Journal of
Eye Movement Research

MDPI

Article

Gaze Information Channels in Head-Mounted Display Eye Tracking: A Pilot Study on the GazeBaseVR Dataset

Micaela Y. Martin [1,*], Marius Vila [2], Miguel Chover [3] and Mateu Sbert [2,*]

[1] Department of Computer Science and Systems Engineering, Escuela Universitaria Politécnica de Teruel, University of Zaragoza, 44003 Teruel, Spain
[2] Department of Computer Science, Applied Mathematics and Statistics, Universitat de Girona, 17003 Girona, Spain; marius.vila@udg.edu
[3] Department of Computer Languages and Systems, Universitat Jaume I, 12071 Castellón de la Plana, Spain; chover@uji.es
* Correspondence: micaelayanet.martin@unizar.es (M.Y.M.); mateu.sbert@udg.edu (M.S.)

Abstract

Eye movements offer one of the most direct observable traces of attention: the size, duration and direction of one fixation constrain the next, and that sequential dependence carries information about both the task and the person. Mutual information (MI) quantifies it, and a family of such gaze channels—saccade amplitude, fixation duration, saccade direction—has been developed for flat-screen viewing. We add vergence, which any binocular eye tracker records, as a proxy for the depth of the fixated point, and evaluate it on a large public virtual-reality dataset of several hundred participants recorded over two same-day sessions on four contrasting tasks. Amplitude MI differs more between the four task means than between participant means, and the results for vergence MI show the reverse by a factor of 1.3 to 5.8 depending on the discretisation; part of vergence's between-participant spread reflects resting vergence rather than behaviour. Vergence MI decreased by approximately 68% after subtracting a rolling median from a within-recording-tercile version of the channel, which is a step that removes slow components (genuine eye movements, baseline changes or measurement drift) that the analysis cannot distinguish, so the channel as defined here quantifies sequential dependence in the binocular signal and has not yet been validated as a specific measure of depth-related visual behaviour. Vergence is the most task-stable of the four channels and the least affected by finite-sample bias. Its between-session ICC is modest; it is highest during reading, where the text changes between sessions, so this is not a clean test–retest estimate.

Keywords: gaze information channel; mutual information; eye tracking; virtual reality; head-mounted display; Markov chain; vergence channel; GazeBaseVR

Check for updates

Academic Editor: Alexander Maedche

Received: 18 October 2026
Revised: 18 November 2026
Accepted: 19 November 2026
Published: 24 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Eye-tracking data from head-mounted displays (HMDs) differs from screen-based data in at least three important ways: the observer is free to rotate their head, gaze must be ray-cast into a three-dimensional scene, and binocular vergence carries a signal related to the depth of the fixated point. The latter is a signal that flat-screen paradigms, with their single fixed depth, rarely exercise. Despite the rapid growth of HMD eye-tracking research, the information-theoretic structure of gaze sequences in this setting remains largely unexplored.

The gaze information channel framework, introduced by Ma et al. [1] and developed in a series of subsequent studies [2–4], offers a principled way to characterise this structure.

By modelling the sequence of fixation states as a first-order Markov chain and computing the mutual information $I(X;Y)$ of the resulting channel, the framework quantifies the sequential predictability of each gaze attribute, that is, how much knowing the current state reduces uncertainty about the next. Applied to a screen-based viewing of paintings, the framework has revealed that different gaze channels (area of interest, amplitude, duration, direction, pupil diameter) carry markedly different amounts of sequential information, and that some channels discriminate primarily across stimuli while others discriminate across observers [4].

The present study extends the framework to HMD eye-tracking for the first time. Because no labelled VR gaming dataset with concurrent immersion ratings yet exists, we conduct a pilot analysis on GazeBaseVR [5], which is a large, publicly available, longitudinal binocular eye-tracking dataset collected in a VR headset. We make four specific contributions (Figure 1 gives an overview of the study):

1. The first computation of gaze information channel MI from HMD eye-tracking data, covering the amplitude, duration, direction, and vergence channels.
2. A cross-task comparison (video viewing, depth alternation, reading, random saccades) that probes how MI varies with task structure in the HMD setting.
3. Adding vergence to the gaze information channel framework and applying it in an exploratory analysis of head-mounted display data.
4. An H(X) vs. MI diagnostic that flags recordings whose state space is effectively reduced so that MI describes transitions among only one or two occupied states.

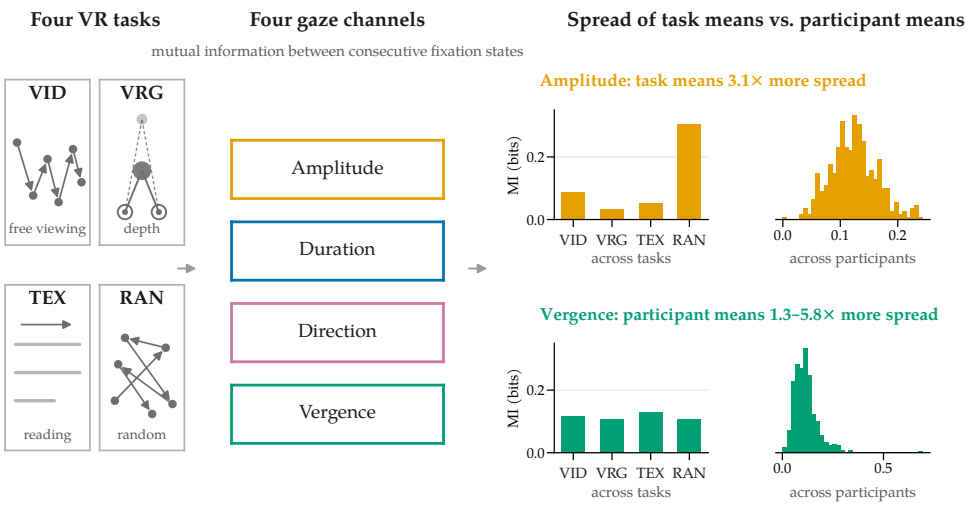


Figure 1. Overview of the study. Four tasks from a public virtual-reality eye-tracking dataset (**left**) are reduced to sequences of discrete gaze states, from which four information channels are computed (**centre**); vergence is introduced here and is computed from binocular disparity. The channels differ in what they are sensitive to (**right**): amplitude's task means are more widely spread than its participant means, and vergence's the reverse. The bars are the mean bias-corrected MI per task and the histograms the distribution of participant means, which are both computed from the analysed recordings; the amplitude bars span 0.033 (VRG) to 0.304 (RAN), which is the range given in Section 5.2. The light grey horizontal line in the two bar charts is a reference gridline at 0.2 bits. The 3.1× and 1.3–5.8× figures are the ratios $\sigma_{\text{task}}/\sigma_{\text{obs}}$ of Section 4.5, which are inverted for vergence. They compare two standard deviations and are descriptive; they do not decompose the variance or attribute it to task or participant. The vergence figure is the range across the six discretisations of Appendix A.6. Part of vergence's between-participant variation reflects differences in resting vergence rather than in behaviour (Section 4.4).

These results establish the empirical baseline against which our forthcoming full study—coupling the channel framework with immersion questionnaires in interactive VR gaming—will be evaluated.

2. Related Work

2.1. Information-Theoretic Gaze Analysis

Shannon's information theory [6] has been applied to eye movements since at least Ellis and Stark [7], who used first-order Markov transition matrices over cockpit display areas of interest (AOIs) to characterise pilot scanning behaviour. Krejtz et al. [8,9] formalised stationary gaze entropy (SGE) and gaze transition entropy (GTE)—the marginal and conditional Shannon entropies of the fixation sequence, respectively—and validated both on gallery viewing data. Shiferaw et al. [10] subsequently reviewed the entropy-based literature and emphasised that SGE and GTE must always be interpreted jointly, because high GTE can arise from either genuine exploration or from stimulus-driven scanning with a flat marginal distribution.

The gaze information channel framework was introduced by Ma et al. [1], who modelled fixation sequences between AOIs as a communication channel and applied it to Van Gogh paintings. The framework was subsequently extended beyond AOIs: Hao et al. [2] introduced the fixation duration channel, finding observer-driven variation dominates over stimulus-driven variation, which is the contrast that gives the discrimination taxonomy used below; Hao et al. [3] introduced the saccade direction channel, finding that horizontal and vertical saccades carry more sequential structure than diagonal ones. Most recently, Vila et al. [4] introduced the saccade amplitude and pupil diameter channels, reporting screen-based baselines of $I(X;Y) = 0.029$ bits (amplitude, participant mean, 3-category) and $I(X;Y) = 0.489 \pm 0.209$ bits (pupil, participant mean), and proved two theoretical results: an upper bound on conditional entropy in terms of transition persistence (Proposition 1 in [4]) and a refinement monotonicity result showing that finer discretisation cannot decrease channel MI (Remark 2 there). Across all five channels, 19 of 20 pairwise Spearman rank correlations between $I(X;Y)$ values, corresponding to ten channel pairs tested at two levels (per observer and per painting), were non-significant, suggesting the channels capture partially distinct aspects of gaze behaviour. For the amplitude channel, they report that observer-driven variation exceeds stimulus-driven variation, placing amplitude with duration and direction rather than with AOI in the discrimination taxonomy.

2.2. Eye Tracking in VR Head-Mounted Displays

Commercial HMDs with integrated eye trackers have enabled a growing body of VR gaze research [11]. Cheng and Chukoskie [12] noted that gaze entropy is increasingly used as a complement to fixation count in VR studies, but that analysis methods established for desktop eye-tracking do not transfer directly to HMDs due to head-movement confounds and calibration drift. GazeBaseVR has been used almost exclusively for questions of identity. Its authors designed it for eye-movement biometrics, and subsequent work has largely followed: end-to-end authentication models [13], breadth-first authentication baselines on very large populations [14], comparisons of headset against high-end-tracker biometric performance [15], continuous authentication over long intervals [16], and de-anonymisation studies demonstrating that gaze in extended reality can reveal identity [17]. The complementary strand is privacy protection, including differentially private synthesis [18]. What these have in common is that gaze is treated as a signature of who a person is. Two exceptions ask instead what the signal itself contains: Zandi and Luhan [19] apply multiscale entropy to all 5,020 recordings to characterise the complexity and predictability of gaze across the five tasks, and a separate strand uses the recordings for visual fatigue

detection [20]. Neither constructs a transition matrix or computes channel measures, so the information carried by the sequence of gaze states—as opposed to the complexity of the raw signal—has not been examined on this dataset. The present study asks that question.

Of this body of work, only one study shares our modelling assumption. Ren et al. [18] applied a Markov chain model to VR gaze data, including GazeBaseVR, to synthesise gaze paths under local differential privacy. To our knowledge, this is the only prior use of Markov chain gaze modelling on HMD data, and because it shares both the modelling assumption and one of our datasets, the relationship to the present work is worth setting out.

The two approaches use the Markov chain for opposite purposes. DPGazeSynth is generative: the transition matrix is a sampling distribution from which surrogate gaze paths are drawn, and the quantity of interest is the fidelity of those surrogates to the original data. Our use is analytic: the transition matrix is not sampled from but summarised, through the entropies and mutual information it induces, and the quantity of interest is how much sequential structure a recording contains. The state spaces differ accordingly. Ren et al. model gaze position, distinguishing fixation from saccade in order to reproduce scan-path geometry; we discard position entirely and model derived attributes of each fixation–saccade pair—amplitude, duration, direction and vergence—so that our states carry no spatial information at all.

The distinction matters for estimation. A generative model tolerates a noisy or biased transition matrix provided the sequences it produces remain plausible; indeed, DPGazeSynth deliberately perturbs transition frequencies through the OUE protocol to obtain its privacy guarantee. Entropy and mutual information are not robust in the same way: they are non-linear functionals of the matrix, and the same perturbation that leaves generation serviceable would inflate the estimator bias quantified in Section 4.3. The two lines of work therefore need different things from the same model, and neither subsumes the other. No channel measures—$H(X)$, $H(Y|X)$ or $I(X;Y)$—are reported in [18], and no prior study has applied the gaze information channel framework to HMD data.

2.3. GazeBaseVR

GazeBaseVR [5] is a large-scale, longitudinal, binocular eye-tracking dataset collected at 250 Hz using an SMI ET-HMD (HTC Vive-based). It comprises 5020 recordings from 407 participants performing five tasks per session: a vergence task (VRG; we call it the depth-alternation task below, to keep it distinct from the vergence channel), a smooth pursuit task (PUR), a video-viewing task (VID), a self-paced reading task (TEX), and a random oblique saccade task (RAN). The present study analyses four of these five tasks (VRG, VID, TEX, RAN); the smooth pursuit task (PUR) is excluded because smooth pursuit does not produce the discrete fixation–saccade sequences required for first-order Markov channel estimation. The dataset is publicly available under a CC BY 4.0 licence on Figshare [21] and is directly loadable via the pymovements library [22], although the analysis reported here reads the released CSV files directly and uses its own event detection (Section 3.3).

3. Methods

3.1. Dataset and Participants

GazeBaseVR comprises 5020 binocular recordings from 407 college-aged participants collected over 26 months in three longitudinal rounds. We analysed Round 1 exclusively, which includes both within-visit sessions (S1 and S2, separated by approximately 30 min) for all 407 participants. Round 1 is the only round with complete data from the full sample; Rounds 2 and 3 are successively smaller subsets of returnees, separated from Round 1 by several months to approximately 26 months [5], and are designed to study the long-term longitudinal stability of eye-movement biometrics—a different research question from the

channel-MI baselines established here. The S1–S2 within-round interval is uniform across all Round 1 participants, making it appropriate for the between-session reliability (ICC) analysis reported in Section 5.1.

Units of analysis.

Three counts recur below and are easily conflated, so we fix the terminology here. A *participant* is one of the 407 individuals. A *session* is one of the two visits, S1 and S2, which are separated by roughly 30 min within Round 1. A *recording* is one participant in one session performing one task—the unit at which every channel measure is estimated. With 407 participants, two sessions and four tasks, the maximum is $407 \times 2 = 814$ recordings per task. The exclusion threshold applies to the *state sequence* rather than to the fixation sequence, and the two are not the same length. Amplitude and direction are inter-fixation quantities, so n fixations yield $n - 1$ states intrinsically. Duration and vergence are fixation-level quantities and would yield n; we deliberately drop the final fixation from both so that all four channels describe the same set of fixation pairs and their MI values are computed over identical intervals. Every channel therefore has $n - 1$ states, from which $n - 2$ consecutive transitions are formed.

We require at least 20 states—equivalently at least 21 fixations. The published archives analysed here supply 812 of the 814 possible recordings for each of VID, TEX and RAN—one participant's two sessions are absent for those three tasks—and the length rule excludes none of the 812. VRG elicits far fewer fixations: of the 814 possible VRG recordings, 813 are present, 104 of these contain fewer than 20 fixations, leaving 709 that meet a 20-fixation floor. Of those 709, 15 contain exactly 20 fixations, giving 19 states—one short. Since the rule and the sequence construction are common to all four channels, the same 15 recordings are excluded from each, and 694 recordings are analysed on VRG for every channel alike, giving 3130 recordings in all.

The vergence channel has one further step: positions without usable binocular samples cannot be assigned a state. They are marked as gaps and left in place rather than deleted so that the sequence retains its temporal alignment and no transition is formed across a gap. This affects 12% of recordings, with a median of zero and a maximum of 69 gap positions, but it excludes no recording that the length rule would have retained—the shortest vergence sequence in the analysed set has exactly 20 valid states. The 694 figure is produced by the length rule alone.

Derived analyses use smaller denominators for structural reasons: test–retest reliability needs a valid estimate in both sessions, the cross-channel correlations need all four channels estimable, and the Friedman test needs all four tasks; each N is stated where it appears.

Participant demographics for the full Round 1 sample are: age $M = 20.9$, $SD = 4.1$, range 18–58 years; 216 female, 188 male, 3 other; 185 participants (45%) wore corrective lenses (see Lohr et al. [5] for complete details).

3.2. Task Descriptions

The four tasks analysed here are described in full in Lohr et al. [5]; the key properties relevant to the present analysis are summarised below and illustrated in Figure 2.

