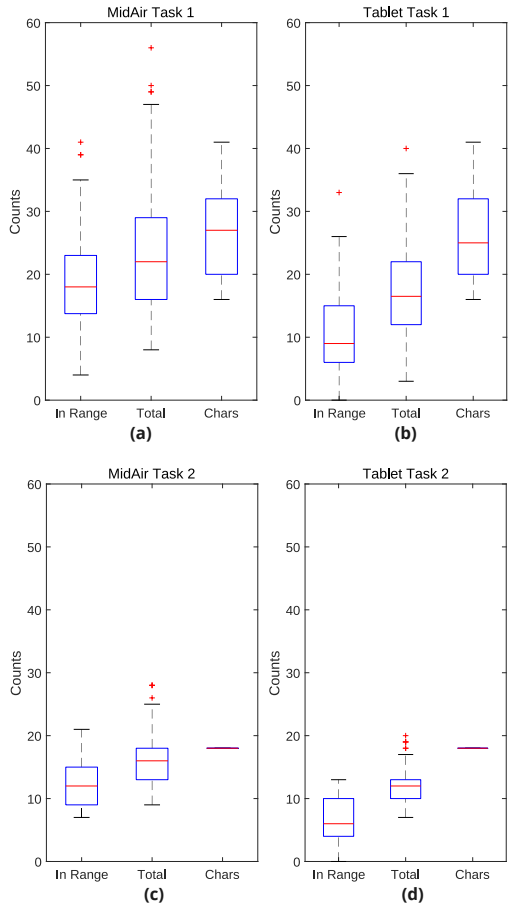

Figure 6: (a)-(d) Boxplots comparing the number of gaze fixation points within the height range of target letter keys (In Range), the total number of fixations per phrase (Total), and the number of characters per phrase (Char) across two conditions and two tasks.

= 3.45) than Tablet typing (M = 12.09, SD = 2.24), $p < 0.05$. The On Keyboard Ratio remained significantly higher in Mid-air (M = 0.95, SD = 0.025) compared to Tablet typing (M = 0.82, SD = 0.13), $p < 0.05$. These findings suggest that while participants become more visually efficient with practice, Mid-air typing still demands more visual attention than Tablet typing. Even as participants gained experience, mid-air typing continued to demand almost all of the visual attention as they spent 95% of the time spent executing a gesture looking at the keyboard.

These findings suggest that mid-air typing demands significantly more visual attention, as reflected by the higher number of fixations, longer fixation durations, and proportion of gaze on the keyboard, particularly in the novice condition. These trends persist, though somewhat reduced, as participants become more experienced in Task 2.

4.4 Eye-finger Coordination

4.4.1 Distance. We used the average Euclidean distance between eye gaze positions and finger positions to measure the spatial proximity. None of the comparisons show statistically significant differences after correcting for multiple comparisons. The results suggest that the two conditions of input and task type do not significantly impact the spatial relationship between eye and finger movements during gesture typing.

4.4.2 Dissimilarity. We used Dynamic Time Warping (DTW) distances to assess the spatial-temporal coordination between eye and finger movements. We normalize the DTW distances using min-max normalization with the keyboard bounds as the min and max values. Using the keyboard's dimensions keeps the spatial relationships intact while ensuring the data is uniformly scaled to [0, 1] and directly comparable across different trials.

Across both tasks, the Mid-air condition exhibited significantly lower dissimilarity between eye and finger movements, but no significant differences were observed between tasks within each modality. For Task 1 (random phrases), the mean DTW distance in the Mid-air condition was significantly lower at 9.34 (SD = 1.64) compared to the Tablet condition (M = 13.14, SD = 3.63; $p < 0.05$). Similarly, in Task 2, the mean DTW distance for Mid-air was 8.29 (SD = 2.02), while for Tablet it was 11.78 (SD = 3.20), again showing a significant difference between the two conditions ($p < 0.05$). These findings indicate that participants in the Mid-air condition exhibited tighter synchronization between eye and finger movements (lower DTW distance) compared to the Tablet condition across both tasks.

No significant differences in DTW distance were found between Task 1 and Task 2 for both the Mid-air ($p = 0.12$) the Tablet ($p = 0.27$) conditions. These results suggest that the type of task does not substantially affect eye-hand coordination within each modality, implying that participants' coordination performance was relatively stable and increased familiarity with phrases does not lead to notable changes in eye-hand coordination patterns within the same input modality.

4.4.3 Variance Analysis. A Levene's test was conducted to assess the equality of variances in DTW distance between the Mid-air and Tablet conditions. The test revealed the variability in DTW distance was significantly greater in the Tablet condition compared to the Mid-air condition. This suggests that participants' eye-hand coordination in the Tablet condition was more inconsistent, while the Mid-air condition exhibited more stable and consistent synchronization between eye and finger movements.

