

Behavioral Markers of Sustained Viewing in VR Museum Exploration: Head Stability, Gaze Dynamics, and Pupillary Response

ANONYMOUS AUTHOR(S)

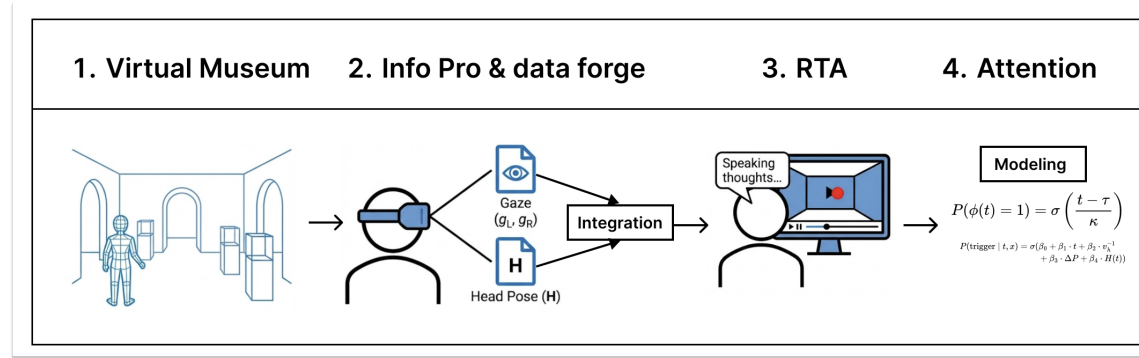


Fig. 1. Overview of our approach. Participants explored a virtual museum while we recorded gaze, pupil diameter, and 6DoF head pose. We analyzed behavioral markers that differentiate sustained viewing from casual glances, finding that head rotational stability, temporal gaze dynamics, and early pupil dilation provide complementary information for attention-based triggering.

We analyze behavioral markers of sustained object viewing during free VR museum exploration using 6DoF head motion, gaze dynamics, and pupil diameter (N=20). Viewing episodes were identified through retrospective think-aloud (RTA) coding, yielding 758 episodes for analysis. In a linear mixed model controlling for subject and exhibit, lower head rotational velocity significantly predicted longer dwell time ($\beta = -0.280$, $p < .001$). For episodes exceeding 3 seconds, gaze behavior exhibits a temporal shift: dispersion decreases while spatial entropy increases, suggesting transition from broad sampling to localized exploration. Event-related pupil responses time-locked to gaze onset were larger for long-dwell than short-dwell episodes (difference=0.107mm, $p = .007$, $d = 0.30$). Head stability and early pupil dilation provide complementary signals beyond dwell-time thresholds for context-aware caption triggering in VR museums.

CCS Concepts: • **Human-centered computing** → **Interactive systems and tools**; *User studies*; **Virtual reality**.

Additional Key Words and Phrases: Virtual Reality, Eye Tracking, Attention, Pupillometry, Head Motion

ACM Reference Format:

Anonymous Author(s). 2026. Behavioral Markers of Sustained Viewing in VR Museum Exploration: Head Stability, Gaze Dynamics, and Pupillary Response. In *Proceedings of Symposium on Eye Tracking Research & Applications (ETRA '26)*. ACM, New York, NY, USA, 8 pages. <https://doi.org/XXXXXXX.XXXXXXX>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Manuscript submitted to ACM

Manuscript submitted to ACM

1 Introduction

Immersive VR applications such as virtual museums increasingly seek to deliver context-aware information, for example, displaying exhibit captions or audio descriptions when users show interest in specific objects [Dondi and Porta 2023; Lu et al. 2021; Shimizu et al. 2025]. A core technical challenge is determining the appropriate trigger for such information delivery. Fixed dwell-time thresholds (e.g., displaying a caption after 500ms of gaze on an object) are widely used [Isomoto et al. 2023; Špakov and Miniotos 2004] but suffer from two problems. First, users may fixate on uninteresting objects due to visual complexity, fatigue, or incidental viewing, leading to false-positive triggers [Zhang et al. 2011]. Second, users may briefly glance at interesting objects they intend to revisit, resulting in missed opportunities.