A note on terminology.

The word *vergence* appears in this paper in two distinct roles, and they should not be conflated. *VRG* always denotes the task, one of the four stimulus conditions, in which the target alternates between two depths. The *vergence channel* always denotes the gaze attribute we compute from binocular disparity. The two are independent: every channel is computed for every task, so the vergence channel is estimated on VID, TEX and RAN exactly as it

is on VRG, and—as it turns out—it does not attain its largest values on VRG. To keep the distinction visible, we refer to the task only by its abbreviation from this point on.

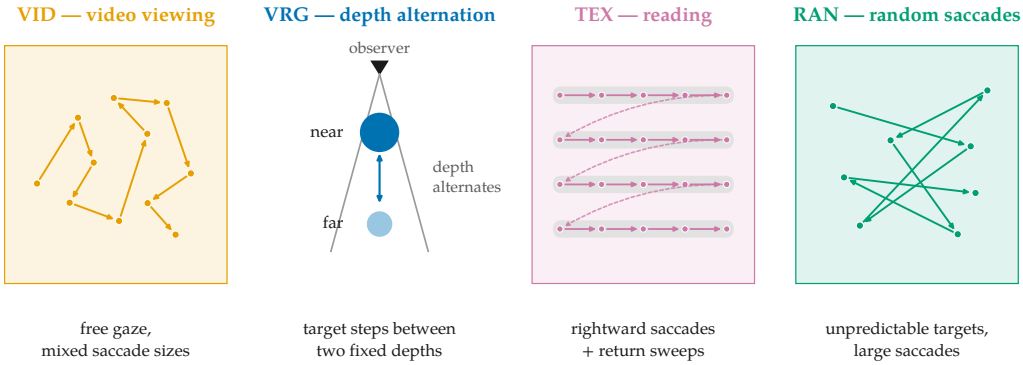


Figure 2. The four GazeBaseVR tasks analysed here. Dots mark fixations, arrows the saccades between them. In TEX, the solid arrows are the rightward saccades along a line of text and the dashed curved arrows the return sweeps to the start of the next line; in VRG, the solid double-headed arrow indicates that the target alternates between the near and far depths. The tasks differ in how strongly they constrain the gaze sequence: VID leaves it free, VRG constrains depth rather than direction, TEX imposes a near-deterministic left-to-right pattern, and RAN imposes large saccades to unpredictable positions.

Depth-Alternation Task (VRG).

A small sphere (1° diameter) is displayed against a random-noise plane and alternates between two depths: approximately 0.44 m and 0.35 m. At these distances, assuming an inter-ocular distance of 62 mm, the ideal vergence angle alternates between 8° and 10° of visual angle. The measured angles reported in Section 4.6 (near ≈ −11°, far ≈ −9°) are the empirical counterparts of these ideal −10° and −8° values. Because the −9° bin boundary coincides with the measured far depth, far-depth VRG fixations may fall on either side of it, which reinforces the exploratory status of the VRG results. The stimulus scales in apparent size to keep the accommodation signal constant, so vergence eye movements are driven by disparity alone. Inter-stimulus intervals are uniformly random between 2 and 3 s, yielding 12 convergent and 12 divergent movements per session (24 movements total) over a duration of 48–72 s. VRG is the only task in GazeBaseVR designed to elicit fixation-depth transitions. It is therefore the condition in which the vergence channel might be expected to register most strongly—an expectation the results do not bear out.

Video Viewing Task (VID).

A 38 s clip from the open-source animated film *Big Buck Bunny* is displayed on a 36 × 21° rectangular plane at 1 m depth; a different clip is shown in each of the two sessions [5]. Participants view the video freely with no task instruction beyond natural viewing. The clip contains a mix of static and moving objects, eliciting a naturalistic variety of fixations and saccades. This task is the most analogous to free-viewing in a VR gaming context and is the primary condition for all channels except vergence.

Reading Task (TEX).

An excerpt of approximately 820 characters from a *National Geographic* article is displayed at 0.6 m depth within a 51.2 × 37.6° viewing region; a different excerpt is used in each of the two sessions [5]. A fixed-width font is used so that each character subtends approximately 1° of the visual angle. Participants read at their own pace and press a controller trigger when finished; task duration ranges from 22 to 141 s (median 52 s). Reading imposes a strongly directional, left-to-right scanpath with periodic return sweeps, providing a

structured baseline for the direction and amplitude channels. Participants press a controller trigger when they have finished, after which a multiple-choice comprehension question is presented. The dataset authors note that the question is not intended to be difficult and that the answer given is discarded; its purpose is to discourage skimming. TEX is therefore the only task in the battery with an explicit comprehension demand.

Random Saccade Task (RAN).

A small sphere (0.5° diameter) jumps to uniformly random positions within ±15° horizontally and ±10° vertically at 1 m depth. Consecutive targets are separated by at least 3°, and inter-stimulus delays are uniformly random between 1 and 1.5 s. Each session comprises 79 stimulus jumps (80 fixation periods) over 80–120 s. The random spatial layout provides an unstructured baseline against which task-driven MI increases can be measured.

3.3. Fixation Detection

Fixations were detected from the centre-gaze coordinates using a self-contained implementation of the I-DT (dispersion-threshold) algorithm [23]. Because I-DT admits several equally common variants that yield materially different segmentations at the same nominal thresholds, the procedure is specified here in full.

Before event detection, missing samples in the horizontal and vertical centre-gaze series were filled by linear interpolation between the nearest valid samples on either side with no limit on gap length; missing samples preceding the first or following the last valid sample were set to that sample's value. No other filtering or smoothing was applied. Appendix A.5 reports how often long gaps occur and repeats the whole analysis with gaps longer than 75 ms left unfilled, which changes no reported ratio by more than 0.01 and no task ordering.

Dispersion was defined following Salvucci and Goldberg as the sum of the two coordinate ranges,

$$D = (\max x - \min x) + (\max y - \min y), \quad (1)$$

so the 1.5° threshold refers specifically to this summed-range measure rather than to the larger of the two ranges or the diagonal extent of the samples, neither of which can exceed the summed range, so at the same nominal value, the summed-range criterion is the strictest of the three. A candidate window beginning at sample i was extended one sample at a time for as long as D remained at or below 1.5°. The duration criterion was applied once to the completed window: extension was never halted by reaching the minimum duration, so accepted fixations are always maximal under the dispersion constraint. A window whose duration reached 80 ms was recorded as a fixation, its centroid taken as the arithmetic mean of its samples and its duration as the interval between its first and last sample timestamps, and detection resumed at the sample that had broken the dispersion criterion. A window falling short of 80 ms was discarded, and the start index advanced by exactly one sample.

For construction of the information channels, the final fixation was omitted from the fixation-level duration and vergence sequences so that these were aligned with the inter-fixation amplitude and direction sequences, and all four channels describe the same fixation pairs. A recording with n detected fixations therefore yields $n - 1$ candidate states in the amplitude, duration and direction channels; in the vergence channel, some of those positions may lack usable binocular samples and therefore carry no state.

Recordings were retained for a channel only when the resulting state sequence contained at least 20 valid states. A run of N consecutive states supplies $N - 1$ adjacent pairs, so this criterion corresponds to a nominal minimum of 19 transition pairs. For amplitude, duration, and direction, the correspondence is exact, because those sequences contain no gaps. For vergence, it is an upper bound rather than an identity: positions without usable

binocular samples are retained as gaps, no pair spanning a gap is counted, and 20 valid states distributed on either side of a gap therefore supply fewer than 19 usable pairs.

We applied the criterion to valid states rather than to usable pairs so that the analysed set remains identical across channels and is fixed before any channel-specific quantity is computed. The empirical consequence is benign: internal gaps cost at least one pair in 11% of retained recordings, a median of two and at most 93, yet the smallest number of usable vergence pairs anywhere in the analysed set is 19. No retained recording therefore falls below the nominal minimum implied by 20 consecutive states, and the state-based criterion turns out to guarantee the pair-based one it was not constructed to enforce.

Event properties (duration, centroid location, amplitude) were computed directly from the detected fixation events. The mean number of fixations per recording was 170 (SD = 86) across all tasks combined. The VRG task yielded substantially fewer fixations (mean 55, SD 37) than the other three tasks (mean 136–241), because the task design elicits only 24 vergence movements (12 convergent, 12 divergent) over 48–72 s [5]. VRG results are therefore treated as exploratory throughout.

3.4. Channel Definitions and Discretisation

Four channels were computed from the fixation sequence of each recording.

Amplitude.

Saccade amplitude was defined as the Euclidean angular distance (°) between the centroids of consecutive fixations. The distribution was partitioned into three fixed categories: short (<1°), medium (1–5°), and long (>5°), following [4], giving $K_{\text{amp}} = 3$ states ordered from shortest to longest saccade.

Duration.

Fixation duration (ms) was partitioned into $K_{\text{dur}} = 3$ fixed categories: short (<150 ms), medium (150–900 ms) and long (≥900 ms), following the same study, which were ordered from shortest to longest dwell time.

Direction.

Saccade direction was defined as the polar angle (0–360°) of the vector from fixation i to fixation $i + 1$, which was computed as $\theta = \arctan_2(\Delta y, \Delta x) \bmod 360°$. The full circle was partitioned into $K_{\text{dir}} = 8$ equal-width octants of 45°.

Vergence depth.

For each fixation, binocular vergence was estimated as the mean horizontal disparity between right and left eye gaze angles during the fixation window:

$$\nu_j = \frac{1}{|W_j|} \sum_{t \in W_j} \left(x_t^R - x_t^L \right), \tag{2}$$

where x_t^R and x_t^L are the right- and left-eye horizontal gaze positions in degrees of visual angle at sample t, and W_j is the set of samples belonging to fixation j.

Because this channel has not been used in the framework before, it is worth setting out what ν_j measures; Figure 3 gives the geometry. When both eyes fixate on a distant object, their optical axes are almost parallel; fixating on a near object requires them to rotate inwards, so the right eye turns left and the left eye turns right. The horizontal difference $x^R - x^L$ therefore becomes more negative the nearer the fixated object: it is the *binocular disparity* of the gaze point, and it is a proxy for fixation depth. Sign convention matters and is fixed once here: ν_j is negative throughout this dataset, more negative means stronger

convergence on a nearer object, and values closer to zero mean divergence towards a farther one. The equivalent distance is

$$z_j = \frac{d_{\mathrm{IPD}}}{2\tan(|\nu_j|\pi/360)}, \tag{3}$$

with $d_{\mathrm{IPD}} = 62\,\mathrm{mm}$ the inter-ocular distance assumed by the dataset authors, though we discretise ν_j directly rather than z_j because the angular quantity is what the tracker measures and the mapping to distance is strongly non-linear at small angles. ν is a simplified horizontal vergence measure computed with an assumed inter-ocular distance of 62 mm; it is not an individualised three-dimensional reconstruction of fixation depth, and z_j is given only to relate angles to nominal distances.

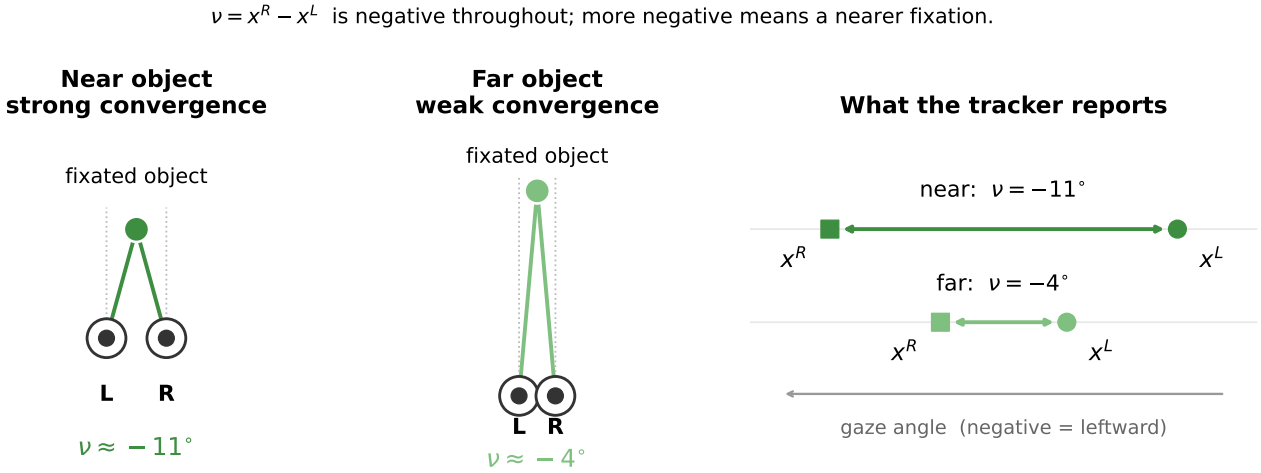


Figure 3. The vergence channel. Fixating a near object rotates both eyes inwards; fixating a far one leaves the optical axes almost parallel. The signed difference between the two reported gaze angles, $\nu = x^R - x^L$, is therefore a proxy for fixation depth, and it is negative throughout this dataset.

The empirical range is approximately -15° to -3°, and the VRG stimulus alternates between two depths whose ideal vergence angles differ by 2°. We partition ν_j into three fixed bins chosen to sit inside that range and to separate the ideal VRG depths: strong convergence ($\nu_j < -9^\circ$, nearer than about 0.39 m), moderate ($-9^\circ \le \nu_j \le -5^\circ$), and weak ($\nu_j > -5^\circ$, beyond about 0.71 m), giving $K_{\mathrm{verg}} = 3$ states. The boundaries are absolute rather than per-participant; Section 4.4 shows that this choice makes part of the between-participant variance in this channel reflect differences in resting vergence rather than differences in gaze behaviour.

A coarser direction scheme.

The eight-octant direction channel requires $8^2 = 64$ transition-matrix cells, which the recording lengths available here sample poorly (Section 4.3). We therefore compute a second, coarser variant in parallel: four equal-width 90° quadrants, reducing the matrix to 16 cells. The quadrant boundaries are placed at the cardinal directions, so the bins are centred on the diagonals, which is the convention the octant scheme already implies. Because the quadrant scheme is a deterministic merging of octant pairs, the data-processing inequality gives $I_4(X;Y) \le I_8(X;Y)$ in the population—the same refinement argument used for the three- and four-category amplitude schemes in [4]. The two are reported side by side throughout: they are alternative discretisations of one signal rather than independent channels. The four-quadrant scheme is the primary one for comparisons of the direction channel between tasks, because the octant transition matrix is the more sparsely sampled at these recording lengths (Section 4.5). The eight-octant scheme is the one used in the

four-channel analyses (Sections 4.2 and 4.5, the octant rows of Table A1 and Appendix A.4, the cross-channel correlations and the $\sigma_{\text{task}}/\sigma_{\text{obs}}$ ratio) and in the comparison with Hao et al. [3], and tables and figures state which scheme they show.

3.5. Markov Chain and Information-Theoretic Measures

For each channel c and recording, let $\mathbf{s} = (s_1, s_2, \ldots, s_N)$ denote the sequence of integer states. Figure 4 summarises the construction and the quantities it yields. The first-order transition matrix $P(Y{=}j \mid X{=}i)$ was estimated from raw consecutive state-pair counts, which were normalised row-wise:

$$\hat{P}(Y{=}j \mid X{=}i) = \frac{n_{ij}}{\sum_k n_{ik}}, \tag{4}$$

where n_{ij} is the observed count of transitions from state i to state j. Rows with zero marginal count are set to the uniform distribution; because $\pi_i = 0$ for any state never visited as a source, these rows carry zero weight in $H(Y|X)$, and the fallback is equivalent to omitting them. The convention $0 \log 0 = 0$ is applied throughout. The following descriptors were then computed:

$$\hat{q}_{ij} = \frac{n_{ij}}{\sum_{kl} n_{kl}}, \qquad \hat{q}_{i\cdot} = \sum_j \hat{q}_{ij}, \qquad \hat{q}_{\cdot j} = \sum_i \hat{q}_{ij} \tag{5}$$

$$H(X) = -\sum_i \hat{q}_{i\cdot} \log_2 \hat{q}_{i\cdot} \tag{6}$$

$$H(Y) = -\sum_j \hat{q}_{\cdot j} \log_2 \hat{q}_{\cdot j} \tag{7}$$

$$H(X,Y) = -\sum_{i,j} \hat{q}_{ij} \log_2 \hat{q}_{ij} \tag{8}$$

$$H(Y|X) = H(X,Y) - H(X) \tag{9}$$

$$I(X;Y) = \sum_{i,j} \hat{q}_{ij} \log_2 \frac{\hat{q}_{ij}}{\hat{q}_{i\cdot}\hat{q}_{\cdot j}} = H(X) + H(Y) - H(X,Y) \tag{10}$$

$$\text{NMI} = \frac{I(X;Y)}{H(X,Y)} \tag{11}$$

where n_{ij} is the observed count of transitions from state i to state j, all logarithms are base 2, and $0 \log_2 0 \equiv 0$.