4.4.4 Word Perplexity. Word perplexity is a metric derived from language models that quantifies the uncertainty or predictability of a sequence of words. It is directly related to entropy, H, a fundamental concept in information theory that measures the average uncertainty inherent in a probability distribution. For a sequence of words, entropy is defined as:

$$H = -\sum_i P(w_i) \log P(w_i)$$

Here, $P(w_i)$ represents the probability of the i-th word in the sequence. Perplexity is the exponential transformation of entropy,

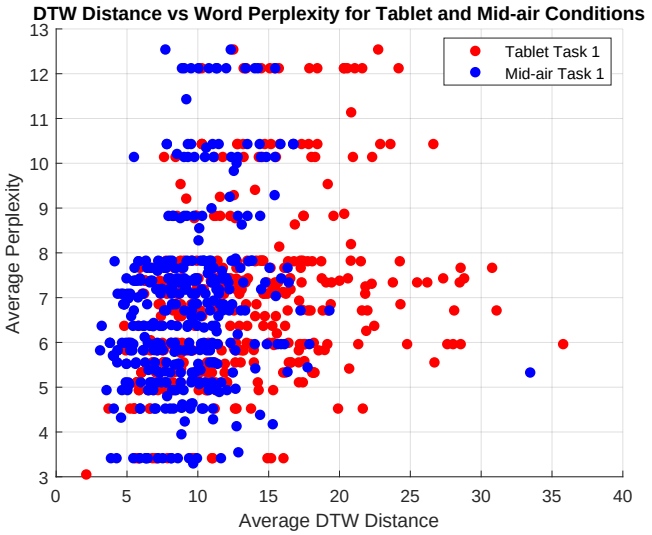

Figure 7: The scatter plot displays the relationship between Average DTW Distance (X-axis) and Word Perplexity (Y-axis) for the Tablet and Mid-air conditions. Two distinct colors represent each condition: blue for the Mid-air condition and red for the Tablet condition.

given by:

$$Perplexity = 2^H$$

This transformation translates the theoretical concept of entropy into a more interpretable scale, representing the effective size of the set of potential next words. Lower perplexity values correspond to sequences that are more predictable, while higher perplexity values indicate greater uncertainty and linguistic complexity. In this study, word perplexity serves as a proxy for word complexity, enabling us to investigate how linguistic uncertainty influences visuomotor coordination patterns during gesture typing under the two conditions.

To explore the relationship between DTW Distance and Word Perplexity, we conducted a Spearman correlation analysis for each condition, as the Shapiro-Wilk test confirmed that Perplexity data was not normally distributed. The analysis revealed a statistically significant positive correlation between DTW Distance and Perplexity ($r = 0.180$, $p < 0.01$) in the Tablet condition, as illustrated in Fig.7. This result suggests that as word complexity (perplexity) increases, DTW Distance tends to increase, indicating more pronounced eye-finger desynchronization. Notably, this relationship was not significant in the Mid-air condition, implying that mid-air gesture typing may provide more stable coordination patterns, even when word complexity increases.

4.4.5 Alignment Steps. The Mean Warping Path Length represents the average number of alignment steps, as shown in Fig.1(c), required to synchronize eye and finger movements over time in the DTW analysis. We normalized the warping path length by the finger and eye movement sequence length for direct comparisons. The mean alignment path was significantly shorter in the Mid-air condition (M = 1.022, SD = 0.004) compared to the Tablet condition (M = 1.034, SD = 0.006), $p < 0.05$. further supporting the tighter synchronization of eye and finger movements in mid-air typing.

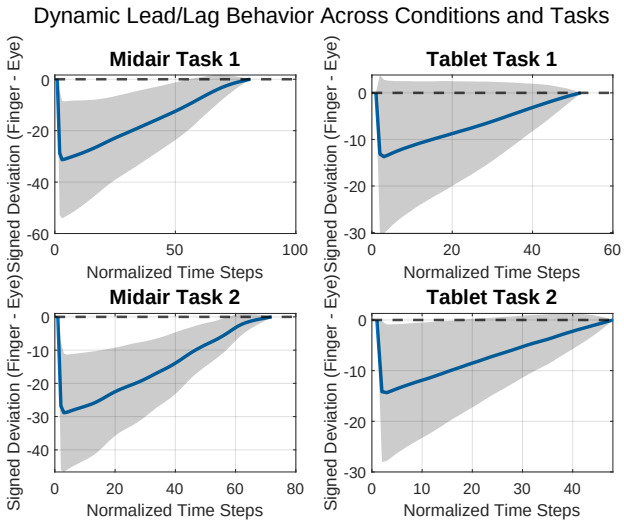

Figure 8: The signed deviations (Finger - Eye) over normalized time steps for both Mid-air and Tablet conditions across Task 1 and Task 2. Each subplot includes the mean deviation trajectory (blue line) with a shaded region representing variability across trials.