Head-mounted displays (HMDs) with integrated eye tracking provide richer behavioral data than traditional 2D displays. In addition to gaze position, VR systems continuously record 6 degrees-of-freedom (6DoF) head pose, enabling analysis of head-gaze coordination. Commercial eye-tracking HMDs also provide pupil diameter measurements, which prior research has linked to cognitive load and arousal [Beatty 1982; Mathôt 2018].

This paper investigates whether these additional signals, specifically head motion stability and pupillary response, provide complementary information for dwell-based triggering. We analyze data from 20 participants exploring a virtual museum and address three research questions.

- (1) **RQ1.** Does head stability during object viewing predict viewing duration?
- (2) **RQ2.** Do viewing episodes exhibit characteristic temporal patterns in gaze behavior?
- (3) **RQ3.** Does pupil dilation at gaze onset differ between long and short viewing episodes?

We make three contributions. First, we show that head rotational stability significantly predicts dwell time in a linear mixed model. Second, we provide evidence of a two-phase viewing pattern with statistically significant temporal differences in gaze dispersion and entropy. Third, we show that event-related pupil response at gaze onset discriminates between long and short viewing episodes, which could serve as an early signal for adaptive triggering.

2 Related Work

Dwell time, defined as the cumulative duration of gaze on an area of interest (AOI), is the most widely used metric for inferring visual attention [Holmqvist et al. 2011]. In VR, dwell-based triggers are commonly used to activate UI elements or display information [Mutasim et al. 2021; Shimizu et al. 2025]. Research on dwell-time optimization has explored speed-accuracy tradeoffs [Zhang et al. 2011] and online adjustment methods [Špakov and Miniotos 2004]. Isomoto et al. [Isomoto et al. 2023] examined dwell time from a cognitive process perspective, suggesting that human processing stages influence optimal thresholds. However, dwell time alone has limited discriminative power for cognitive states such as interest or comprehension [Just and Carpenter 1976].

Eye tracking in VR HMDs presents unique challenges. Recent evaluations have characterized signal quality limitations in commercial devices such as the Meta Quest Pro [Aziz and Komogortsev 2024], while deep learning approaches have improved tracking robustness and precision [Barkevich et al. 2024], enabling behavioral signal extraction for real-time applications.

VR systems provide behavioral signals beyond gaze position. Unlike desktop eye tracking, VR continuously tracks 6DoF head pose [Hu 2020]. Land [Land 2009] characterized head-eye coordination during visual attention, and recent work has modeled this coordination in VR contexts [Pan et al. 2025], showing that head motion patterns differ between task types [Clay et al. 2019]. Head stability during viewing may reflect focused attention, as users naturally minimize body movement when concentrating.

Pupil dilation is an established marker of cognitive load, arousal, and interest [Beatty 1982; Sweller 1988]. Event-related pupil responses (ERPRs) time-locked to fixation onset have been used to study attention capture and stimulus salience [Mathôt 2018]. Recent VR pupillometry work includes Pielage et al. [Pielage et al. 2025], who validated VR-based pupillometry for listening effort, and Grootjen et al. [Grootjen et al. 2024], who demonstrated real-time pupil-based adaptation of reading speed. These results indicate that pupil signals can drive adaptive VR interfaces despite HMD-specific artifacts.

Recent work has explored gaze-based information presentation in VR museums and exhibitions [Dondi and Porta 2023; Lu et al. 2021]. Shimizu et al. [Shimizu et al. 2025] developed a floating caption system that displays exhibit information when users fixate on designated gaze points. The spatial contiguity principle suggests that integrating information near its referent improves learning [Ginns 2006]. However, existing systems rely on fixed dwell thresholds without considering additional behavioral signals. Scene viewing research has identified characteristic phases in viewing behavior, with an initial overview phase followed by detailed examination [Unema et al. 2005]. Whether similar patterns emerge in VR object viewing remains unclear.