Equations (5)–(11) estimate every quantity from the same empirical distribution of consecutive state pairs. This differs from the common practice of computing $I(X;Y) = H(X) - H(Y|X)$ with $H(X)$ taken from the marginal state occupancy, which is exact only if the observed occupancy is stationary for the estimated transition matrix. On a finite sequence, it generally is not: the first observation never appears as a destination and the last never as a source, so $\hat{q}_{i\cdot}$ and $\hat{q}_{\cdot j}$ differ, and with them $H(X)$ and $H(Y)$. The discrepancy is negligible for long sequences but not for short ones. In this dataset, it is below 0.01 bits for VID, TEX and RAN, but reaches 0.10 bits for vergence and 0.07 bits for amplitude on VRG, where only ~53 transitions are available and the state distribution is highly concentrated—enough to inflate the VRG estimates materially. We therefore use the pair-distribution form throughout, which requires no stationarity assumption. Under stationarity, the two coincide, so Equation (10) remains consistent with the definition used in [1,4].

State-space collapse control.

A recording with very low $H(X)$ has an effectively reduced state space. Since $I(X;Y) \leq H(X)$, low marginal entropy bounds rather than inflates the mutual information; what it signals is that the estimated sequential dependence concerns only one or two substantially occupied states, and it should not be read as evidence of a rich multi-state

structure. We flag recordings where $H(X) < H_{\min} = 0.5$ bits as collapse candidates and report their frequency by channel and task.

(**a**) A gaze attribute as an information channel (**b**) Shannon quantities

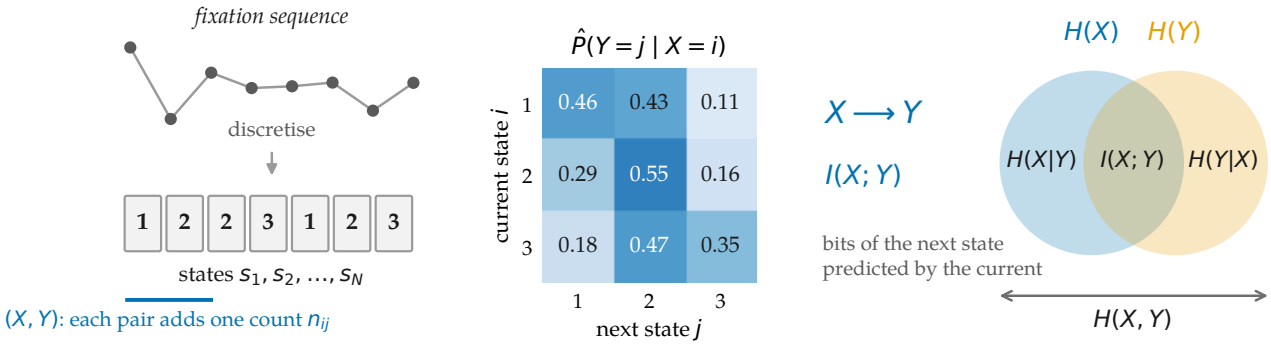


Figure 4. (**a**) How a gaze attribute becomes an information channel. The fixation sequence is reduced to a sequence of discrete states by fixed binning; every consecutive pair of states contributes one count n_{ij}; the row-normalised counts give the channel $\hat{P}(Y{=}j \mid X{=}i)$, whose mutual information measures how many bits of the next state are predicted by the current one. (**b**) The Shannon quantities. $H(X)$ and $H(Y)$ are the entropies of the two marginals, and their overlap is $I(X;Y)$, the non-overlapping parts are the conditional entropies, and the union is the joint entropy $H(X,Y)$. Estimating all of these from the same pair distribution (Equations (5)–(11)) keeps the identities in the diagram exact in the sample. In (**a**), the shading of the matrix cells is proportional to the transition probability (darker blue, higher probability); in (**b**), the blue and yellow circles represent $H(X)$ and $H(Y)$, respectively.

3.6. Finite-Sample Bias Correction

The plug-in estimator $\hat{I}(X;Y)$ of Equation (10) is biased upward when the number of observed transitions is small relative to the number of transition-matrix cells. This is a general property of information estimation from finite samples and is independent of the gaze channel framework; it is treated here because the recording lengths in GazeBaseVR make it material. The bias arises because the sampling noise in $\hat{P}(Y|X)$ makes rows appear more deterministic than they are, deflating $\hat{H}(Y|X)$. This matters here: an eight-state direction channel has 64 cells, and the VRG task supplies roughly 53 transitions per recording—fewer than one per cell.

We estimate the bias non-parametrically, using the shuffling approach established in the spike-train information literature [24,25], where the same undersampling problem arises when the number of trials approaches the number of discrete response states. For each recording, the state sequence is randomly permuted $B = 500$ times and Equation (10) is re-evaluated on each permutation. Permutation preserves the state counts exactly while destroying the original temporal ordering, so the MI recovered from a permuted sequence provides an empirical finite-sample null for the plug-in estimator at that sequence length and marginal. (A permutation of a finite sequence is a draw without replacement, so adjacent positions are not strictly independent; the null is therefore an empirical reference distribution rather than an exactly-zero-MI process.) We define

$$\hat{I}_{\text{null}} = \frac{1}{B}\sum_{b=1}^{B} \hat{I}^{(b)}(X;Y), \qquad \hat{I}_{\text{corr}} = \max\left(0,\ \hat{I} - \hat{I}_{\text{null}}\right), \tag{12}$$

where $\hat{I}^{(b)}(X;Y)$ is Equation (10) evaluated on the b-th permutation. We additionally record the 95th percentile of the null distribution, allowing each recording to be tested against its own null rather than against a pooled threshold. All results below report $\hat{I}_{\text{corr}}$; raw values are given alongside where the distinction is material. The normalised form is built on

the same corrected numerator, NMI = $\hat{I}_{\text{corr}}/H(X,Y)$: Equation (11) defines the ratio, and every NMI reported below takes its numerator from Equation (12) rather than from the plug-in estimator. The distinction is not cosmetic—for the eight-octant direction channel on VRG, the two differ by a factor of eight—so NMI computed from the uncorrected $\hat{I}$ will not reproduce the NMI values reported in Section 4.2.

Both shuffling and analytic corrections are long-established, and Panzeri et al. [24] recommend using them alongside one another. We use the permutation estimate for the correction itself and compute the Panzeri–Treves analytic estimate [26,27] as an independent diagnostic of finite-sample bias. It is never subtracted: $\hat{I}_{\text{corr}}$ is Equation (12) alone, and no recording is corrected twice. The analytic estimate is

$$\widehat{\Delta I}_{\text{PT}} = \frac{(R_x - 1)(R_y - 1)}{2N_{\text{tr}} \ln 2}, \tag{13}$$

where N_{tr} is the number of observed transitions and R_x, R_y are the numbers of occupied rows and columns of the transition-count matrix. Counting occupancy rather than substituting the nominal state count S matters here: vergence occupies on average only 2.4 of its three states and amplitude on VRG only 2.3, so $(S-1)^2$ would overstate the bias for exactly those channels whose occupancy is most concentrated.

The two estimates rest on different principles—a series expansion in N_{tr}^{-1} against a resampling procedure—so their agreement is informative. Across all recordings, they agree to a mean absolute difference of 0.003 bits for the three-state channels (18.6 transitions per cell on average) and 0.009 bits for the four-quadrant direction channel (10.5 per cell), but they diverge to 0.037 bits for the eight-octant channel (2.6 per cell), where the analytic expansion is outside its domain of validity. The divergence therefore tracks sampling adequacy, as expected, and is confined to the one channel already identified as undersampled. We report the shuffling estimate throughout, since it carries no distributional assumption; the analytic values are given alongside in Table A1, Appendix A.1. The permutation correction reduces finite-sample effects but does not guarantee their complete removal, particularly for the short eight-octant sequences, and corrected MI remains associated with recording length in some channel–task cells (Appendix A.7).

Nothing in either procedure is new. What is new is their application to gaze information channels, where channel measures have to date been reported from the uncorrected plug-in estimator.

3.7. Statistical Analysis

The comparison in Section 4.9 uses a random forest (250 trees, default depth) with five-fold cross-validation grouped by participant so that no participant appears in both training and test folds; accuracy is averaged over folds.

Cross-channel correlations are computed on one value per participant and task: the mean across that participant's available valid sessions, which is both sessions where both are retained and the single session otherwise. Under VRG, 317 of the 377 contributing participants have two sessions and 60 have one. Averaging is preferred to selecting a single session because it is fully specified and uses every retained recording; where two sessions are averaged, it also reduces the variance of an estimate whose between-session reliability is modest (Section 5.1), though no such reduction applies to the single-session cases.

Because finite-sample bias scales with fixation count, and that count is shared by all four channels of a recording, we additionally report partial Spearman correlations controlling for the participant's mean fixation count. These rank each participant-level variable and then take the first-order partial correlation of the two channel ranks given the ranked fixation count; significance uses the t approximation on $N-3$ degrees of freedom, which is Bonferroni-corrected over the six channel pairs within a task. The partial values are the

ones we interpret; the zero-order values are given beside them so the size of the adjustment is visible.

Comparisons across tasks used the Friedman test rather than Kruskal–Wallis. Each participant performs all four tasks within a session, so the four samples are paired and the independence assumption of Kruskal–Wallis does not hold; Friedman is its repeated-measures analogue. It is applied to participants with valid data in all four tasks, each contributing the mean of their available valid sessions—the same aggregation used for the cross-channel correlations (Section 4.11). A participant's two sessions are not independent, so using participant sessions as blocks would violate the independence the test assumes between blocks; averaging first removes that problem at the cost of the within-participant session variation. The effect size is Kendall's coefficient of concordance $W = \chi^2_F/[N(k-1)]$. Pairwise contrasts use the Wilcoxon signed-rank test with Holm correction within each channel, Holm being preferred to Benjamini–Hochberg here because the six contrasts are few and mutually dependent.

Test–retest reliability was quantified with ICC(2,1) (two-way random effects, absolute agreement, single measures) between Sessions 1 and 2 [28]. Associations between participant-level MI and the self-reports of Section 4.8 are Spearman correlations on participant means, which are Bonferroni-corrected over the 32 tests (four channels × four tasks × two ratings). Demographic covariates (Section 4.10) were tested against each participant's mean MI: age by Spearman correlation, self-reported race and ethnicity by Kruskal–Wallis test, and gender and dominant eye by Mann–Whitney U test, which were Bonferroni-corrected over 16 tests (four channels × four covariates). Unless stated otherwise, confidence intervals are 95% percentile bootstrap intervals over 2000 resamples of participants, each resampled participant contributing all of their data.

All tests are two-tailed and non-parametric, as the MI distributions are bounded below at zero and right-skewed. Analyses were conducted in Python 3 using scipy.stats and pandas.

4. Results

4.1. Descriptive Statistics

Table 1 summarises fixation counts and raw signal ranges per task.

Table 1. Descriptive statistics per task. N_{rec}: valid recordings (participant × session), which are lower for VRG because it elicits few fixations (~55 per session) and so loses more to the 21-fixation minimum (20 states). $\bar{n}_{\text{fix}}$: mean fixations per recording (SD).

Task	N_{rec}	$\bar{n}_{\text{fix}}$ (SD)	Mean Dur. (ms)	Mean Amp. (°)
VID	812	136 (20)	256	3.93
VRG	694	55 (37)	1419	0.81
TEX	812	232 (60)	188	6.86
RAN	812	241 (41)	393	5.13

4.2. Channel Mutual Information

Table 2 presents the mean (SD) bias-corrected MI and NMI for each channel × task combination pooled across participants. Figure 5 shows the raw values, the permutation null, and the corrected difference side by side.

Table 2. Mean (SD) bias-corrected mutual information $\hat{I}_{corr}$ (bits) and its normalised form per channel and task with 95% percentile bootstrap intervals over 2000 resamples of participants (each resampled participant contributes all their recordings, preserving the repeated-measures structure). Bold: highest value per task. Raw values and the permutation null are in Table A1. Direction is the eight-octant scheme (Section 3.4).

Channel	Metric	Task			
		VID	VRG	TEX	RAN
Amplitude	MI	0.087 (0.059)	0.033 (0.051)	0.052 (0.034)	**0.304 (0.136)**
	95% CI	[0.083, 0.092]	[0.029, 0.037]	[0.049, 0.055]	[0.292, 0.316]
	NMI	0.030 (0.021)	0.040 (0.082)	0.018 (0.012)	**0.113 (0.055)**
	95% CI	[0.028, 0.031]	[0.034, 0.047]	[0.017, 0.019]	[0.108, 0.118]
Duration	MI	0.009 (0.015)	0.045 (0.069)	0.004 (0.008)	0.057 (0.049)
	95% CI	[0.008, 0.010]	[0.040, 0.051]	[0.004, 0.005]	[0.053, 0.062]
	NMI	0.004 (0.007)	0.020 (0.034)	0.002 (0.005)	0.021 (0.017)
	95% CI	[0.004, 0.005]	[0.018, 0.023]	[0.002, 0.003]	[0.019, 0.022]
Direction	MI	0.074 (0.065)	**0.115 (0.111)**	0.076 (0.051)	0.126 (0.076)
	95% CI	[0.070, 0.079]	[0.106, 0.124]	[0.073, 0.080]	[0.120, 0.133]
	NMI	0.014 (0.013)	0.026 (0.026)	0.018 (0.012)	0.023 (0.014)
	95% CI	[0.013, 0.015]	[0.024, 0.028]	[0.017, 0.019]	[0.022, 0.024]
Vergence	MI	**0.115 (0.095)**	0.107 (0.143)	**0.129 (0.125)**	0.106 (0.075)
	95% CI	[0.108, 0.122]	[0.096, 0.120]	[0.119, 0.140]	[0.100, 0.112]
	NMI	**0.090 (0.071)**	**0.150 (0.166)**	**0.094 (0.078)**	0.075 (0.044)
	95% CI	[0.085, 0.095]	[0.136, 0.163]	[0.088, 0.101]	[0.072, 0.078]

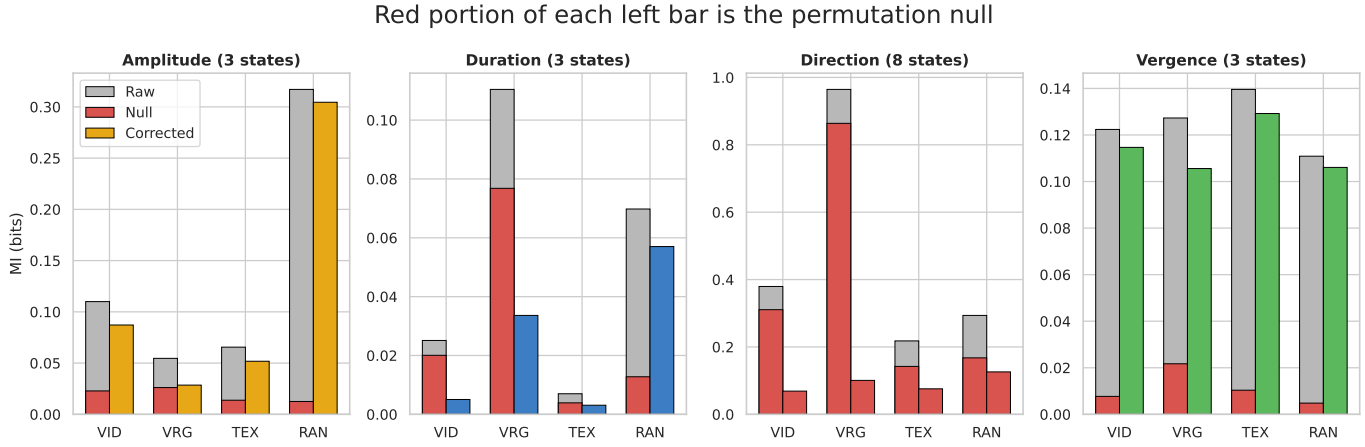


Figure 5. Raw MI, permutation null and corrected MI by channel and task. The red portion of each left-hand bar is the null, the MI of sequences with identical marginals but no temporal structure: a thin sliver for the three-state channels and most of the raw value for the eight-state direction channel. The right-hand bar of each pair is the corrected MI, drawn in that channel's colour: orange for amplitude, blue for duration, red for direction and green for vergence. In the legend of the first panel, "Raw" and "Null" apply to all four panels, while the colour shown for "Corrected" is the amplitude one. Direction is the eight-octant scheme (Section 3.4).