4.4.6 Lead/Lag Behavior—Temporal Alignment. The signed deviation refers to the difference between two time-aligned sequences, indicating the extent and direction of the deviation at each step of the alignment path. In both conditions, eye gaze led finger movements, with more pronounced time lag in the Mid-air condition (M = -15.809, SD = 5.353) compared to the Tablet condition (M = -8.502, SD = 4.096), $p < 0.05$. This lag remained consistent across Tasks (Fig. 8).

The larger magnitude of the mean signed deviation in the Mid-air condition, combined with the smaller DTW distance, suggests that this eye-finger lag is more systematic and consistent over time. As a result, it likely contributes less to the overall DTW distance, reflecting a stable but delayed coordination pattern in mid-air gesture typing.

4.4.7 Segmentation. To investigate spatial alignment further, we segmented swipe sequences into five equal stages. This approach seeks to show how spatial alignment and synchronization evolve throughout a swiping gesture. Each sequence was divided into five equally timed segments, ensuring that each stage represents a consistent portion of the gesture's duration. This method allows for a detailed analysis of eye-hand coordination at different stages. It helps identify patterns such as strong alignment early in the gesture and the deterioration of synchronization late in the gesture. Analyzing these segments provides deeper insights than aggregate metrics alone.

A three-way repeated measures ANOVA (between Segment, Condition and Task) revealed significant main effects for segment ($F(4, 304) = 8.33$, $p < 0.05$) and a significant condition-segment

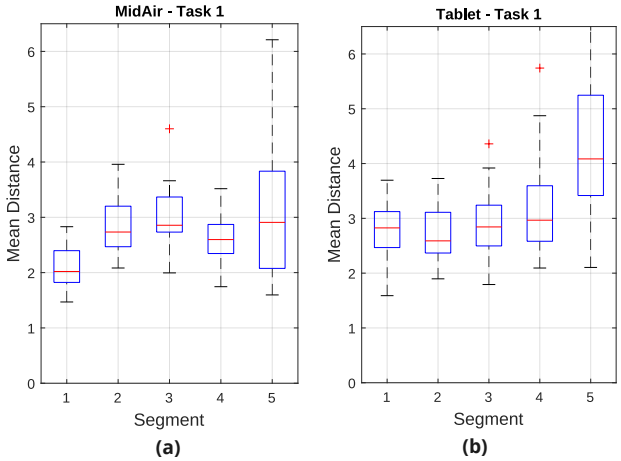

Figure 9: Box plots showing the distribution of mean eye-finger distance across five segments of the swiping sequence for (a) Mid-air gesture typing in Task 1 and (b) Tablet typing in Task 1.

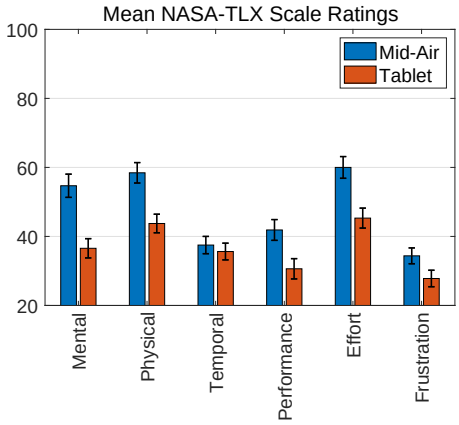

Figure 10: Mean NASA-TLX scale ratings. Note that a lower 'Performance' rating indicates 'better' perceived performance. The ratings for each item could range from 0 to 100.

interaction ($F(4, 304) = 3.38$, $p < 0.05$). As illustrated in Fig.9, post-hoc analyzes indicated that in Mid-air typing, Segment 1 exhibited significantly lower distances compared to Segment 3 ($p < 0.053$) and Segment 5 ($p < 0.05$). For Tablet typing, Segment 5 consistently demonstrated significantly higher distances compared to all other segments (all $p < 0.05$).

The observed trends suggest that eye-hand synchronization deteriorated toward the final stages of both conditions, with larger gaps emerging between eye and finger movements. This may be attributed to participants visually checking the text area to confirm the typing results toward the end of each gesture, leading to a brief desynchronization. The significantly lower eye-finger distance in the first segment of Mid-air typing suggests that the initial touch-on movement to reach the typing threshold of the virtual keyboard requires additional attention without haptic feedback. Interestingly, in the Mid-air condition, there was a notable fluctuation during the middle segments of the swiping sequence. We suspect that this is due to visual search for the next key while maintaining finger position stable until the next letter is located. Another possible explanation for the dynamic changes in eye-finger synchronization is that the finger moves faster between key targets and slows down when approaching the keys in Mid-air typing. This variable speed causes fluctuations in the eye-finger distance, with tighter alignment around key presses and more significant gaps during transitions between keys.