3 Methods

Twenty participants (N=20; 13 male, 7 female; mean age 31.2 years, SD=5.4) with normal or corrected-to-normal vision completed the study. We used a PICO 4 Enterprise HMD with integrated binocular eye tracking (90Hz sampling) and 6DoF inside-out tracking. The VR application was developed in Unity.

The virtual museum contained 6 exhibits arranged in a 2×3 grid (Figure 2). Participants navigated freely using joystick locomotion. After eye tracker calibration, participants explored the museum for 5–6 minutes. They then completed a retrospective think-aloud (RTA) session watching a replay of their session to verbalize their thoughts. Viewing episodes were automatically segmented based on gaze target ID transitions (continuous gaze on a single exhibit, minimum 200ms). Two coders independently reviewed the RTA transcripts to identify episodes where participants expressed interest or engagement with an exhibit; automatic boundaries were adjusted when verbal reports indicated viewing continuation. Inter-coder agreement was high (Cohen’s $\kappa = 0.89$); disagreements were resolved through discussion, yielding 758 validated episodes for analysis.

We logged synchronized time-series data at 60Hz (downsampled from 90Hz eye tracking for synchronization with Unity frame rate), including timestamp, gaze target ID, pupil diameter (left/right), world gaze point coordinates, head pose (position and quaternion), and blink count. Blinks were detected via the PICO SDK; gaps <200ms were linearly interpolated, and binocular pupil diameter was computed as the mean of left and right eyes. Head translational velocity was computed as $v_{pos} = \|\mathbf{p}_t - \mathbf{p}_{t-1}\| / \Delta t$ (m/s). For rotational velocity, we computed the relative quaternion $q_{rel} = q_{t-1}^{-1} \otimes q_t$ and extracted angular velocity as $\omega = 2 \cdot \arccos(|q_{rel,w}|) / \Delta t$ (rad/s). For each viewing episode, we computed dispersion (SD of gaze positions), spatial entropy (normalized entropy of 10×10 gaze histogram), and scanpath length. For episodes ≥ 3 seconds, we divided each into three equal temporal phases (Early, Middle, Late). For event-related pupil response (ERPR), we extracted epochs from gaze onset to 1000ms after, excluding epochs with >20% missing data. To avoid baseline confounds from luminance differences between objects, we used the 0–200ms window after gaze onset as baseline (before task-evoked dilation typically emerges) and measured pupil change in the 200–700ms window.

Statistical analysis used linear mixed models (LMM) implemented via statsmodels, with crossed random intercepts for subject and exhibit to account for the nested structure. Post-hoc comparisons used paired t-tests with Bonferroni correction.



Fig. 2. Virtual museum environment. Participants freely explored 6 exhibits (3D artworks on pedestals with rope barriers) using joystick locomotion. The pink dot indicates the current gaze point, which was recorded along with head pose and pupil diameter at 60Hz. The environment featured consistent lighting to minimize luminance-related pupil artifacts.

4 Results

The dataset contained 758 viewing episodes across 20 participants and 6 exhibits. Mean dwell time was 9.41s (SD=8.14). Table 1 summarizes key metrics.

Table 1. Descriptive statistics for viewing episodes (N=758)

Metric	Mean	SD
Dwell time (sec)	9.41	8.14
Head pos velocity (m/s)	0.54	0.36
Head rot velocity (rad/s)	0.18	0.12
Pupil SD (mm)	0.26	0.16
Spatial entropy (normalized)	0.47	0.14

We fit a linear mixed model predicting log-transformed dwell time from standardized head velocity metrics, with crossed random intercepts for subject and exhibit. Head rotational velocity significantly predicted dwell time ($\beta=-0.280$, $SE=0.078$, $p<.001$), with the negative coefficient indicating that lower velocity (greater stability) was associated with longer viewing (Table 2, Figure 3). Head translational velocity showed a weaker effect ($\beta=-0.158$, $p=.047$), which may partially reflect locomotion rather than pure head stability (see Limitations). The substantial random effect variance for subjects (0.512) indicates meaningful individual differences.