4.3. Magnitude of the Estimator Bias

Figure 5 shows the raw MI, the permutation null and the corrected difference for every channel and task. Table A1, Appendix A.1 gives the numerical values with the analytic Panzeri–Treves estimate, the proportion of the raw value retained, the transitions per transition-matrix cell and the per-recording significance rate.

Sampling adequacy, not channel identity, drives the pattern. Amplitude and vergence retain 79–96% of their raw MI wherever at least ~15 transitions per cell are available. Direction, with 64 cells, never exceeds 3.7 transitions per cell and retains only 12–43%; on

VRG, with fewer than one transition per cell, 88% of the raw value is bias. Testing each recording against its own null agrees: 57–94% of vergence recordings and 100% of amplitude recordings on RAN exceed their null against 28–36% of direction recordings on VID and VRG. Duration is affected structurally instead: on VID and TEX, it retains 35% and 62% because its true MI is near zero there, and only 11% and 14% of recordings clear their null.

4.4. Pooled Estimates and the Level of Analysis

Every value above comes from a single recording of 53–239 transitions. Summing transition counts within a task would instead give 10^4–10^5 transitions and reduce the bias of Equation (13) to a negligible 10^{-5}–10^{-3} bits, but a pooled matrix yields one value per channel and task, where the reliability analysis needs one per participant per session, the cross-channel correlations one per participant, and the task comparisons a distribution: the bias is the price of estimating at the level the questions demand. The comparison is still informative (Table A2, Appendix A.2). The kinematic channels agree within a factor of about two (ratios 0.49–1.49), so the bias-free aggregate and the bias-corrected individual estimate reach comparable magnitudes by different routes. Vergence pools three to five times higher, and the discrepancy diagnoses the baseline problem: pooled $H(X)$ exceeds the mean individual $H(X)$ by a factor of 1.45–2.02 for vergence against 1.01–1.24 for the others, because the bins are absolute while resting vergence differs between participants, so the pooled matrix fills with each participant's self-transitions in their own preferred state. The same mechanism explains the vergence collapse rates of Section 4.7.

4.5. Do the Channels Separate Tasks or People?

The companion study [4] characterised each channel by whether its MI varied more across stimuli or across observers, computing one value per observer pooled over paintings and one per painting pooled over observers. The same question can be asked here with tasks in place of paintings. For each channel, we take the standard deviation across the 407 participants of their mean bias-corrected MI (σ_{obs}) and the standard deviation across the four task means (σ_{task}); both are reported in Table 3. A ratio above one means the task means are more spread than the participant means; below one indicates the reverse.

Table 3. Whether each channel varies more across tasks or across participants. σ_{obs}: SD across the 407 participants of their mean bias-corrected MI. σ_{task}: SD across the four task means. Direction is the eight-octant scheme (Section 3.4).

Channel	σ_{obs}	σ_{task}	Ratio	Larger Spread
Amplitude	0.0411	0.1255	3.057	task means
Duration	0.0174	0.0264	1.520	task means
Direction	0.0362	0.0267	0.737	participant means
Vergence	0.0625	0.0107	0.172	participant means

The ratio compares two spreads; it does not say the channel distinguishes tasks or people, which Section 4.9 addresses directly. Two properties limit how far it can be read. σ_{task} is the spread of four numbers and is itself imprecise, and σ_{obs} carries measurement error as well as genuine between-participant differences, which is an error that Section 5.1 shows to be substantial. The ratio is descriptive of this dataset rather than a population variance-component ratio. Read that way, the comparison separates the extremes clearly. The spread of amplitude's four task means is about three times the spread of its participant means. Vergence runs the other way by about a factor of six, its per-participant means spanning 0.002 to 0.695 bits while the four task means lie within 0.023 bits of each other.

Duration and direction fall between these extremes and separate only weakly in opposite directions: duration leans towards tasks (1.52) and direction towards participants (0.74).

One caveat is specific to vergence. Because its bins are absolute, part of its between-participant spread reflects resting vergence (Section 4.4), and part reflects measurement error, so the ratio should not be read as a measure of behavioural individual differences. We tested this directly, recomputing the channel under five alternatives to the boundaries of Section 3.4: values centred on each participant's median vergence, per-participant terciles, per-recording terciles, and two other absolute boundary pairs (Appendix A.6). σ_{obs} stays between 0.058 and 0.066 across all six discretisations, but the ratio does not, because it is driven by σ_{task}, which is small under the $-9°/-5°$ boundaries (0.0107) and about 1.7 to 4 times larger under every alternative discretisation, so the ratio rises from 0.17 to between 0.32 and 0.75. Vergence varies more across participants than across tasks under all six discretisations, but by a factor of between about 1.3 and 5.8, and the $-9°/-5°$ boundaries used here give the extreme of that range. Characterising individuals with the channel would need the baseline-corrected form.

Octants versus quadrants.

Table A1 in Appendix A.1 reports both direction schemes. Halving the angular resolution quarters the transition matrix and roughly doubles the proportion of MI surviving correction, from 12–43% to 37–76% by task, with sampling adequacy rising from 0.8–3.7 transitions per cell to 3.3–14.9. Two observations support the correction itself: on VRG, the schemes carry very different bias (0.864 against 0.200 bits) yet return almost the same corrected MI (0.115 against 0.113), and the corrected quadrant value lies below the corrected octant value in all four tasks, which is consistent with the data-processing inequality for the underlying population mutual information. Neither was imposed on the estimator.

The two schemes do not rank the tasks identically. With octants, the order is RAN > VRG > TEX > VID, with quadrants VRG and RAN exchanging places (VRG > RAN > TEX > VID), and Kendall's W is larger for quadrants (0.239 against 0.113, $N = 376$ participants). Of the six pairwise contrasts, four are significant with the same sign under both schemes (Holm-adjusted Wilcoxon): the direction of MI is higher on VRG and on RAN than on either VID or TEX. The two that differ are VID versus TEX, which are significant only with quadrants, and VRG versus RAN, which are significant only with quadrants and in the opposite direction to the octant point estimate. We therefore rest the interpretation on the four contrasts that hold under both schemes and draw no conclusion from the VRG–RAN ordering. Appendix A.7 gives both sets of tests.

Because the octant transition matrix is the more sparsely sampled (0.87 transitions per occupied cell on VRG against 3.3 for quadrants), quadrants are the primary scheme for comparing the direction channel between tasks, as set out in Section 3.4. The four-channel analyses retain octants, for comparability with the published direction channel [3], and the task orderings that differ between the two schemes are reported as such rather than resolved.

Relation to gaze entropy measures.

The two quantities most reported in the gaze-entropy literature [9,10] are stationary gaze entropy, $H(X)$, measuring how evenly attention is distributed over the states, and gaze transition entropy, $H(Y|X)$, measuring how unpredictable the transitions are. Both follow from the channel (Equations (6) and (9)) with $I(X;Y) = H(X) + H(Y) - H(X,Y)$ relating them: a channel carries information to the extent that its transition entropy falls below its marginal entropy. Reporting both makes our values comparable with that literature and with multiscale-entropy analyses of this dataset [19].

Normalising $H(Y|X)$ by $\log_2 S$ puts the four alphabets on one scale. So normalised, transition entropy is high for the kinematic channels (amplitude 0.77–0.91 except on VRG, direction 0.61–0.89, duration 0.60–0.82) and markedly lower for vergence (0.20–0.41). Vergence transitions are therefore far more predictable than kinematic ones—the same fact its high self-transition rate expresses. The lower transition entropy is also associated with greater retention after bias correction, although the relative robustness depends on both transition structure and effective state occupancy. The lowest values overall are VRG on amplitude (0.34) and VRG on vergence (0.20), both reflecting the stereotypy the depth-alternation stimulus imposes. Because $I(X;Y) \leq H(X)$, stereotypy lowers MI when it collapses occupancy onto few states, as it does here; it need not do so when persistent transitions occur among states that all remain occupied.

Channel rank ordering.

Pooled across all four tasks, the rank ordering of bias-corrected MI follows: amplitude (0.122) > vergence (0.115) > direction (0.097) > duration (0.028). The uncorrected ordering was direction (0.445) > amplitude (0.140) > vergence (0.125) > duration (0.051); the inversion is attributable to the direction channel's estimator bias (Section 4.3), which is the clearest illustration of why the correction is needed. Per-task rankings vary (Table 2 and Figure 6): vergence is highest on VID and TEX, direction on VRG, and amplitude is highest on RAN. In normalised terms (NMI), which expresses MI relative to the joint entropy, vergence is highest on VID, VRG and TEX and amplitude on RAN. The pooled ranking uses weighted means across all valid recordings rather than unweighted averages of the four task means.

Neither ordering matches the screen-based hierarchy reported in the companion study [4] (AOI ≫ pupil > direction > amplitude ≈ duration). The difference is one of task and modality rather than estimation: the bias depends on the ratio of transition-matrix cells to observed transitions, and the two studies sit in very different regimes on that axis (Section 5.2).

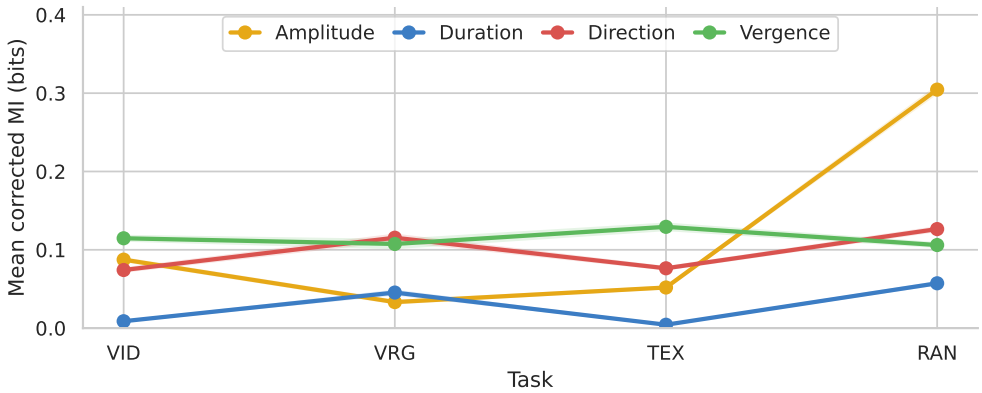


Figure 6. Mean bias-corrected MI per channel across tasks (shading: ±1 SEM). Vergence leads on VID and TEX, direction on VRG, and amplitude on RAN, where the task by construction imposes varied saccade magnitudes. Vergence is the flattest of the four, spanning only 0.106–0.129 bits. Direction is the eight-octant scheme (Section 3.4).

Task effects.

Each participant performs all four tasks within a session, so the task samples are paired rather than independent. We therefore test task effects with the Friedman test on the $N = 376$ participants with valid data in all four tasks, each contributing the mean of their available valid sessions, and report Kendall's W as the effect size (Table 4). All four channels show a significant task effect (all $p < 0.001$), but the magnitudes separate sharply:

amplitude ($W = 0.698$) and duration ($W = 0.432$) are strongly task-modulated, direction is modest ($W = 0.239$ with quadrants, the primary scheme for task comparisons; 0.113 with octants), and vergence is small ($W = 0.017$).

Pairwise contrasts use the Wilcoxon signed-rank test with Holm correction within each channel. For amplitude, all six contrasts are significant, RAN exceeding every other task by a median of 0.19–0.26 bits (95% CI widths below 0.032; matched-pairs rank-biserial correlation 0.99–1.00), which is consistent with a task that imposes varied saccade magnitudes by construction. VRG yields the lowest amplitude MI (0.033 bits) and TEX the lowest duration MI (0.004 bits), the latter reflecting the stereotyped scan path of reading. For vergence, three of six contrasts reach significance with median differences an order of magnitude smaller (0.006–0.023 bits, rank-biserial 0.18–0.22): VID differs from neither TEX ($p = 0.22$) nor RAN ($p = 0.24$), and VRG not from RAN ($p = 0.11$). Vergence is statistically task-sensitive but practically stable, spanning only 0.106–0.129 bits.

Table 4. Friedman test for task effect on bias-corrected MI per channel, over the $N = 376$ participants with valid data in all four tasks, each contributing the mean of their available valid sessions. Friedman replaces Kruskal–Wallis because the task samples are paired, and averaging the sessions first keeps the blocks independent. Effect size is Kendall's $W = \chi^2/[N(k-1)]$ with $k = 4$ tasks with a 95% percentile bootstrap interval over 2000 resamples of participants; W ranges from 0 (no agreement in task ranking across participants) to 1 (complete agreement). Both direction schemes are listed; the four-quadrant scheme is the primary one for comparisons between tasks and the eight-octant scheme the one used in the four-channel analyses (Section 3.4).

Channel	χ^2_F	p	Kendall's W	95% CI
Amplitude	787.0	<0.001	0.698	[0.665, 0.730]
Duration	487.8	<0.001	0.432	[0.387, 0.483]
Direction (8-oct)	127.1	<0.001	0.113	[0.083, 0.151]
Direction (4-quad)	269.9	<0.001	0.239	[0.194, 0.293]
Vergence	18.8	<0.001	0.017	[0.005, 0.038]

4.6. The Vergence Channel

The vergence channel is the novel contribution of this study, and after bias correction, it is also the most robust of the four to finite-sample bias. Figure 7 shows the distribution of per-participant corrected MI across channels and tasks.

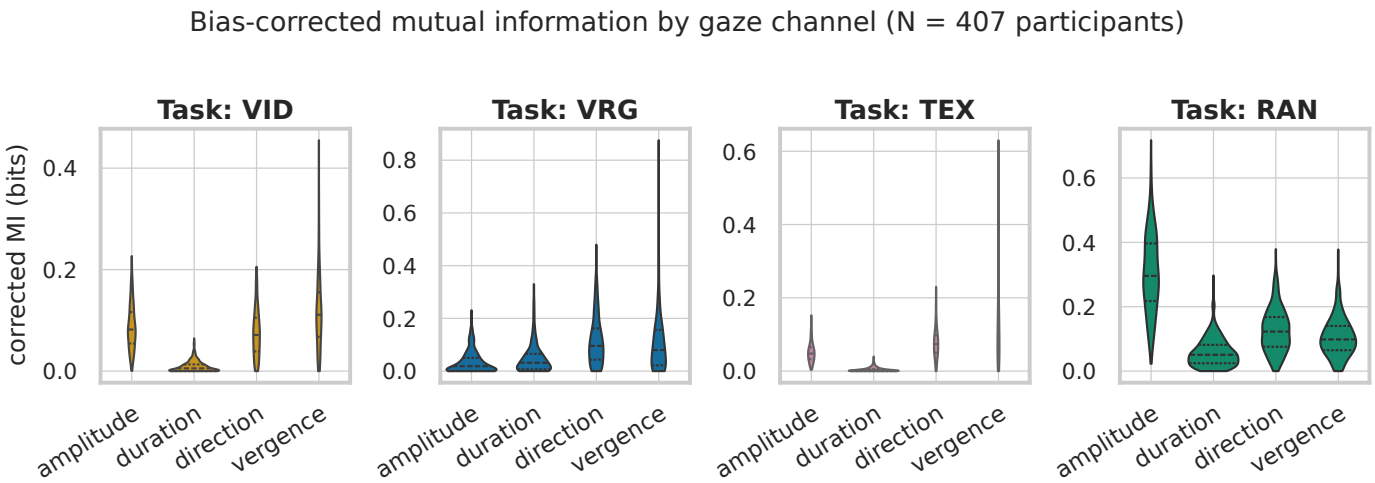


Figure 7. Distribution of bias-corrected mutual information across channels and tasks. Each violin shows the kernel density estimate of per-participant MI; horizontal lines mark quartiles. Direction is the eight-octant scheme (Section 3.4).