Although the overall DTW distances for each swipe in the Mid-air condition are lower in both mean and variance compared to the Tablet condition, the dynamic behavior during the swipes may differ significantly.

4.5 Perceived Workload

Participants completed the NASA Task Load Index (NASA-TLX) questionnaire at the end of each condition. Fig.10 shows the mean NASA-TLX ratings across all participants for both conditions. The Mid-air condition was considered by participants to be significantly more physically ($p = 0.002$) and mentally demanding ($p = 0.02$). The Mid-air condition also induced significantly more perceived effort ($p = 0.02$). Participants also perceived themselves as performing better and experiencing less frustration in the Tablet condition. However, the Friedman's test revealed neither of these measured differences were statistically significant.

5 Discussion

The key finding of our study is that mid-air gesture typing necessitates tighter and more consistent eye-hand coordination, characterized by lagged synchronization between gaze and finger movements (RQ1).

Interestingly, this lagged synchronization persists even as participants become more familiar with the interface. Unlike prior work on touchscreen typing [15] or physical keyboard use [9], where increased familiarity often leads to reduced visual reliance and more automatic motor responses, the current findings in mid-air typing suggest that the coordination pattern remains largely consistent across experience levels.

While the interaction becomes faster with practice, the foundational eye-finger coordination pattern appears to speed up in unison rather than demonstrating a clear shift towards reduced visual dependence. This speaks directly to whether users exhibit common or different eye-finger coordination patterns after practice: we find that while performance improves, the fundamental eye-leading-finger coordination pattern does not markedly change. Instead, the entire pattern accelerates uniformly, rather than shifting towards a more touch-typist-like mode with less visual dependency (RQ5). However, given the relatively limited familiarity achieved in this study, it's unclear how these patterns might evolve with extended practice.

5.1 Performance Gaps between Mid-air and Tablet Typing

Participants typed significantly slower in Mid-air compared to Tablet typing in Task 1. They also required significantly longer time to resume swiping and made more error corrections in Task 1. These performance gaps diminished after practice in Task 2, where Mid-air typing performance became comparable to Tablet typing in terms of speed and accuracy. Over time, they adjusted their strategies to balance the speed and accuracy trade-off while coping with the lack of tactile feedback, the need for greater visual reliance and gradually improving their efficiency. This improvement in performance despite persistent coordination patterns addresses how eye and hand behaviors relate to typing performance: as users become more adept, they type faster and more accurately, even though they still rely on visual guidance (RQ4).

A major challenge in mid-air typing is the absence of tactile feedback, which complicates depth regulation during finger movements. Dudley et al. [5] analyzed 3D finger motions on a virtual keyboard and highlighted that the lack of a physical surface in mid-air typing significantly hinders depth control. This difficulty leads to longer index finger travel times and higher error rates compared to keyboards aligned with a physical surface. Our findings support these observations, showing that participants in the *Mid-air* condition exhibited prolonged gesture durations and a greater reliance on visual feedback (RQ3). While our analysis primarily focused on eye and finger movements, prior work demonstrates that depth-related inefficiencies play a critical role in reducing typing performance and increasing cognitive load in 3D environments.

It is important to note that participants were also not fully accustomed to typing on the tablet, as this was not a typical touchscreen interface. However, we chose this particular keyboard size to ensure a fair comparison between conditions, leaving the absence of physical touch as the main distinguishing factor between Mid-air and Tablet typing. This allowed us to isolate the impact of touch on performance and coordination.

5.2 Visual Attention Demands in Mid-Air Typing

Mid-air typing required more visual attention, as reflected in a higher number of fixations, longer fixation durations compared to Tablet typing in both Tasks and almost all of the time focusing the visual attention on the keyboard. Even as participants gained experience (Task 2), Mid-air typing continued to demand more visual resources (RQ3). Several factors contribute to the high reliance on visual attention in Mid-air gesture typing.

First, participants must visually ensure their finger movements stay within the virtual keyboard's boundaries. Unlike physical or touchscreen keyboards, which provide tactile feedback, the absence of physical constraints in Mid-air typing requires constant visual guidance.

Second, since Mid-air interfaces lack tactile landmarks, visual feedback becomes critical for locating the next key—participants frequently engaged in visual search to maintain accuracy since they could not rely on touch to guide their finger movements. One particular difficulty was the sense of depth when interacting with virtual interfaces in mixed reality. The absence of physical boundaries not only complicated horizontal navigation across the keyboard but also made it challenging to manage finger movements in the depth dimension. Without clear spatial references, participants had to rely heavily on visual feedback to avoid unintentional actions, such as crossing virtual boundaries.