Spatial entropy showed a moderate positive correlation with viewing duration ($r=0.27$, $p<.001$). For viewing episodes with dwell time ≥ 3 seconds ($N=312$, 41% of total episodes), we analyzed temporal dynamics by dividing each episode into three equal phases (Early, Middle, Late). A linear mixed model with Phase (3 levels) as fixed effect and random intercepts for subject and exhibit revealed significant effects of temporal phase on dispersion ($\chi^2(2)=9.82$, $p=.007$) and entropy ($\chi^2(2)=21.4$, $p<.001$). Pairwise comparisons showed that dispersion decreased from Early to Middle phase ($d=0.59$), while entropy increased from Early to Middle ($d=0.73$) and from Early to Late ($d=0.75$). This pattern suggests a

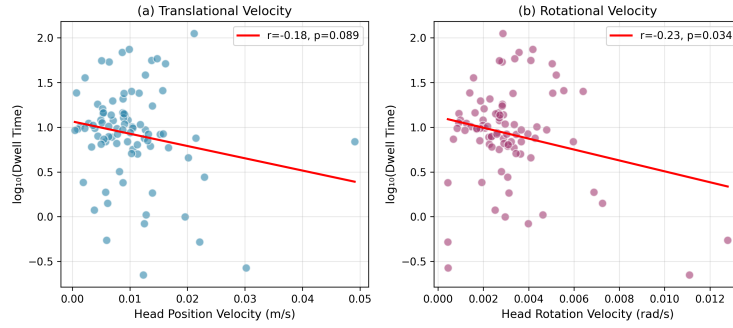


Fig. 3. Head stability and viewing duration. Each point represents one subject \times exhibit mean ($N=83$; pairs without valid viewing episodes excluded). (a) Head translational velocity shows a negative trend with log-transformed dwell time. (b) Head rotational velocity shows a stronger negative relationship. Note: Statistical inference is from LMM on episode-level data ($N=758$), not these aggregated correlations.

Table 2. Linear Mixed Model results predicting log(Dwell Time)

Predictor	β	SE	z	p
Intercept	1.797	0.228	7.87	<.001
Head Pos Velocity _z	-0.158	0.080	-1.98	.047*
Head Rot Velocity _z	-0.280	0.078	-3.60	<.001***
Pupil SD _z	0.093	0.093	1.00	.316
Random Effects	Subject Var=0.512, Exhibit Var=0.118			

* $p < .05$, ** $p < .01$, *** $p < .001$

two-phase viewing structure: an initial overview phase with broad gaze sampling, transitioning to detailed exploration with spatially distributed fixations within a localized region (Figure 4).

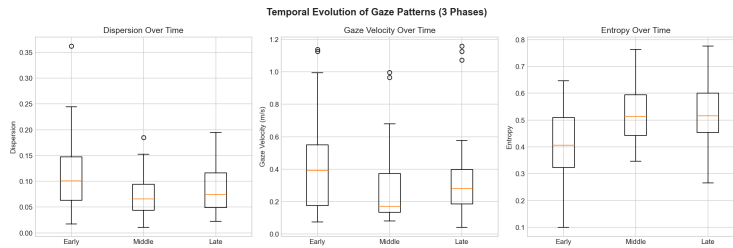


Fig. 4. Temporal evolution of gaze patterns across viewing phases. Each viewing episode ($\geq 3s$) was divided into Early, Middle, and Late thirds. Left: Dispersion (SD of gaze positions) decreases over time, indicating transition from broad sampling to focused viewing. Center: Gaze velocity shows similar reduction. Right: Spatial entropy increases, reflecting more distributed exploration within increasingly localized regions. This two-phase pattern (overview \rightarrow exploration) parallels findings from scene viewing research.