The corrected vergence MI is highest on TEX (0.129 bits, SD = 0.125), which is followed by VID (0.115), VRG (0.107) and RAN (0.106). The task effect is statistically significant but small (Friedman $\chi^2_F = 18.8$, $p < 0.001$, $W = 0.017$), and vergence is by this measure the most task-invariant of the four channels. In normalised terms, it is the highest-scoring channel on three of the four tasks (Table 2), the fairer comparison against direction, whose eight states inflate the joint entropy. Against amplitude and duration, it is not neutral: all three have three states, and dividing by $H(X, Y)$ favours whichever has the most concentrated occupancy, which is vergence.

Its robustness to estimator bias matters more: vergence retains 84–96% of its raw MI in every task, more than any other channel, and 57–94% of recordings exceed the 95th percentile of their own permutation null. The channel introduced here is thus the one least vulnerable to the estimation problem that dominates direction.

The vergence signal is $v_j = x^R - x^L$ (Equation (2)), negative throughout and ranging from approximately $-15°$ to $-3°$ across participants, with more negative values indicating stronger convergence at nearer distances. The VRG task alternates between two fixed depths (near $\approx -11°$, far $\approx -9°$). That this scripted alternation does not produce the highest vergence MI is initially surprising. Sequence length does not account for it: VRG supplies only ~55 fixations, but the correction reduces the length effect: vergence retains 84% of its raw MI on VRG, and on VRG, its corrected MI is not reliably associated with the number of fixations ($\rho = 0.07$, 95% cluster-bootstrap interval $[-0.01, 0.16]$; Appendix A.7). An association survives correction in ten of the sixteen channel–task cells (per-cell intervals, unadjusted; six after Bonferroni adjustment), so the correction reduces the dependence on length without removing it. One possible contributor is the discretisation: the far depth lies at $\approx -9°$ on a bin boundary, so the bins may resolve the alternation only partly. Showing that vergence angle follows target distance does not establish that the binning caused the lower MI. In VRG, the per-fixation vergence correlates with the target depth at $\rho = 0.67$, and fixations at the near target are on average 1.56° more convergent than those at the far one, against the 2° nominal separation. The low MI is therefore not a failure of the measure to register depth. The reason for the low value is not settled. Under none of the discretisations of Appendix A.6 does VRG have the highest vergence MI, which argues against the particular boundaries as the explanation, and a detector that also segments on vergence returns about three times as many fixations on the same recordings (Appendix A.5), so under-segmentation of the depth alternation is a second possible contributor. We treat the VRG results as exploratory throughout.

4.7. State-Space Collapse Diagnostic

Figure 8 plots $H(X)$ against the bias-corrected MI for each channel and task. Collapse candidates ($H(X) < 0.5$ bits) account for 1196 of the 12,520 channel–recording pairs formed by the four principal channels (9.6%), and they fall into two groups. Amplitude on VRG: 224 of 694 recordings (32%), because VRG saccades are nearly all <1° and genuinely concentrated in the short bin. Vergence on every task: 19% on RAN, 26% on VID, 29% on TEX and 54% on VRG, because vergence varies over a narrower range than the kinematic variables. Both patterns reflect a real concentration of state occupancy rather than a discretisation artefact, so we retain the recordings, reading their MI as describing transitions among a reduced set of occupied states. One further recording (duration on VRG) falls below the threshold.

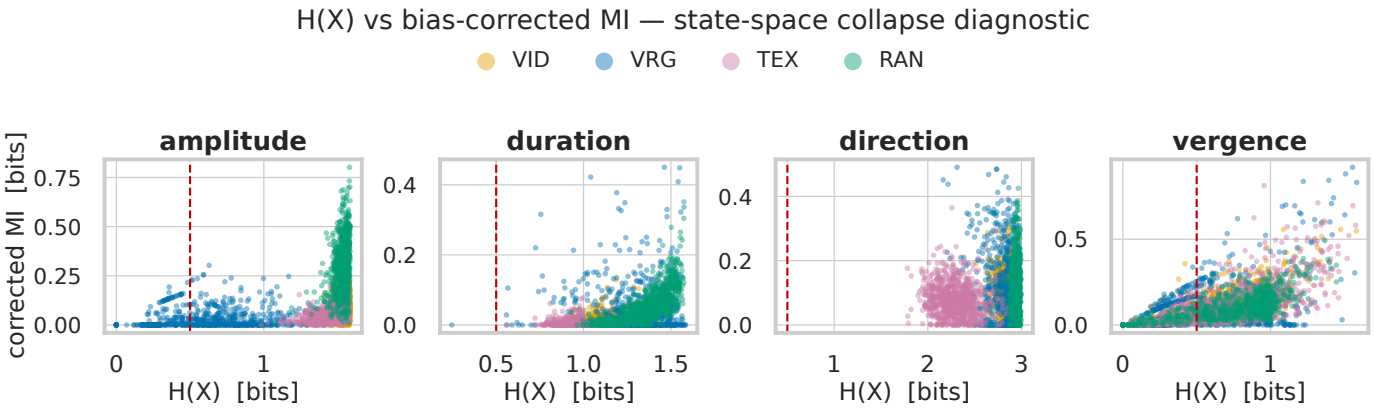


Figure 8. Source-marginal entropy $H(X)$ versus bias-corrected MI for all channels and tasks. Recordings to the left of the red vertical dashed line at $H(X) = 0.5$ bits are collapse candidates: their state space is effectively reduced, so the MI describes transitions among one or two occupied states. They concentrate in amplitude on VRG and in vergence across all tasks (Section 4.7).

4.8. Do the Channels Track Self-Reported State?

The dataset includes per-session self-reports that bear on the interpretation of the channels [5]. Participants rated sleepiness on a seven-point scale before, between and after the two sessions, and they also rated the mental effort exerted during each session on a five-point scale. We therefore tested each channel's bias-corrected MI against the reported state, assigning the "between-sessions" effort rating to Session 1 and the "after-sessions" rating to Session 2, and sleepiness at the start of each session correspondingly. Round 1 supplies these ratings for 405 of the 407 participants.

Both the ratings and the channel measures are then averaged to one value per participant before correlating. A participant's two sessions are not independent observations, and correlating at the session level would treat them as though they were; the unit of analysis is therefore the participant, $N = 405$ (376 on VRG, where fewer participants have a retained recording). The cost is that averaging discards whatever within-person variation exists between the two sessions.

No association survives correction for the 32 tests performed (Bonferroni threshold $p < 0.0016$); the smallest p obtained is 0.003. The two largest coefficients are vergence against mental effort on TEX ($\rho = +0.146$, 95% CI $[+0.054, +0.238]$, $p = 0.003$) and amplitude against sleepiness on RAN ($\rho = -0.131$, $[-0.234, -0.029]$, $p = 0.008$). Eight of the 32 exceed $|\rho| = 0.08$ and none reaches 0.15. Hours slept and caffeine consumption likewise show no relationship with any channel (all $|\rho| \leq 0.10$); these two were tested separately and are not part of the family of 32. Intervals here and throughout are percentile bootstrap intervals over 2000 resamples of participants.

The null should be read with the distribution of the ratings in view. Sleepiness before the sessions is 1 or 2 for 69% of participants (mean 2.19 of 7) and mental effort is 1 or 2 for 65% (mean 2.18 of 5); a single participant reported the maximum sleepiness. Effort changes between the two sessions for only 32% of participants and sleepiness for 38%, so most of the sample contributes no within-person variation at all. The channels therefore fail to detect state differences across a young, rested sample performing short and undemanding tasks, which is weaker than saying they cannot detect state; testing that needs a protocol that manipulates load. The scope is narrow in another way too: effort and alertness are not measures of understanding or performance, and the reading task's comprehension responses were not retained in the released data [5].

4.9. What the Channels Do Not Measure

Channel MI is not proposed as a way of identifying which task a recording came from. To show what it is and is not good for, we trained a random forest to classify the four tasks from each recording, using participant-disjoint five-fold cross-validation, and compared two feature sets: the four bias-corrected channel MI values and three elementary descriptive statistics—mean saccade amplitude, mean fixation duration, and the number of fixations. Against a chance level of 25%, the channel measures reach 68.4% accuracy (SD across the five folds 0.9); the three descriptive statistics reach 99.0% (0.5), and adding the channel measures to them changes nothing (98.9%, 0.6).

The gap is unsurprising once the quantities are distinguished. Mean amplitude records how large the saccades were; the amplitude channel records how much the size of one saccade constrains the size of the next. A task that fixes saccade magnitude by construction is trivially identifiable from magnitude itself, and no sequential measure competes with a direct one on that question. We report the comparison because an information-theoretic measure that correlates with task is easily mistaken for a better task descriptor. Its interest lies elsewhere: vergence, whose mean MI is nearly constant across tasks, has a wider spread of participant means than of task means in its sequential structure by a factor that ranges from about 1.3 to 5.8 with the discretisation (Section 4.5).

4.10. Demographic Covariates

This is a secondary, exploratory analysis, which is reported because one of its results bears directly on the interpretation of the vergence channel, whose absolute bin boundaries may interact with anatomy. The dataset supplies participant details expressly so that questions of this kind can be examined [5]. We tested each channel's mean bias-corrected MI against age, gender, self-reported race and ethnicity, and dominant eye over the 407 Round 1 participants (216 female, 188 male, 3 other; median age 20, interquartile range 18–22, range 18–58). Sixteen tests were performed, giving a Bonferroni threshold of $p < 0.0031$. Gender was compared between female and male participants; the three participants who reported "other" are too few to form a group and were not included in that comparison.

Race and ethnicity and dominant eye show nothing. Across the four groups with $n \geq 20$ (Hispanic or Latino 148, White 140, Mixed 62, Black or African American 41), no channel differs (Kruskal–Wallis, all $p > 0.03$, none surviving correction), and left- against right-eye dominance differs on no channel (all $p > 0.16$).

Three differences by gender survive correction, and their interpretation differs. Amplitude MI is higher for male participants (0.133 against 0.116; difference +0.017, 95% CI [+0.009, +0.025], $p = 0.0002$) as is duration (0.031 against 0.026; +0.005, [+0.002, +0.008], $p = 0.002$); we can offer no account of these and do not propose one. Vergence runs the other way, higher for female participants (0.126 against 0.103; +0.023, [+0.012, +0.035], $p = 0.0003$), and its source is uncertain. The stationary entropy $H(X)$ of the vergence channel separates the groups more sharply than the MI does (0.723 against 0.644; +0.078, [+0.046, +0.111], $p = 3 \times 10^{-6}$), meaning the states themselves are occupied differently. One possible explanation is that because the vergence bins are fixed absolute angles, a systematic group difference in inter-ocular distance would place $v_j = x^R - x^L$ differently relative to those boundaries so that occupancy would shift and MI would follow. Inter-ocular distance was not measured here, so this explanation cannot be tested, and behavioural differences and other measurement-related sources cannot be excluded: the differences are observed, and their origin remains unknown. The result nonetheless strengthens the case for centring vergence on each participant before discretisation as a precaution against boundaries that could register group differences other than differences in gaze behaviour.

Age correlates weakly with direction MI ($\rho = +0.163$, $p = 0.001$), rising monotonically across age bands from 0.094 for participants aged 18–20 to 0.123 for those over 35. The relation persists within the 22-or-younger majority ($\rho = +0.117$, $p = 0.034$), so it is not produced solely by the older tail, but 81% of the sample is 22 or younger and only 18 participants are over 30. This age range cannot support an inference about ageing and gaze; we report the association for the record.

4.11. Cross-Channel Associations

Table A3 reports Spearman rank correlations between all six channel pairs within each task, on bias-corrected MI across the participants with all four channels estimable (406 for VID, TEX and RAN; 377 for VRG), alongside the same correlations partialled on mean fixation count. Bonferroni correction within each task sets the threshold at $p < 0.0083$. The correction family is the six channel pairs within a task. Were it taken instead over all 24 contrasts, the threshold would fall to $p < 0.0021$, and two of the eight associations that survive within-task correction would no longer be significant: direction–vergence on VID ($\rho \mid n_{\text{fix}} = +0.14$, $p \approx 0.004$) and direction–vergence on TEX (-0.15, $p \approx 0.003$). The remaining six survive either family. These two p-values are approximate: recovered from the partial coefficient and N rather than recomputed from the permutation distribution. Full values are in Table A3, Appendix A.3; Figure 9 shows the matrices.

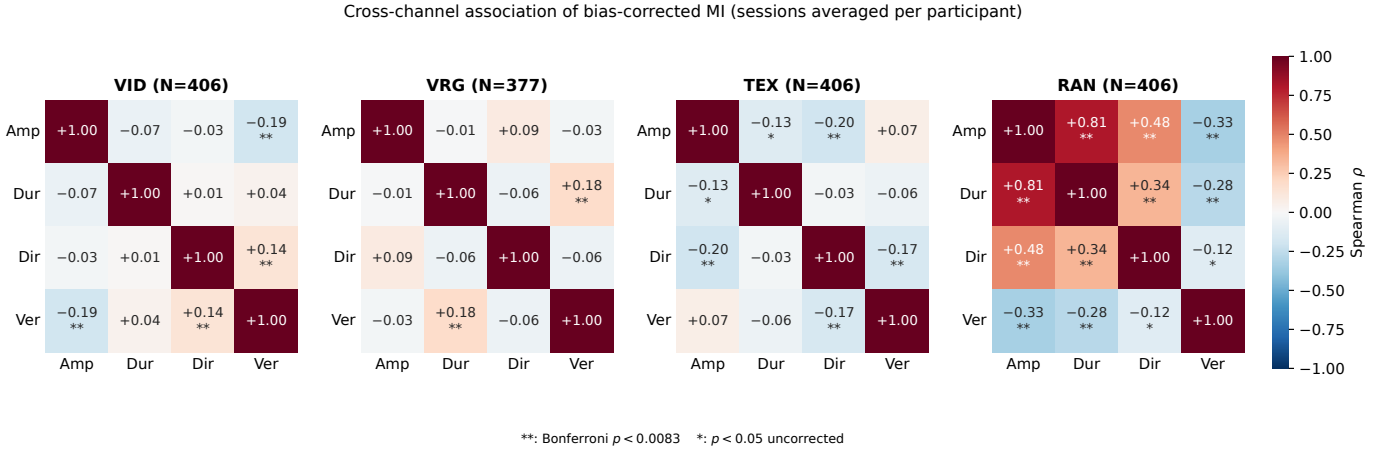


Figure 9. Cross-channel Spearman ρ matrices of bias-corrected MI, one per task (red = positive, blue = negative). Values are zero-order; partial correlations controlling for fixation count are in Table A3. RAN shows a positive kinematic cluster that survives the length control and negative vergence associations that do not; VID two weak Bonferroni-corrected vergence associations, TEX several weak negative ones, VRG one weak positive one. **: Bonferroni-corrected $p < 0.0083$; *: $p < 0.05$ uncorrected. Direction is the eight-octant scheme (Section 3.4).

Bias correction matters here. On raw MI, both RAN and VRG show all six pairs significant; after correction, VRG retains one. Estimator bias scales with fixation count, shared by all four channels within a recording, so short sequences manufacture positive correlation, and VRG is the shortest task.