The larger size of both the mid-air and tablet keyboards necessitates greater arm movements compared to traditional touchscreen keyboards, further complicating the reliance on motor memory. The expansive gestures required for swiping across these larger keyboards make it difficult for participants to develop the kind of muscle memory they would in a smaller, more confined space, such as on a standard touchscreen. Training motor memory to perform such large, precise movements without visual guidance appears challenging, which may explain why participants continue to depend on visual feedback even with practice. The increased physical demands of covering a larger area could also contribute to this reliance on visual feedback, as participants need to ensure accurate finger placement over a wider space.

Moreover, despite the presence of auditory feedback to signal key events (such as when a finger lifts off the keyboard), participants still relied heavily on visual cues. This suggests that auditory feedback alone is insufficient to guide accurate movements in mid-air typing, likely due to the lack of tactile feedback, the sense of depth, and the larger space required for gestures. Visual feedback remains essential for maintaining control and precision throughout the interaction.

While this study focuses on gesture typing, other input methods such as direct touch typing, bimanual typing, and indirect input using a raycast interaction also impose unique visual demands. Understanding how users deploy their gaze during these methods, particularly in mid-air or virtual environments, could reveal insights into the visual-motor strategies required for efficient interaction and guide the design of more usable input systems.

5.3 Eye-Finger Coordination

In Mid-air typing, the coordination between eye and finger movements is tighter and more consistent compared to Tablet typing (RQ2). This means that while the eyes and fingers maintain a similar spatial relationship in both conditions, they are more synchronized in time during mid-air typing. The lower DTW distances in the mid-air condition show that eye and finger movements align more closely throughout the typing sequence.

We used the DTW distance to quantify both spatial and temporal dissimilarities between eye and hand movement sequences. Unlike the Partial Curve Mapping (PCM) method used in prior work [15], which only provides a dissimilarity score based on local alignment between portions of the sequences, DTW aligns the entire length of the sequences by warping their temporal alignments. This makes DTW more suitable for comparing eye-hand coordination during gesture typing, where continuous alignment across the entire swipe sequence is critical. In contrast, PCM's focus on local alignment makes it less appropriate for capturing the full complexity of eye-hand interactions in this context.

This synchronization is characterized by a consistent lag, where the eyes lead the fingers. The larger negative signed deviation indicates that the eyes consistently anticipate the next key before the fingers move to it. This lagged synchronization reflects a pattern

where the eyes guide the fingers, ensuring more precise timing, but with the fingers always following slightly behind. This tighter coordination suggests that, in mid-air typing, users rely heavily on visual feedback to control their finger movements, with the eyes setting the pace for the fingers to follow.

The cautious typing approach observed in mid-air gesture typing aligns with the findings of Liu et al. [27], who found that participants engaged their eyes with the next target earlier and for longer during more challenging tasks. This adaptation suggests that participants deliberately slowed their hand movements to manage the cognitive demands of visually searching for target information. Similarly, in the Mid-air condition of our study, participants exhibited longer fixations and a higher number of visual adjustments, reflecting their reliance on visual guidance to maintain accuracy in the absence of tactile feedback. This deliberate strategy likely prioritizes precision and mitigates errors in visually demanding environments.

Beyond synchronization, mid-air gesture typing imposes a higher cognitive load compared to tablet typing due to its reliance on visual attention and the lack of tactile feedback. The increased fixation frequency and duration observed in the mid-air condition highlights the necessity of visual guidance to compensate for the absence of physical boundaries. The NASA-TLX results further confirm the higher perceived mental and physical demands of mid-air typing. The lagged synchronization between eye and finger movements reflects a deliberate strategy to manage visuomotor challenges, with the eyes guiding finger movements to ensure precision.

5.4 Design Implications

Our findings suggest two possible design paths to improve mid-air gesture typing systems (RQ6):

Leveraging Eye-Hand Coordination for Predictive Models. Despite efforts to enhance mid-air gesture typing through gaze-based predictive models, previous methods have achieved limited success in boosting the typing performance [36, 50]. Our study provides insights into *why* these gaze-based methods may have achieved only limited improvements. First, as evidenced by our findings (see Fig.6), users do not fixate on all the letters in a word, especially during touchscreen typing. In fact, less than 40% of the letter keys in each phrase were covered by the participants' fixation ranges. This suggests that gaze is not always directly tied to finger movements, particularly when users are typing efficiently. In mid-air typing, the coordination we observed is a result of users adopting compensatory strategies to overcome the system's constraints, such as the lack of tactile feedback, rather than reflecting their natural typing behavior.