For event-related pupil response (ERPR), we used the 0–200ms window after gaze onset as baseline (before task-evoked dilation emerges) and computed mean pupil change in the 200–700ms window. This baseline choice avoids confounds from luminance and vergence differences between the previously viewed object and the current target. Of 758 episodes, 537 yielded valid epochs (excluding those with $>20\%$ missing data). A median split of these 537 epochs

classified episodes as Long-dwell (N=269) or Short-dwell (N=268). We fit a linear mixed model with dwell category as fixed effect and random intercepts for subject and exhibit. Long-dwell episodes showed significantly larger early pupil dilation than short-dwell episodes (estimate=0.107mm, SE=0.040, $\chi^2(1)=7.16$, $p=.007$, $d=0.30$), suggesting that pupil response shortly after gaze onset may serve as an early signal of sustained viewing (Figure 5).

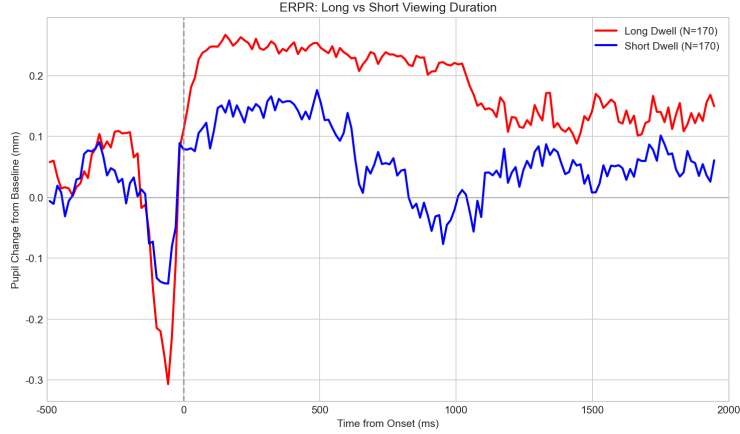


Fig. 5. Event-related pupil response (ERPR) time-locked to gaze onset. Red: Long-dwell episodes (above-median, N=269). Blue: Short-dwell episodes (below-median, N=268). The dashed line marks gaze onset ($t=0$). Statistical inference uses the mean pupil change in the 200–700ms window relative to a 0–200ms baseline (same-object baseline avoids cross-object luminance confounds). Long-dwell episodes show larger dilation (estimate=0.107mm, $p=.007$, $d=0.30$).

5 Discussion

Head stability as a predictor of viewing duration has direct applications in VR. Unlike desktop eye tracking, VR systems have continuous access to 6DoF head pose at no additional cost. The negative relationship between head velocity and dwell time suggests that when users engage in sustained viewing, they stabilize their head to facilitate detailed visual inspection. This aligns with research on head-eye coordination during focused attention [Land 2009]. In VR, where users can move their entire body, choosing to stabilize the head may be a strong signal of attentional engagement. VR applications could incorporate head stability as a feature in attention-based triggering, potentially reducing false positives during casual glances when the head is still moving.

The two-phase viewing pattern we observed parallels findings from scene viewing research [Unema et al. 2005]. The transition from low to high entropy reflects increasingly distributed gaze within a progressively narrower spatial region. Users first broadly sample the exhibit, then explore specific details with more varied fixation patterns. This temporal structure suggests that caption timing could be adaptive, with information presented during the Middle or Late phase (after approximately 1 second) rather than immediately upon gaze arrival. Pupil dilation at gaze onset also predicts subsequent dwell time. Pupil dilation reflects cognitive load and arousal [Beatty 1982]. Although the effect size was modest ($d=0.30$), pupil response could serve as an early signal to adjust trigger probability in combination with other features.

Designers of context-aware VR systems could combine these behavioral signals. Head stability could weight trigger probability by inverse head velocity. Viewing phase information could delay triggers until the Middle or Late phase.

Pupil response could increase trigger probability when dilation exceeds baseline, while dwell time remains the primary feature with adaptive thresholds. The substantial individual differences captured by our random effects suggest that personalized thresholds may outperform fixed values.