Two patterns survive the length control. Under RAN, the three kinematic channels form a correlated cluster ($\rho \mid n_{\text{fix}} = +0.31$ to $+0.55$); since these are correlations between MI values, the reading is not that some observers make larger saccades but that participants with stronger sequential dependence in one kinematic channel tend to show it in the others. Under free viewing, the associations are weak but not absent: amplitude and vergence are opposed (-0.19) and direction and vergence weakly aligned ($+0.14$). TEX shows two weak negative associations surviving correction, amplitude–direction and direction–vergence

with amplitude–duration significant only when it is uncorrected; VRG retains one weak positive duration–vergence association (+0.19).

We did not test whether the coefficients differ between tasks; the comparison below is descriptive. Two apparent patterns weaken under the control. The opposition of vergence to amplitude and duration under RAN falls from −0.33 and −0.28 to −0.12 and −0.05 once fixation count is partialled out: both remain negative, but neither is significant after adjustment and Bonferroni correction, and little of either is independent of recording length. Whether that reflects estimator bias, behaviour that co-varies with recording length, or both is not settled by this adjustment. And the amplitude–duration correlation under RAN, the largest in the study, falls from +0.81 to +0.55. A substantial association remains after adjustment; the reduction measures how much of the zero-order value co-varies with recording length, which the two estimates share. Conditioning on fixation count is a sensitivity analysis rather than a causal decomposition—fixation count is itself partly behavioural.

These are correlations between quantities that are themselves estimated with error, and Section 5.1 shows that error to be substantial: single-session reliability is poor for every channel under VID and VRG. Unreliability in the correlated variables attenuates observed associations, so the near-null results under those two tasks are weak evidence of absence rather than evidence of independence. We do not convert the reliabilities into numerical bounds on the correlations: the coefficients in Table A4 are single-measure values for per-session estimates while the correlations use participant means, the classical attenuation relation is defined for Pearson correlations under assumptions we have not established, and several channels have negative point estimates for which the square root is undefined.

5. Discussion

5.1. Between-Session Reliability of Channel MI

ICC(2,1) between Session 1 and Session 2 quantifies the share of total variance attributable to stable between-participant differences (two-way random effects, absolute agreement, single measures; benchmarks following Koo and Li [28]). Values for every channel and task are given in Table A4, Appendix A.4, and shown in Figure 10. *N* counts participants with a valid estimate in both sessions: 406 for VID, TEX and RAN, 317 for VRG.

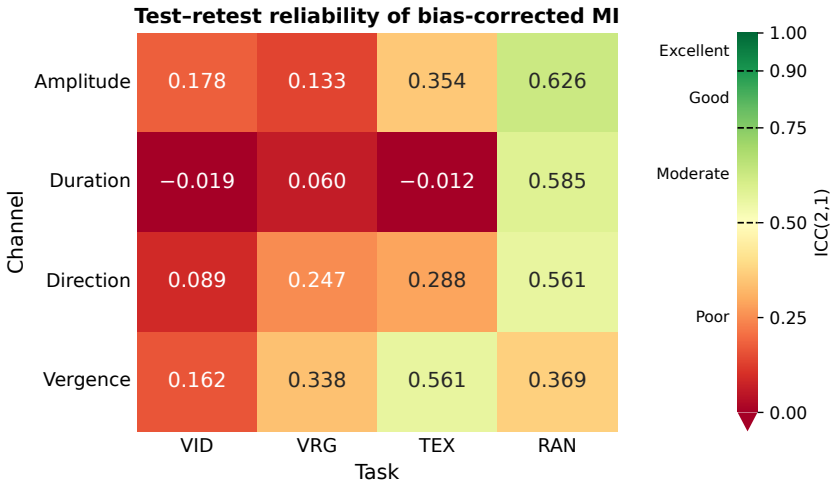


Figure 10. ICC(2,1) heatmap by channel and task. Colour scale runs from red (Poor, ICC < 0.50) through yellow (Moderate, 0.50–0.75) to green (Good, 0.75–0.90; Excellent, >0.90). Dashed lines on the colour bar mark the benchmark thresholds of Koo and Li [28]. The category labels are given beside the colour bar, and the arrow at its lower end indicates that negative estimates (duration on VID and TEX) are shown in the darkest red. Direction is the eight-octant scheme (Section 3.4).

Reliability is governed less by the channel than by whether the stimulus repeats. The RAN task reaches moderate values for all three kinematic channels (amplitude 0.626, duration 0.585, direction 0.561) and 0.369 for vergence, while the other tasks are predominantly poor. This is expected: RAN presents the same random-position statistics in both sessions, whereas VID shows a different clip and TEX a different text excerpt in each session of Round 1 [5], and VRG delivers only 24 depth events. The highest single value is amplitude on RAN (0.626); duration collapses to near zero on VID, TEX and VRG, where pacing and content both differ between sessions.

Vergence behaves differently from the rest, reaching moderate on TEX (0.561) rather than on RAN (0.369) or its own dedicated task (0.338). Reading at a fixed 0.6 m holds the depth of the stimulus constant across sessions, which may stabilise the vergence sequence more than a purpose-built depth probe delivering few events. That alone does not explain why TEX exceeds VID and RAN, whose stimuli also lie at constant depth (1 m); one untested possibility is that the resting vergence expected at 0.6 m ($\approx -6^\circ$ to -7°) falls inside the moderate bin, whereas at 1 m ($\approx -4.6^\circ$), it falls close to the -5° boundary (Section 5.7). We cannot distinguish these alternatives here. ICC(2,1) with absolute agreement reflects both the consistency of participants' relative positions and the agreement of their values across sessions: a systematic shift between sessions lowers it even when rankings are preserved. A stable between-participant offset in resting vergence would raise it without reflecting behaviour (Sections 4.4 and 5.7).

These ICCs are not a clean measure of channel reliability. Only RAN and VRG repeat the same stimulus in both sessions, and VRG delivers only 24 depth events, so RAN's values (ICC $\approx$ 0.56–0.63 for amplitude, duration and direction, which are moderate values by the benchmarks of Koo and Li [28]) are the only ones that approximate the repeatability of the measure. Elsewhere, the stimulus changes between sessions, and the lower values mix measurement instability with the response to different content. Whether the remaining instability tracks a participant's state is a possibility this design cannot test, since state, stimulus and measurement error are not separable here; the same caution applies to the prediction that the gaming study would show higher ICC values within a game than across games.

The results show that the gaze information channel framework can be computed on HMD eye-tracking data and yields interpretable differences between tasks. Whether the channels carry the same behavioural meaning as in screen-based viewing is not established by a single dataset with a chin rest, and we treat transfer as demonstrated computationally rather than behaviourally. Separately from that conclusion, the plug-in MI estimator needs bias correction whenever observed transitions are few relative to the size of the transition matrix. That is a property of finite samples rather than of the framework or the display, and it becomes salient here only because GazeBaseVR segments its recordings by task. Section 5.2 sets the two studies side by side, including the one numerical comparison the shared bin boundaries permit; two caveats bound it throughout. The chin rest [5] makes GazeBaseVR effectively head-fixed, so none of this tests immersive viewing, and the two studies differ in viewing duration by an order of magnitude.

5.2. Comparison with the Screen-Based Study

Table 5 sets the two studies side by side.

The two studies occupy different sampling regimes.

The bias characterised in Section 4.3 is a function of sequence length relative to state-space size, and on this axis, the screen-based study is comfortably placed. Pooling across 12 paintings at 45 s each, it accumulates on the order of 1200 fixations per observer—roughly

130 transitions per cell for its three-state channels, against 27 for our best-sampled task and 0.8 for the direction channel on VRG. Applying the analytic bias estimate at that sequence length gives under 1% of the reported pupil MI and about 3% of the nine-state AOI MI, so the values in [4] are essentially unaffected. The estimation problem addressed here is specific to the short, task-segmented recordings characteristic of HMD protocols, where a single task yields 53–239 transitions rather than several thousand. It follows from the experimental design rather than from the framework.

Table 5. The present study alongside the screen-based companion study [4]. Amplitude MI is the 3-category participant mean in both cases; VR values span the four tasks.

	Vila et al. [4]	**This Work**
Display	Screen (SMI iViewETG2w)	HMD (SMI ET-HMD, chin rest)
Stimulus	12 Van Gogh paintings	4 structured tasks
Viewing	Free, 45 s per painting	Task-directed
Sample	10 observers	407 participants
Channels	AOI, amp, dur, dir, pupil	amp, dur, dir, vergence
Amplitude MI	0.029 bits (uncorrected)	0.033 (VRG)–0.304 (RAN)
Cross-channel	19/20 pairs n.s. ($N = 10$)	Varies descriptively by task ($N \approx 400$)
Bias handling	Plug-in estimator	Permutation-corrected
Repeatability	Not assessed	ICC(2,1), two sessions

Amplitude follows the task, but task and display cannot be separated.

The screen-based amplitude baseline (0.029 bits) sits close to our lowest value (VRG, 0.033 bits), below TEX (0.052 bits), and an order of magnitude below RAN (0.304 bits). Both studies use the same fixed boundaries (<1°, 1°–5°, >5°), so the gap is not a discretisation artefact. Free viewing of a static painting imposes far less sequential amplitude structure than a random-saccade task, which imposes it by construction, which is compatible with reading the channel as a measure of task-imposed sequential constraint. The comparison cannot isolate the display, however: trackers, stimuli, sample sizes and estimators all differ (Section 5.4), and VID, the HMD task closest to free viewing, gives 0.087 bits, which is about three times the screen-based value.

The discrimination taxonomy is only partly consistent across displays.

Vila et al. [4] report that amplitude MI varies more across observers than across paintings, grouping it with duration and direction rather than with AOI. Only direction agrees here. Amplitude and duration have task means more spread than participant means in our data ($\sigma_{\text{task}}/\sigma_{\text{obs}} = 3.06$ and 1.52, Table 3), which is the opposite of their placement. The likely reason is that our four "tasks" differ far more from one another than paintings by a single artist do, so the task means are more widely spread here by construction. Nor does repeatability support the agreement: amplitude reaches moderate ICC values only on RAN (0.626) and is poor on the other three tasks.

Cross-channel association was limited by N.

With $N = 10$, Vila et al. [4] find 19 of 20 correlations non-significant—ten channel pairs tested at two levels per observer and per painting—and are appropriately cautious about inferring independence. At $N \approx 400$, we can separate regimes that a small sample would merge (Section 4.11): VID and VRG come closest to their result (two and one weak associations surviving correction, respectively), TEX shows two weak negative associations surviving correction, and RAN five of six, of which only the kinematic cluster survives the length control. The four patterns are therefore described rather than formally tested for

differences, which matters for the Channel Immersion Index (CII): weights fitted under free viewing need not transfer to constrained gameplay.

Vergence has no counterpart in the screen-based study.

Vergence is the one channel absent from [4], whose flat display offered no changing stereoscopic depth, so it has to be validated on its own terms rather than against a screen-based precedent.

5.3. Marginal Entropy as a Quality Control

Reporting marginal entropy alongside MI earns its place. The collapse candidates of Section 4.7 fall in two places only, amplitude on VRG and vergence on every task, and both reflect a genuine concentration of state occupancy: VRG saccades are almost all below 1°, and vergence varies over a narrower range than the kinematic variables. The MI values there nonetheless rest on few occupied states and should be read with that in mind.

5.4. Task-Driven Modulation of Channel MI

Amplitude and duration MI show clear differences across tasks (Tables 3 and 4), much less so in direction and least in vergence. We cannot separate task from modality, since the battery contains no screen-based equivalent of any of its four tasks. Each task organises the four channels differently: RAN produces the highest amplitude (0.304 bits) and duration (0.057 bits) MI under its externally imposed sequence of large, varied target-directed saccades; TEX produces the highest vergence MI (0.129 bits) for reasons the data do not settle, since VID and RAN also present a constant-depth stimulus (Section 5.6 lists candidate explanations); with quadrants, the primary scheme for task comparisons, the direction of MI is highest on VRG (0.113 bits) and then RAN (0.090), while with octants, the two exchange places (0.126 and 0.115), and we draw no conclusion from the VRG–RAN ordering (Section 4.5); the direction channel is undersampled in every task either way (Section 4.3); and vergence stays within 0.106–0.129 bits across all four tasks, which is the narrowest range of any channel. Tasks therefore appear to organise different components of gaze dynamics, which bears directly on the gaming study: if the difficulty, genre and interaction mode modulate specific channels, the CII can weight each channel by its task-specific capacity to carry information.

5.5. Independence of Channels and Implications for the CII

The cross-channel correlation analysis (Table A3) shows patterns of association that differ descriptively between tasks though less sharply than the zero-order correlations suggest and without a formal test of the difference (Section 4.11). Under free viewing, the associations are weak but present; under RAN, the three kinematic channels form a correlated cluster that survives control for recording length, while the apparent opposition of vergence to amplitude and duration does not. Bias correction and the length control each remove some of the apparent coupling, most of it under VRG and RAN, but neither identifies whether estimator bias or length-dependent behaviour is responsible.

For the CII, this is consistent with the *simple addition* of channel MI values being least problematic under free-viewing conditions, where participant-level associations are weak, and most problematic where an externally imposed task couples the kinematic channels, as under RAN, although it does not establish it. A weak measured correlation is not evidence that two channels carry non-overlapping information, because measurement error attenuates the correlation towards zero whether or not the underlying quantities overlap, and the free-viewing case should not be over-read for the reason given in Section 4.11: repeatability under VID is poor for all four channels, which is the condition under which

attenuation is largest. Whether channel MI values may be added therefore remains an open question for the gaming study. CII weights should be validated against the correlation structure measured during gameplay rather than assumed.

5.6. The Special Role of the Vergence Channel

Vergence sits apart from the other three channels on every measure we applied: its correlations with the kinematic channels vary descriptively between tasks, its task effect is the weakest by a wide margin ($W = 0.017$ against 0.698 for amplitude), and it is the most bias-robust, retaining 84–96% of its raw MI where direction retains 12–43%. Its ICC of 0.561 on TEX is the first between-session ICC reported for this channel.

That combination makes it plausible that vergence MI captures something not reducible to saccadic kinematics—proximity-related depth engagement—but the present data motivate the hypothesis rather than establish it, and the sensitivity analysis bounds how far it can be taken: Vergence MI decreased by approximately 68% after subtracting a rolling median from a within-recording-tercile version of the channel, which is a step that removes slow components (genuine eye movements, baseline changes or measurement drift) that the analysis cannot distinguish (Appendix A.6). Three of the four tasks present their stimulus on a plane at constant depth, so vergence MI there may instead reflect the persistence of resting vergence, slow drift (whose non-stationarity inflates the MI of a first-order Markov chain), an eccentricity-dependent disparity, or where resting vergence falls relative to the fixed bin boundaries, which differs between the 0.6 m and 1 m planes (Section 5.7). If the hypothesis holds, the channel should separate immersive gameplay from passive viewing, since interactive three-dimensional environments engage vergence more continuously than video does.

5.7. Limitations

Two estimation caveats apply throughout. The correction $\max(0, \hat{I} - \hat{I}_{\text{null}})$ rectifies at zero, so where the true MI is near zero—duration on VID and TEX—the reported means are biased slightly upward. And the linear interpolation applied before I-DT places no limit on gap length, so long gaps may create fixations or lengthen their durations artificially. Repeating the analysis with gaps longer than 75 ms left unfilled leaves the analysed set and every task ordering unchanged and moves no reported ratio by more than 0.01, so this does not drive the results (Appendix A.5).

Four groups of limitation bound these findings.

The paradigm is not immersive viewing.

Participants performed structured laboratory tasks on a chin rest with the view inside the headset held fixed so that the stimulus kept its position relative to the headset irrespective of residual head movement [5]. The recordings contain no head–eye coordination, the head rotation channel planned for the gaming study cannot be piloted here, and the constraint may suppress gaze variability. Pupil diameter is unrecorded, so a full seven-channel baseline is unavailable.

The tasks index oculomotor constraint rather than cognitive load.