Second, even among the 40% of accurate fixations, the system does not yet know if they were made immediately before the finger moved toward that key. Given the extended lag observed between eye and finger movements in the mid-air condition, it is highly likely that gaze behavior is not always predictive of immediate finger actions. This further complicates the effectiveness of gaze-based predictive methods, as the gaze does not consistently lead finger movements, limiting the potential for gaze direction to accurately forecast typing actions. Moreover, when gaze and finger movements are closely synchronized, the model operates in a more closed-loop fashion, providing limited opportunities for predictive enhancements. This means that gaze behavior often lacks additional predictive signals beyond what is already reflected in finger movements, further reducing the effectiveness of gaze-based models.

The fluctuating nature of eye-finger coordination during different segments of the swiping gesture further complicates the use of gaze for prediction. Our segmentation analysis revealed that eye-hand synchronization deteriorated toward the final stages of gestures in both conditions. In Mid-air typing, there was notable fluctuation during the middle segments, possibly due to users visually searching for the next key or adjusting their finger speed. This variability indicates that gaze behavior is not always a straightforward predictor of the next intended key, making it challenging for models to accurately anticipate user input based on gaze alone.

Therefore, the limited success of these gaze-based methods can be attributed to the fundamental coordination challenges in mid-air typing, where gaze and finger movements are tightly coupled but often out of sync in ways that are difficult to predict. While these challenges limit the effectiveness of current gaze-based methods, the findings from this study offer an opportunity to inform the design of more sophisticated predictive models for mid-air word-gesture typing. By incorporating an understanding of the temporal lag and variability in gaze and finger coordination, future systems could dynamically adapt their predictions to better align with users' natural input patterns. Such designs could help address the inherent constraints of mid-air typing and improve efficiency in ways that existing models have yet to achieve.

Decoupling Visuomotor Requirements. Decoupling visuomotor coordination refers to reducing the heavy reliance on visual feedback and mitigating the constraints that force users to coordinate eye and finger movements so tightly. One of the primary limitations forcing the tight coordination between eye and hand is the hard threshold for touch activation on the virtual keyboard. A rigid boundary makes users overly cautious, slowing down their movements and increasing visual attention. By dynamically adjusting the threshold based on finger movement speed and proximity, accidental key activations could be minimized, allowing users to move more fluidly and confidently without constantly checking their finger position. Additionally, expanding the flexibility of gesture input to allow for more forgiving and adaptive recognition of swipes and finger movements would also help reduce the reliance on visual confirmation.

6 Conclusion

In summary, this paper provides a comprehensive empirical investigation into eye-hand coordination during mid-air gesture typing in mixed reality. By examining the absence of tactile feedback and its effect on typing performance, we observed that while users achieved comparable typing rates, mid-air typing introduced a lagged but tighter synchronization between eye gaze and finger movements. This consistent lag suggests that users adapt their visuomotor strategies to compensate for the absence of physical boundaries.

Our study also offers a rich dataset of visuomotor movements in both mid-air and tablet conditions, providing valuable insights into

the dynamics of gesture typing. These findings contribute actionable design implications for improving mid-air typing interfaces, particularly by addressing the cognitive and motor challenges of MR environments. The recommendations derived from our results lay the groundwork for future developments in optimizing MR text entry systems, aiming to enhance both performance and user experience.

Acknowledgments

John J. Dudley and Per Ola Kristensson were supported by EPSRC (Grant EP/W02456X/1).

References

[1] Jiban Adhikary and Keith Vertanen. 2021. Text entry in virtual environments using speech and a midair keyboard. *IEEE Transactions on Visualization and Computer Graphics* 27, 5 (2021), 2648–2658.

[2] Samantha Aziz and Oleg Komogortsev. 2022. An assessment of the eye tracking signal quality captured in the HoloLens 2. In *2022 Symposium on eye tracking research and applications*. 1–6.

[3] Jennifer K Bertrand and Craig S Chapman. 2023. Dynamics of eye-hand coordination are flexibly preserved in eye-cursor coordination during an online, digital, object interaction task. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13.

[4] Sibo Chen, Junce Wang, Santiago Guerra, Neha Mittal, and Soravis Prakkamakul. 2019. Exploring Word-gesture Text Entry Techniques in Virtual Reality. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–6. doi:10.1145/3290607.3312762

[5] John Dudley, Hrvoje Benko, Daniel Wigdor, and Per Ola Kristensson. 2019. Performance envelopes of virtual keyboard text input strategies in virtual reality. In *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 289–300.

[6] John J Dudley, Amy Karlson, Kashyap Todi, Hrvoje Benko, Matt Longest, Robert Wang, and Per Ola Kristensson. 2024. Efficient Mid-Air Text Input Correction in Virtual Reality. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 893–902.