5.1 Toward a Unified Trigger Model

Our findings suggest that multiple behavioral signals can be combined for context-aware triggering. We propose a composite trigger probability that integrates dwell time, head stability, and early pupil response:

$$P(\text{trigger} \mid t, \mathbf{x}) = \sigma(\beta_0 + \beta_1 \cdot \log(t) + \beta_2 \cdot S_h + \beta_3 \cdot \Delta P) \quad (1)$$

where $\sigma(\cdot)$ is the sigmoid function, t is current dwell time, $S_h = 1/(v_{\text{rot}} + \epsilon)$ is head stability (inverse rotational velocity in rad/s), and ΔP is pupil dilation from baseline (mm).

Each variable requires calibration. Pupil baseline P_0 is computed as the mean pupil diameter during the 0–200ms window after gaze onset (before task-evoked dilation emerges); $\Delta P = P_{200-700} - P_0$ then captures task-evoked dilation. Head stability S_h is computed from rotational velocity averaged over a sliding 500ms window; $\epsilon = 0.01$ prevents division by zero during complete stillness. The log-transform on time captures diminishing information gain from additional dwell duration.

The coefficients $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$ can be calibrated via logistic regression using RTA-coded binary labels (interested/not interested) as ground truth. Given a labeled dataset of viewing episodes with features \mathbf{x}_i and labels $y_i \in \{0, 1\}$, the optimal coefficients minimize cross-entropy loss. Leave-one-subject-out cross-validation provides unbiased estimates of generalization performance. For online personalization, user-specific coefficients $\beta^{(u)}$ can be initialized from population priors and updated via maximum a posteriori estimation as new labeled episodes accumulate, enabling the system to adapt to individual viewing styles within a single session.

5.2 Limitations

Several limitations constrain interpretation. First, head translational velocity may confound head stability with locomotion; joystick movement was not masked from the analysis, so the weaker positional velocity effect ($\beta = -0.158$) should be interpreted cautiously. Future work should separate head-in-world motion from head-relative-to-body micro-movements. Second, the temporal phase analysis focused on episodes with dwell time $\geq 3s$ ($N=312$), which improves reliability but may bias toward more engaged viewing; effects for shorter episodes remain unclear. We modeled phase effects with an LMM including subject and exhibit random intercepts, but future work should test robustness under alternative phase definitions (e.g., fixed-time windows). Third, pupillometry in VR is susceptible to artifacts from vergence, accommodation, and luminance changes despite our controlled lighting; we did not model object luminance or viewing distance. Fourth, we used dwell time as a proxy for engagement without ground-truth interest labels; although RTA data exists, systematic coding for interest was not completed. Fifth, current analyses use episode-level aggregate features; real-time triggering requires evaluation of early-window features (e.g., first 500ms) with proper cross-validation.

6 Conclusion

We investigated behavioral markers of sustained viewing during VR museum exploration beyond traditional dwell time metrics. Head rotational stability significantly predicts dwell time, viewing episodes show a characteristic two-phase temporal pattern, and pupil dilation at gaze onset is larger for long-dwell episodes. By incorporating head stability and

early pupil response alongside dwell time, VR applications could achieve more accurate attention-based triggering with fewer false positives. Future work should validate these markers against explicit interest ratings and develop composite triggering algorithms with early-window prediction.

Acknowledgments

Privacy and Ethics Statement: Gaze, head motion, and pupil data were collected with informed consent and stored as anonymized identifiers. Because attention inference can be misused for surveillance, deployments should provide clear disclosure and opt-in consent.

Use of AI Writing Tools: We used AI-assisted writing tools to improve clarity in early drafts; all technical claims, analyses, and interpretations were reviewed and verified by the authors.