GazeBaseVR was built for eye-movement biometrics: VRG and RAN require only that a target be followed, VID is passive, and TEX alone carries an explicit comprehension demand. Critically, the two run in opposite directions across the battery—TEX combines the highest comprehension requirement with the most stereotyped scan path, RAN the lowest with the most varied saccades—so we cannot say whether the task effects reflect cognitive state or motor demand, though the pattern leans towards the latter, since the channel responding

most strongly is the most purely motor. GazeBaseVR records no measure of comprehension or performance, and the self-reports we do test (Section 4.8) index effort and alertness, so comprehension cannot be addressed here in either direction. Separating load from motor demand needs a design that varies the first at constant oculomotor demand, which the gaming study should incorporate: difficulty manipulations otherwise vary both together.

The vergence channel carries a baseline confound.

ν is a simplified horizontal vergence measure computed with an assumed inter-ocular distance of 62 mm rather than an individualised three-dimensional reconstruction of fixation depth. Its bins are absolute while resting vergence differs between participants, so part of the between-participant variance is baseline rather than behaviour (Section 4.4). The demographic analysis is consistent with this: the channel's state occupancy differs by gender more sharply than its MI does (Section 4.10), which an interaction between fixed angular boundaries and inter-ocular distance could produce, although behavioural or other measurement-related sources cannot be excluded. Centring each participant's vergence on their own median would separate the two at the cost of comparability with the fixed depth geometry of VRG. Tracker accuracy adds a further uncertainty: the ET-HMD is specified at about 1° of spatial accuracy per eye, which is an offset from the target rather than a sample-to-sample fluctuation. Adding the two eyes in quadrature, which assumes their errors are independent, gives roughly 1.4° for $\nu = x^R - x^L$. Neither assumption is tested here: an offset shared by both eyes cancels in the difference, independent offsets do not, and the dataset reports no test–retest precision that would distinguish them. We therefore read 1.4° as an order of magnitude rather than a measured noise floor; it cannot by itself establish whether the device separates the two VRG depths. Near-target fixations are, however, 1.56° more convergent than far-target ones: against 2° nominally (Section 4.6). Two further contributions can be sized against a common reference (Appendix A.6). Subtracting a centred rolling median over eleven fixations lowers the mean corrected vergence MI from 0.190 to 0.061 bits, which is a decrease of approximately 68%; what the subtraction removes may include genuine eye movements, baseline changes and measurement drift, which the analysis cannot separate. Partialling out horizontal gaze position lowers the mean to 0.172 bits, a reduction of 9.5%, and within a recording, ν and the gaze x coordinate have a median $|\rho|$ of 0.22 to 0.25 by task with over half of recordings above 0.2. Much of what the vergence channel registers is therefore slow change in the measured angle rather than fixation-to-fixation change, and a smaller part co-varies with where the eyes are pointing. The decrease is largest in the three tasks held at constant depth (62% to 78%) and smallest on VRG (54%), which is consistent with part of the VRG sequence being genuine scripted depth change. One untested consequence would make the channel's individual differences partly geometric. At an inter-ocular distance of 62 mm, the ideal vergence angle is ≈3.6° at the 1 m depth of VID and RAN and ≈5.9° at the 0.6 m depth of TEX. On VRG, the measured angles are about 1° more convergent than the ideal ones (−11° and −9° against −10° and −8°; Section 3.2). Were that offset to hold across tasks, VID and RAN would sit near −4.6°, close to the −5° bin boundary, and vergence MI would then depend in part on how near each participant's resting vergence falls to that edge. A descriptive table giving the median ν and the percentage of fixations in each bin by task would settle it.

Two channels rest on inadequate samples.

The direction channel needs 64 transition-matrix cells but no task supplies more than 3.7 transitions per cell, and VRG supplies 0.8, so its corrected values carry lower confidence than the three-state channels, and we recommend the four-quadrant variant at these sequence lengths. The constraint binds state-space size and recording length together:

screen-based protocols accumulating thousands of fixations satisfy it comfortably, while task-segmented HMD recordings do not. States are not in one-to-one correspondence with stimulus events, either: RAN presents 80 fixation periods but yields ≈ 241 fixations per recording, and VRG presents 24 depth movements against ≈55 fixations (Table 1), so I-DT with a 1.5° dispersion threshold returns several states per event. In RAN, those extra states are small movements within a target presentation rather than mis-segmented target saccades: saccades that stay within one presentation have a median amplitude of 0.88° against 9.44° for those crossing a target change. Varying the detector's dispersion threshold and minimum duration changes how many such states are returned without changing the conclusions, and on VRG, a detector that also watches vergence returns about three times as many fixations on the recordings both detectors retain, so the 2D detector may under-segment the depth alternation (Appendix A.5). Smooth pursuit is excluded altogether, since I-DT presumes stable fixation periods that pursuit does not produce; extending the framework to it would require states defined on pursuit velocity or target error, which is a different formulation left to future work.

6. Conclusions

We have introduced the binocular vergence channel and evaluated it on head-mounted display data alongside the three channels the gaze information channel framework [1–4] already provides, using GazeBaseVR (407 participants, two same-day sessions, four tasks). Vergence is new to the framework; HMDs make it especially relevant because they present stereoscopic depth. The contribution is bringing vergence into the gaze information channel framework and applying that framework to head-mounted display data.

Vergence behaves unlike the other channels. The spread of its participant means exceeds the spread of its four task means under every discretisation we tested by a factor of between about 1.3 and 5.8; the $-9°/-5°$ absolute boundaries used here give the extreme of that range ($\sigma_{\text{task}}/\sigma_{\text{obs}} = 0.17$), and participant-centred or participant-relative boundaries give about 1.3 to 2.1 (Appendix A.6); changing the fixation detector's parameters moves it further (Appendix A.5). Amplitude is the reverse at 3.06. It is the most task-stable of the four ($W = 0.02$ against 0.70) and the most robust to finite-sample bias, retaining 84–96% of its raw MI where direction retains 12–43%, and its highest between-session ICC occurs on reading (0.561), whereas for VID, the stimulus content changes between sessions. Two qualifications travel with it: its state occupancy is narrow, and because part of the between-participant spread reflects resting vergence and measurement error rather than behaviour, the ratio should not be read as a measure of behavioural individual differences. Vergence MI decreased by approximately 68% after subtracting a rolling median from a within-recording-tercile version of the channel (Appendix A.6), which is a step that removes slow components (genuine eye movements, baseline changes or measurement drift) that the analysis cannot distinguish. Therefore, the channel as defined here quantifies sequential dependence in the binocular signal and has not yet been validated as a specific measure of depth-related visual behaviour.

Channel MI is not a task descriptor. It distinguishes the four tasks far less well than mean amplitude and duration do (68.4% against 99.0%), as it should: the channels measure how predictable successive gaze states are rather than how large or long they were. We use the task contrast to vary sequential structure rather than to recover it.

The established channels differ clearly across tasks. Task effects are strong for amplitude ($W = 0.70$) and duration ($W = 0.43$) but modest for direction ($W = 0.24$ with quadrants, 0.11 with octants) and small for vergence ($W = 0.02$); the amplitude and duration effects are large enough to be of practical rather than merely statistical size, whereas those for direction and vergence are not.

Channel dependence differs descriptively between the four tasks. Free viewing shows two weak vergence associations. Under RAN, a positive kinematic cluster remains after adjustment for fixation count, whereas the negative associations involving vergence are attenuated and are not significant after that adjustment and Bonferroni correction. Four tasks in one dataset cannot establish which task properties produce these patterns, so the finding is that dependence is not constant across conditions: we do not provide an account of what drives it. For the CII, this matters directly: weights fitted under free viewing need not transfer to constrained gameplay.

One methodological prerequisite underlies all of this. The plug-in estimator is biased upward when observed transitions are few relative to transition-matrix cells, and task-segmented HMD protocols make that unavoidable: the bias is negligible for well-sampled three-state channels yet dominant for the eight-state direction channel, which the uncorrected analysis ranks first of the four and the corrected analysis third. We recommend that bias correction, or at minimum a report of transitions per cell, accompany channel MI values whenever sequences are short.

These results inform a forthcoming study pairing the framework with immersion questionnaires during interactive VR gaming, where vergence should be engaged more continuously than in any task examined here.

Author Contributions: Conceptualisation, M.Y.M., M.V. and M.S.; methodology, M.Y.M., M.V. and M.S.; software, M.Y.M. and M.V.; validation, M.Y.M. and M.V.; formal analysis, M.Y.M.; resources, M.S.; data curation, M.Y.M. and M.V.; writing—original draft preparation, M.Y.M. and M.S.; writing—review and editing, M.V., M.C. and M.S.; visualisation, M.Y.M. and M.S.; supervision, M.S. and M.C.; project administration, M.S.; funding acquisition, M.S. and M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by grants PID2023-149976OB-C21 and PID2023-149976OB-C22 from the Spanish Ministry of Science and Innovation (MCIN/AEI/10.13039/501100011033) and by the European Regional Development Fund (ERDF). This work was also supported by AGAUR, grant 2025 SGR 01671.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The GazeBaseVR dataset is publicly available under a CC BY 4.0 licence on Figshare [21]. This study is a secondary analysis of these data; ethical approval for the original collection was obtained by the dataset authors, and the recordings were anonymised in accordance with the informed consent of the original participants [5]. The analysis code and the derived per-recording channel measures supporting the findings of this study are openly available on Zenodo at https://doi.org/10.5281/zenodo.24157580, which is the DOI of the deposit version used here. The concept DOI https://doi.org/10.5281/zenodo.23222235 resolves to the latest version. Because fixation detection is sensitive to implementation details that nominal parameter values do not capture (Section 3.3), the detector is specified algorithmically in the Methods and the code is the definitive reference for any attempted replication.

Acknowledgments: The authors thank the GazeBaseVR team (Lohr, Aziz, Friedman, Komogortsev) for making their dataset openly available. During the preparation of this manuscript, the authors used Claude Opus 5 (Anthropic) for analysis code development, statistical reanalysis, manuscript revision and figure preparation, and GPT-5.5 (OpenAI) for critical revision of the manuscript. All analyses were re-executed from the raw recordings, and every reported value was checked against the deposited output files. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CII	channel immersion index
ET	eye tracking
GTE	gaze transition entropy
HMD	head-mounted display
ICC	intraclass correlation coefficient
MI	mutual information
NMI	normalised mutual information
PT	Panzeri–Treves
RAN	random oblique saccade task
SGE	stationary gaze entropy
TEX	self-paced reading task
VID	video-viewing task
VR	virtual reality
VRG	depth-alternation task

Appendix A. Supplementary Tables

This appendix collects the supplementary tables referenced from Sections 4 and 5. Each supports a finding stated in the main text and is reproduced here in full for readers who wish to check the underlying numbers.

Appendix A.1. Finite-Sample Bias by Channel and Task

Table A1 accompanies Section 4.3: for every channel and task, it presents the raw MI, the permutation null, the analytic Panzeri–Treves estimate of the same quantity, the corrected value, the proportion of the raw value retained, the sampling adequacy in transitions per transition-matrix cell, and the percentage of recordings whose raw MI exceeds the 95th percentile of their own null. It also gives the four-quadrant direction variant, which is not reported elsewhere in full.

Table A1. Finite-sample bias by channel and task. "Retained" is $\hat{I}_{\text{corr}}/\hat{I}$. "T/cell" is the mean number of observed transitions per transition-matrix cell ($(n_{\text{fix}}-2)/S^2$); values below ∼5 indicate undersampling. "PT" is the analytic Panzeri–Treves estimate of the same bias (Equation (13)), computed independently; "Sig." is the percentage of recordings whose raw MI exceeds the 95th percentile of their own null. Both direction schemes are shown, which are labelled by row (Section 3.4).

Channel	Task	Raw	Null	PT	Corrected	Retained	T/Cell	Sig.
Amplitude	VID	0.110	0.023	0.022	0.087	79%	14.9	82%
	VRG	0.055	0.026	0.027	0.033	61%	5.9	30%
	TEX	0.066	0.014	0.013	0.052	79%	25.5	83%
	RAN	0.317	0.013	0.012	0.304	96%	26.5	100%
Vergence	VID	0.123	0.008	0.008	0.115	94%	14.8	89%
	VRG	0.127	0.022	0.022	0.107	84%	5.9	57%
	TEX	0.140	0.010	0.011	0.129	93%	25.4	89%
	RAN	0.111	0.005	0.005	0.106	96%	26.5	94%
Direction (8-oct)	VID	0.380	0.311	0.271	0.074	20%	2.1	36%
	VRG	0.965	0.864	0.846	0.115	12%	0.8	28%
	TEX	0.218	0.142	0.155	0.076	35%	3.6	71%
	RAN	0.294	0.168	0.152	0.126	43%	3.7	80%
Direction (4-quad)	VID	0.092	0.052	0.050	0.043	47%	8.4	40%
	VRG	0.301	0.200	0.166	0.113	37%	3.3	35%
	TEX	0.086	0.031	0.030	0.055	64%	14.3	74%
	RAN	0.119	0.028	0.028	0.090	76%	14.9	89%
Duration	VID	0.025	0.020	0.020	0.009	35%	14.9	11%
	VRG	0.110	0.077	0.070	0.045	41%	5.9	19%
	TEX	0.007	0.004	0.004	0.004	62%	25.5	14%
	RAN	0.070	0.013	0.012	0.057	82%	26.5	81%

Appendix A.2. Pooled Estimates

Table A2 accompanies Section 4.4 and presents the channel MI estimated from transition counts summed across all recordings within a task as well as the mean of the bias-corrected per-recording estimates.

Table A2. Pooled channel MI (transition counts summed across all recordings within a task) against the mean bias-corrected per-recording estimate. N_{tr}: pooled transitions. The pooled estimate is essentially bias-free but describes the aggregate process rather than the typical recording. Bold: the vergence ratios, which are well above 1 because resting vergence differs between participants (Section 4.4). Direction is the four-quadrant scheme (Section 3.4).

Channel	Task	N_{tr}	T/cell	Pooled MI	Mean Corrected	Ratio
Amplitude	VID	108,662	12,074	0.068	0.087	0.78
	VRG	36,837	4093	0.043	0.033	1.30
	TEX	186,382	20,709	0.034	0.052	0.66
	RAN	193,806	21,534	0.243	0.304	0.80
Duration	VID	108,662	12,074	0.004	0.009	0.49
	VRG	36,837	4093	0.067	0.045	1.49
	TEX	186,382	20,709	0.005	0.004	1.12
	RAN	193,806	21,534	0.037	0.057	0.65
Direction (4-quad)	VID	108,662	6791	0.040	0.043	0.92
	VRG	36,837	2302	0.066	0.113	0.59
	TEX	186,382	11,649	0.043	0.055	0.77
	RAN	193,806	12,113	0.085	0.090	0.94
Vergence	VID	108,224	12,025	0.369	0.115	**3.22**
	VRG	36,626	4070	0.546	0.107	**5.08**
	TEX	185,615	20,624	0.387	0.129	**2.99**
	RAN	193,312	21,479	0.333	0.106	**3.14**

Appendix A.3. Cross-Channel Correlations

Table A3 accompanies Section 4.11, presenting the Spearman rank correlations between the bias-corrected MI values of each channel pair, computed separately within each task.

Table A3. Spearman ρ between bias-corrected channel MI values across participants, by task, with one value per participant obtained by averaging their available valid sessions. $\rho \mid n_{fix}$ is the same correlation partialled on the participant's mean fixation count, which all four channels share and which drives finite-sample bias. N counts participants with all four channels estimable in that task: 406 for VID, TEX and RAN, 377 for VRG. Bracketed values are 95% percentile bootstrap intervals over 2000 resamples of participants. ** $p < 0.0083$ (Bonferroni); * $p < 0.05$ uncorrected. Amp = amplitude, Dur = duration, Dir = direction, Ver = vergence. Direction is the eight-octant scheme (Section 3.4).