[7] John J Dudley, Keith Vertanen, and Per Ola Kristensson. 2018. Fast and precise touch-based text entry for head-mounted augmented reality with variable occlusion. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 6 (2018), 1–40.

[8] John J Dudley, Jingyao Zheng, Aakar Gupta, Hrvoje Benko, Matt Longest, Robert Wang, and Per Ola Kristensson. 2023. Evaluating the performance of hand-based probabilistic text input methods on a mid-air virtual qwerty keyboard. *IEEE Transactions on Visualization and Computer Graphics* (2023).

[9] Anna Maria Feit, Daryl Weir, and Antti Oulasvirta. 2016. How we type: Movement strategies and performance in everyday typing. In *Proceedings of the 2016 chi conference on human factors in computing systems*. 4262–4273.

[10] Jens Grubert, Lukas Witzani, Eyal Ofek, Michel Pahud, Matthias Kranz, and Per Ola Kristensson. 2018. Text Entry in Immersive Head-Mounted Display-Based Virtual Reality Using Standard Keyboards. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, Reutlingen, 159–166. doi:10.1109/VR.2018.8446059

[11] Eve Hoggan, Stephen A Brewster, and Jody Johnston. 2008. Investigating the effectiveness of tactile feedback for mobile touchscreens. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1573–1582.

[12] Jinghui Hu, John J Dudley, and Per Ola Kristensson. 2022. An evaluation of caret navigation methods for text editing in augmented reality. In *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 640–645.

[13] Jinghui Hu, John J Dudley, and Per Ola Kristensson. 2024. LookUP: Command Search Using Dwell-free Eye Typing in Mixed Reality. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 981–989.

[14] Jinghui Hu, John J. Dudley, and Per Ola Kristensson. 2024. SkiMR: Dwell-free Eye Typing in Mixed Reality. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE, Orlando, FL, USA, 439–449. doi:10.1109/VR58804.2024.00065

[15] Xinhui Jiang, Yang Li, Jussi P.P. Jokinen, Viet Ba Hirvola, Antti Oulasvirta, and Xiangshi Ren. 2020. How We Type: Eye and Finger Movement Strategies in Mobile Typing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–14. doi:10.1145/3313831.3376711

[16] Xinhui Jiang, Yang Li, Jussi PP Jokinen, Viet Ba Hirvola, Antti Oulasvirta, and Xiangshi Ren. 2020. How we type: Eye and finger movement strategies in mobile typing. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[17] Roger Johansson, Åsa Wengelin, Victoria Johansson, and Kenneth Holmqvist. 2010. Looking at the keyboard or the monitor: relationship with text production processes. *Reading and writing* 23 (2010), 835–851.

[18] Roland S Johansson, Göran Westling, Anders Bäckström, and J Randall Flanagan. 2001. Eye–hand coordination in object manipulation. *Journal of neuroscience* 21, 17 (2001), 6917–6932.

[19] Jussi Jokinen. 2017. Touch screen text entry as cognitively bounded rationality. In *Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

[20] Florian Kern, Florian Niebling, and Marc Erich Latoschik. 2023. Text input for non-stationary XR workspaces: investigating tap and word-gesture keyboards in virtual and augmented reality. *IEEE Transactions on Visualization and Computer Graphics* 29, 5 (2023), 2658–2669.

[21] Per-Ola Kristensson. 2002. Design and evaluation of a shorthand aided soft keyboard. *Master's thesis, Linköping University, Sweden* (2002), 96.

[22] Per Ola Kristensson. 2007. *Discrete and continuous shape writing for text entry and control*. Ph. D. Dissertation. Institutionen för datavetenskap.

[23] Per-Ola Kristensson and Shumin Zhai. 2004. SHARK ^{2}: a large vocabulary shorthand writing system for pen-based computers. In *Proceedings of the 17th annual ACM symposium on User interface software and technology*. ACM, Santa Fe NM USA, 43–52. doi:10.1145/1029632.1029640

[24] Per-Ola Kristensson and Shumin Zhai. 2005. Relaxing stylus typing precision by geometric pattern matching. In *Proceedings of the 10th international conference on Intelligent user interfaces*. 151–158.

[25] Michael Land, Neil Mennie, and Jennifer Rusted. 1999. The roles of vision and eye movements in the control of activities of daily living. *Perception* 28, 11 (1999), 1311–1328.

[26] Luis A. Leiva, Sunjun Kim, Wenzhe Cui, Xiaojun Bi, and Antti Oulasvirta. 2021. How We Swipe: A Large-scale Shape-writing Dataset and Empirical Findings. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*. ACM, Toulouse & Virtual France, 1–13. doi:10.1145/3447526.3472059

[27] Xin Liu, Yao Zhang, Xianta Jiang, and Bin Zheng. 2023. Human eyes move to the target earlier when performing an aiming task with increasing difficulties. *International Journal of Human–Computer Interaction* 39, 6 (2023), 1341–1346.