References

- Shafayat Aziz and Oleg V Komogortsev. 2024. Evaluation of Eye Tracking Signal Quality for Virtual Reality Applications: A Case Study in the Meta Quest Pro. In *Proc. ETRA Adjunct*.
- Nicholas Barkevich, Reynold J Bailey, and Jeff B Pelz. 2024. Using Deep Learning to Increase Eye-Tracking Robustness, Accuracy, and Precision in Virtual Reality. *Proc. ACM Comput. Graph. Interact. Tech.* 7, 1 (2024).
- Jackson Beatty. 1982. Task-Evoked Pupillary Responses, Processing Load, and the Structure of Processing Resources. *Psychological Bulletin* 91, 2 (1982), 276–292.
- Viviane Clay, Peter König, and Sabine U König. 2019. Eye Tracking in Virtual Reality. *Journal of Eye Movement Research* 12, 1 (2019).
- Piercarlo Dondi and Marco Porta. 2023. Gaze-Based Human-Computer Interaction for Museums and Exhibitions: Technologies, Applications and Future Perspectives. *Electronics* 12 (2023), 3064.
- Paul Ginns. 2006. Integrating Information: A Meta-Analysis of the Spatial Contiguity and Temporal Contiguity Effects. *Learning and Instruction* 16, 6 (2006), 511–525.
- Marc Grootjen, Anna Maria Feit, Daniel Bachmann, Daniel Weiskopf, and Andreas Bulling. 2024. Your Eyes on Speed: Using Pupil Dilation to Adaptively Select Speed-Reading Parameters in Virtual Reality. *Proc. ACM Hum.-Comput. Interact.* 8, ETRA (2024).
- Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. 2011. *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford University Press.
- Zhiming Hu. 2020. Gaze Analysis and Prediction in Virtual Reality. In *IEEE VR Workshops*. 543–544.
- Takuto Isomoto, Shota Yamanaka, and Buntarou Shizuki. 2023. Exploring Dwell-time from Human Cognitive Processes for Dwell Selection. *Proc. ACM Hum.-Comput. Interact.* 7, ETRA (2023).
- Marcel Adam Just and Patricia A Carpenter. 1976. Eye Fixations and Cognitive Processes. *Cognitive Psychology* 8, 4 (1976), 441–480.
- Michael F Land. 2009. Vision, Eye Movements, and Natural Behavior. *Visual Neuroscience* 26, 1 (2009), 51–62.
- Haoyang Lu, Hao Ren, Yansong Feng, Shuhui Wang, Shuai Ma, and Wen Gao. 2021. GazeTance Guidance: Gaze and Distance-Based Content Presentation for Virtual Museum. In *IEEE VR Workshops*.
- Sebastiaan Mathôt. 2018. Pupillometry: Psychology, Physiology, and Function. *Journal of Cognition* 1, 1 (2018).
- Abdul Karim Mutasim, Anil U Batmaz, and Wolfgang Stuerzlinger. 2021. Pinch, Click, or Dwell: Comparing Different Selection Techniques for Eye-Gaze-Based Pointing in Virtual Reality. In *Proc. ETRA*.
- Xufeng Pan, Mel Slater, and Maria V Sanchez-Vives. 2025. Head-EyeK: Head-Eye Coordination and Control Learned in Virtual Reality. *IEEE Trans. Vis. Comput. Graph.* 31, 1 (2025), 1–12.
- Hidde Pielage, Adriana A Zekveld, Niek J Versfeld, and Sophia E Kramer. 2025. Using Pupillometry in Virtual Reality as a Tool for Speech-in-Noise Research. *Ear and Hearing* 46, 3 (2025), 678–690.
- Kosuke Shimizu, Shoya Komatsu, and Hidenori Watanabe. 2025. Floating Captions: Eyetracking-based Context-aware Captioning System for Immersive VR Contents. In *Proc. ETRA*.
- John Sweller. 1988. Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science* 12, 2 (1988), 257–285.
- Pieter JA Unema, Sebastian Pannasch, Markus Joos, and Boris M Velichkovsky. 2005. Time Course of Information Processing during Scene Perception. *Visual Cognition* 12, 3 (2005), 473–494.
- Xuan Zhang, Peng Xu, Qiang Zhang, and Hongbin Zha. 2011. Speed-Accuracy Trade-off in Dwell-Based Eye Pointing Tasks at Different Cognitive Levels. In *Proc. PETMEI*. 37–42.
- Oleg Spakov and Darius Miniotas. 2004. On-Line Adjustment of Dwell Time for Target Selection by Gaze. In *Proc. NordiCHI*. 203–206.