[28] Lu Lu, Pengshuai Duan, Xukun Shen, Shijin Zhang, Huiyan Feng, and Yong Flu. 2021. Gaze-pinch menu: Performing multiple interactions concurrently in mixed reality. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 536–537.

[29] 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 (2022), 1–18.

[30] 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 (May 2022), 1–16. doi:10.1145/3530882

[31] Anders Markussen, Mikkel Rønne Jakobsen, and Kasper Hornbæk. 2014. Vulture: a mid-air word-gesture keyboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1073–1082.

[32] Anders Markussen, Mikkel Rønne Jakobsen, and Kasper Hornbæk. 2014. Vulture: a mid-air word-gesture keyboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Toronto Ontario Canada, 1073–1082. doi:10.1145/2556288.2556964

[33] Kseniia Palin, Anna Maria Feit, Sunjun Kim, Per Ola Kristensson, and Antti Oulasvirta. 2019. How do people type on mobile devices? Observations from a study with 37,000 volunteers. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–12.

[34] Alexandra Papoutsaki, Aaron Gokaslan, James Tompkin, Yuze He, and Jeff Huang. 2018. The eye of the typer: a benchmark and analysis of gaze behavior during typing. In *Proceedings of the 2018 acm symposium on eye tracking research & applications*. 1–9.

[35] 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*. 99–108.

[36] Yunlei Ren, Yan Zhang, Zhitao Liu, Yi Li, Li Yuan, and Ning Xie. 2024. Eye-Hand Typing: Eye Gaze Assisted Finger Typing via Bayesian Processes in AR. *IEEE Transactions on Visualization and Computer Graphics* (2024).

[37] Dario D Salvucci and Joseph H Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*. 71–78.

[38] Junxiao Shen, Jinghui Hu, John J Dudley, and Per Ola Kristensson. 2022. Personalization of a mid-air gesture keyboard using multi-objective bayesian optimization. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 702–710.

[39] Danqing Shi, Yujun Zhu, Jussi PP Jokinen, Aditya Acharya, Aini Putkonen, Shumin Zhai, and Antti Oulasvirta. 2024. CRTypist: Simulating Touchscreen Typing Behavior via Computational Rationality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.

[40] 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.

[41] Florian Spiess, Philipp Weber, and Heiko Schuldt. 2022. Direct interaction word-gesture text input in virtual reality. In *2022 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, 140–143.

[42] Keith Vertanen and Per Ola Kristensson. 2011. A versatile dataset for text entry evaluations based on genuine mobile emails. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*. 295–298.

[43] James Walker, Bochao Li, Keith Vertanen, and Scott Kuhl. 2017. Efficient Typing on a Visually Occluded Physical Keyboard. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, Denver Colorado USA, 5457–5461. doi:10.1145/3025453.3025783

[44] Wenge Xu, Hai-Ning Liang, Anqi He, and Zifan Wang. 2019. Pointing and Selection Methods for Text Entry in Augmented Reality Head Mounted Displays. In *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, Beijing, China, 279–288. doi:10.1109/ISMAR.2019.00026

[45] Naoki Yanagihara, Buntarou Shizuki, and Shin Takahashi. 2019. Text Entry Method for Immersive Virtual Environments Using Curved Keyboard. In *25th ACM Symposium on Virtual Reality Software and Technology*. ACM, Parramatta NSW Australia, 1–2. doi:10.1145/3359996.3365026

[46] Chun Yu, Yizheng Gu, Zhican Yang, Xin Yi, Hengliang Luo, and Yuanchun Shi. 2017. 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*. 4479–4488.

[47] Shumin Zhai and Per-Ola Kristensson. 2003. Shorthand writing on stylus keyboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Ft. Lauderdale Florida USA, 97–104. doi:10.1145/642611.642630

[48] Shumin Zhai and Per Ola Kristensson. 2012. The word-gesture keyboard: reimagining keyboard interaction. *Commun. ACM* 55, 9 (Sept. 2012), 91–101. doi:10.1145/2330667.2330689

[49] Shumin Zhai, Carlos Morimoto, and Steven Ihde. 1999. Manual and gaze input cascaded (MAGIC) pointing. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 246–253.

[50] Maozheng Zhao, Alec M Pierce, Ran Tan, Ting Zhang, Tianyi Wang, Tanya R. Jonker, Hrvoje Benko, and Aakar Gupta. 2023. Gaze Speedup: Eye Gaze Assisted Gesture Typing in Virtual Reality. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. ACM, Sydney NSW Australia, 595–606. doi:10.1145/3581641.3584072