Task		Amp–Dur	Amp–Dir	Amp–Ver	Dur–Dir	Dur–Ver	Dir–Ver
VID	ρ	−0.07	−0.03	−0.19 **	+0.01	+0.04	+0.14 **
		[−0.17, +0.02]	[−0.13, +0.06]	[−0.28, −0.09]	[−0.09, +0.11]	[−0.06, +0.13]	[+0.05, +0.24]
	$\rho \mid n_{fix}$	−0.07	−0.03	−0.19 **	−0.01	+0.03	+0.14 **
		[−0.17, +0.02]	[−0.13, +0.06]	[−0.28, −0.09]	[−0.11, +0.09]	[−0.07, +0.13]	[+0.04, +0.24]
VRG	ρ	−0.01	+0.09	−0.03	−0.06	+0.18 **	−0.06
		[−0.11, +0.10]	[−0.02, +0.19]	[−0.13, +0.08]	[−0.16, +0.04]	[+0.07, +0.27]	[−0.16, +0.05]
	$\rho \mid n_{fix}$	−0.01	+0.09	−0.03	−0.05	+0.19 **	−0.06
		[−0.12, +0.10]	[−0.01, +0.20]	[−0.13, +0.08]	[−0.15, +0.05]	[+0.08, +0.28]	[−0.17, +0.05]
TEX	ρ	−0.13 *	−0.20 **	+0.07	−0.03	−0.06	−0.17 **
		[−0.23, −0.03]	[−0.30, −0.11]	[−0.03, +0.17]	[−0.13, +0.06]	[−0.16, +0.03]	[−0.26, −0.07]
	$\rho \mid n_{fix}$	−0.12 *	−0.20 **	+0.06	−0.05	−0.02	−0.15 **
		[−0.22, −0.02]	[−0.29, −0.11]	[−0.04, +0.15]	[−0.15, +0.04]	[−0.12, +0.08]	[−0.24, −0.05]
RAN	ρ	+0.81 **	+0.48 **	−0.33 **	+0.34 **	−0.28 **	−0.12 *
		[+0.77, +0.84]	[+0.40, +0.55]	[−0.41, −0.24]	[+0.25, +0.43]	[−0.37, −0.18]	[−0.23, −0.03]
	$\rho \mid n_{fix}$	+0.55 **	+0.53 **	−0.12 *	+0.31 **	−0.05	−0.07
		[+0.45, +0.63]	[+0.46, +0.60]	[−0.21, −0.02]	[+0.21, +0.40]	[−0.15, +0.05]	[−0.16, +0.03]

Appendix A.4. Between-Session Reliability

Table A4 accompanies Section 5.1, presenting the ICC(2,1) between Sessions 1 and 2 of Round 1 for every channel and task.

Table A4. ICC(2,1) between-session reliability between Sessions 1 and 2 of Round 1. *N* counts *participants* contributing a valid estimate in both sessions (not recordings); lower for VRG (317) than other tasks (~406) because more VRG recordings are excluded by the 21-fixation minimum (20 states). CI: 95% percentile bootstrap interval, resampling participants 2000 times with each participant's two session values kept together. Labels: Poor < 0.50, Moderate 0.50–0.75, Good 0.75–0.90, Excellent > 0.90. Both direction schemes are shown, which are labelled by row (Section 3.4).

Channel	Task	*N*	ICC(2,1)	95% CI	Label
Amplitude	VID	406	0.178	[0.092, 0.260]	Poor
	VRG	317	0.133	[0.004, 0.267]	Poor
	TEX	406	0.354	[0.256, 0.443]	Poor
	RAN	406	0.626	[0.564, 0.683]	Moderate
Duration	VID	406	−0.019	[−0.085, 0.066]	Poor
	VRG	317	0.060	[−0.051, 0.159]	Poor
	TEX	406	−0.012	[−0.085, 0.076]	Poor
	RAN	406	0.585	[0.508, 0.653]	Moderate
Direction (8-oct)	VID	406	0.089	[0.015, 0.168]	Poor
	VRG	317	0.247	[0.122, 0.358]	Poor
	TEX	406	0.288	[0.194, 0.379]	Poor
	RAN	406	0.561	[0.483, 0.628]	Moderate
Direction (4-quad)	VID	406	0.113	[0.041, 0.191]	Poor
	VRG	317	0.090	[−0.020, 0.197]	Poor
	TEX	406	0.449	[0.368, 0.527]	Poor
	RAN	406	0.560	[0.497, 0.618]	Moderate
Vergence	VID	406	0.162	[0.071, 0.249]	Poor
	VRG	317	0.338	[0.150, 0.496]	Poor
	TEX	406	0.561	[0.472, 0.631]	Moderate
	RAN	406	0.369	[0.255, 0.471]	Poor

Appendix A.5. Sensitivity to Preprocessing

This appendix examines two preprocessing choices: filling missing samples without limiting gap duration and the I-DT parameters. We measured the gaps and then repeated the whole analysis under five alternative settings.

Missing samples are uncommon but not negligible, and their distribution is strongly skewed. They account for 0.47% of samples on RAN, 0.75% on VRG, 1.06% on TEX and 1.52% on VID. Between 70% and 93% of recordings contain at least one gap, with a median gap of 12–16 ms, or three to four samples. Of the 79,727 gaps, 5397 (6.8%) exceed 75 ms and 787 exceed 150 ms, and between 38% and 60% of recordings contain at least one gap longer than 75 ms with the longest reaching 3.3 s. A gap longer than the 80 ms minimum fixation duration could by itself be detected as a fixation, so we repeated the analysis leaving every gap longer than 75 ms unfilled with no window and no transition allowed to span one.

Table A5 gives the outcome. With the long gaps left unfilled, the same 3130 recordings are analysed, every reported ratio changes by at most 0.01 and every task ordering is preserved. The I-DT parameters move the numbers more, mostly by changing how many recordings clear the 21-fixation criterion (2931 at a 2.0° dispersion against 3242 at 1.0°), but the main results hold. Amplitude varies several times more across tasks than across participants in every setting (ratio 2.51 to 3.46) and is highest on RAN in every setting; the four-quadrant direction channel gives the identical task ordering in all six settings; and vergence's ratio stays far below one throughout (0.09 to 0.26). The orderings that do move

under reparameterisation are the ones already identified as unstable: the lower three tasks for amplitude and the eight-octant direction ordering.

Table A5. The analysis repeated under five alternative preprocessing settings. "Ratio" is $\sigma_{task}/\sigma_{obs}$. N is the number of recordings clearing the 21-fixation criterion.

Setting	*N*	Amp. Ratio	Amplitude Order	Direction-4 Order	Verg. Ratio
Main analysis	3130	3.06	RAN>VID>TEX>VRG	VRG>RAN>TEX>VID	0.17
Gaps > 75 ms unfilled	3130	3.07	RAN>VID>TEX>VRG	VRG>RAN>TEX>VID	0.17
Dispersion 1.0°	3242	3.24	RAN>TEX>VID>VRG	VRG>RAN>TEX>VID	0.15
Dispersion 2.0°	2931	3.05	RAN>VID>VRG>TEX	VRG>RAN>TEX>VID	0.26
Minimum 60 ms	3139	2.51	RAN>VID>TEX>VRG	VRG>RAN>TEX>VID	0.23
Minimum 100 ms	3125	3.46	RAN>VID>VRG>TEX	VRG>RAN>TEX>VID	0.09

In RAN, across the 812 recordings, the task presents a median of 80 target positions and the detector returns a median of 231 fixations or 2.89 per target presentation. The two populations of saccades are separated. Of the 194,618 saccades, 66.2% stay within one target presentation and have a median amplitude of 0.88°, while those that cross a target change have a median amplitude of 9.44°; 55% of the former are below 1° against 19% of the latter. The additional fixations are therefore small movements within a target presentation rather than mis-segmented target-driven saccades. Whether those small movements are physiological or an artefact of the detector is not settled by their amplitude alone, and the parameter sweep shows that loosening or tightening the detector changes how many such events are returned without changing the conclusions.

Finally, the vergence channel on VRG was rebuilt with a detector that requires a window to be compact in vergence as well as in gaze position (vergence range $\leq 1°$), since a change of fixation depth need not move the 2D gaze point. Across all recordings each detector retains, it returns 147 fixations per recording against 55, and 803 recordings clear the length criterion against 694. Because method and sample change together there, the comparison is repeated on the *common set*: the 684 VRG recordings both detectors retain. There, with each of the 375 participants contributing the mean of their matched sessions under each detector, as in Section 3.7, the vergence-aware detector returns 153 fixations against 52, the corrected vergence MI is 0.156 bits against 0.107 (Wilcoxon signed-rank over participants, rank-biserial 0.34 $[0.23, 0.45]$, $p < 0.001$; 204 participants higher, 144 lower, 27 tied), and the amplitude MI is 0.020 against 0.034. The difference is therefore a property of the segmentation rather than of which recordings each detector happens to retain, and the 2D detector may under-segment the depth alternation. We report this as a robustness check rather than adopting it, because segmenting on vergence and then measuring the vergence channel share information, so the higher MI is not independent evidence of a stronger signal.

Appendix A.6. Sensitivity of the Vergence Channel

The vergence channel is built on absolute angular boundaries, while resting vergence differs between participants. To separate the behavioural signal from that baseline, we recomputed the channel on the same 3130 analysed recordings, with the same permutation null, under five alternatives to the $-9°/-5°$ boundaries of Section 3.4: two other absolute boundary pairs; values centred on each participant's median vergence, pooled over all of that participant's fixations, with boundaries at $\pm 2°$; boundaries at each participant's own terciles; and boundaries at each recording's own terciles.

Two further variants ask what the channel is responding to rather than how it is cut: the sequence with slow components removed by subtracting a centred rolling median over eleven fixations, and the residual of v after regressing out the horizontal gaze coordinate. Both are discretised on within-recording terciles, because neither transformed quantity is

on the original angular scale. Terciles equalise state occupancy and move MI on their own, so these two are compared against raw v cut at within-recording terciles rather than against the absolute boundaries.

Table A6 gives the result. The between-participant spread is stable: σ_{obs} lies between 0.058 and 0.066 under all six discretisations, so the finding that vergence MI differs between participants does not depend on the boundaries. The ratio is not stable, because the four task means are close together under the $-9°/-5°$ boundaries ($\sigma_{task} = 0.0107$) and about 1.7 to 4 times more spread under every alternative discretisation, so $\sigma_{task}/\sigma_{obs}$ rises from 0.17 to between 0.32 and 0.75. The direction of the comparison holds throughout (vergence always varies more across participants than across tasks), but its magnitude ranges from about 1.3-fold to 5.8-fold, and the boundaries used here sit at the extreme. The rank order of the four tasks is not stable either; what holds is the weaker statement that VRG is never the highest.

Against their matched reference of 0.190 bits, the two transformed variants separate. Subtracting the rolling median lowers the mean corrected MI to 0.061 bits, which is a decrease of approximately 68%. Removing the horizontal gaze coordinate lowers it to 0.172 bits, which is a reduction of 9.5%. Both effects differ by task, which are each measured against that task's own per-recording-tercile reference. The rolling-median reduction is 78% on VID, 70% on TEX and 62% on RAN, the three tasks at constant depth, and 54% on VRG, the one task whose stimulus changes depth; the gaze reduction is close to 8% on the three fixed-depth tasks and 18% on VRG. Slow components are therefore a substantial contributor to what the channel registers—whether they arise from genuine vergence change, baseline changes or measurement drift, which this analysis cannot separate—and horizontal gaze position is a modest one. Within a recording, the two are related more strongly than a signed average suggests, the median $|\rho|$ between v and the gaze x coordinate being 0.22 to 0.25 by task, with 55% to 59% of recordings above 0.2, while the signed means (0.01 to 0.06) are small only because the positive and negative associations cancel. Subtracting the rolling median also lowers σ_{obs}, from 0.064 on the matched reference to 0.025, so a large part of the between-participant spread lies in these slow components as well.

Table A6. Vergence channel under six discretisations of v (upper block) and two transformations of it (lower block), 3130 recordings. σ_{obs}: SD across participants of their mean corrected MI. σ_{task}: SD across the four task means. "Top" is the task with the highest mean MI and "VRG" its rank among the four. The two transformations are discretised on within-recording terciles, so their reference is the last row of the upper block, not the first. Italics mark the two block headings and the note identifying the reference row for the lower block.

	σ_{obs}	σ_{task}	Ratio	Inverse	Mean MI	Top	VRG
Discretisations of v							
Absolute $-9°/-5°$ (used here)	0.0626	0.0107	0.17	5.8	0.1148	TEX	3
Absolute $-10°/-6°$	0.0579	0.0185	0.32	3.1	0.1032	TEX	3
Absolute $-8°/-4°$	0.0593	0.0279	0.47	2.1	0.0995	TEX	2
Participant-centred, $\pm 2°$	0.0662	0.0315	0.48	2.1	0.1043	TEX	3
Participant terciles	0.0595	0.0443	0.75	1.3	0.1372	TEX	4
Recording terciles (*reference below*)	0.0636	0.0359	0.56	1.8	0.1901	VID	4
Transformations, cut at recording terciles							
Rolling median subtracted	0.0252	0.0104	0.41	2.4	0.0612	RAN	2
Gaze-x partialled out	0.0601	0.0394	0.66	1.5	0.1721	VID	4

Appendix A.7. Recording Length and Direction Categorisation

Two robustness questions are reported here: whether the permutation correction removes the dependence of MI on recording length and which direction results survive both the four- and the eight-state categorisation.

Table A7 gives the Spearman correlation between a recording's fixation count and its MI before and after correction. Correction reduces the association substantially, the median $|\rho|$ across the sixteen channel–task cells falling from 0.35 to 0.08 and for direction on VRG from -0.87 to $+0.07$, but it does not remove it. Ten of the sixteen per-cell intervals exclude zero, and six of those still do when the same replicates are read at Bonferroni-adjusted percentiles; the same six cells are significant under Holm-adjusted recording-level tests that ignore the clustering, so correcting the dependence leaves the set unchanged and the step from six to ten reflects the absence of multiplicity adjustment. The strongest association is amplitude on RAN ($\rho = -0.75$, $[-0.79, -0.71]$). Not all residual associations are negative; vergence MI rises with length on TEX and RAN. For the shortest task, VRG, no channel shows a residual association that survives adjustment, vergence included ($\rho = 0.07$, $[-0.01, 0.16]$). Corrected values should therefore not be compared across recordings of very different length without regard to this dependence, and the correction should not be described as removing the length effect.

Table A7. Spearman ρ between the number of fixations in a recording and its MI, raw and permutation-corrected. The point estimate is computed over recordings; the 95% interval is a percentile *cluster* bootstrap over 2000 resamples of participants, each resampled participant contributing all of their recordings, as in Section 3.7 (seed fixed per cell). * marks an interval excluding zero without adjustment for the sixteen comparisons; † marks the six that still exclude zero when the same bootstrap replicates are read at Bonferroni-adjusted percentiles ($\alpha = 0.05/16$). Intervals are given to three decimals (five in one cell) where two would round a bound to zero.

Channel	Task	N_{rec}	N_{part}	ρ_{raw}	ρ_{corr}	95% CI	
Amplitude	VID	812	406	−0.14	−0.08	[−0.147, −0.005]	*
	VRG	694	377	−0.06	−0.02	[−0.10, +0.05]	
	TEX	812	406	−0.03	+0.08	[+0.005, +0.156]	*
	RAN	812	406	−0.76	−0.75	[−0.79, −0.71]	*†
Duration	VID	812	406	−0.44	−0.20	[−0.26, −0.13]	*†
	VRG	694	377	−0.46	−0.07	[−0.14662, +0.00004]	
	TEX	812	406	−0.13	−0.09	[−0.156, −0.008]	*
	RAN	812	406	−0.70	−0.68	[−0.72, −0.63]	*†
Direction (8-oct)	VID	812	406	−0.49	+0.07	[−0.004, +0.138]	
	VRG	694	377	−0.87	+0.07	[−0.004, +0.149]	
	TEX	812	406	−0.42	−0.08	[−0.156, −0.008]	*
	RAN	812	406	−0.47	−0.17	[−0.25, −0.09]	*†
Vergence	VID	812	406	+0.04	+0.05	[−0.03, +0.12]	
	VRG	694	377	+0.02	+0.07	[−0.01, +0.16]	
	TEX	812	406	+0.19	+0.22	[+0.14, +0.29]	*†
	RAN	812	406	+0.27	+0.28	[+0.20, +0.35]	*†

Table A8 compares the two direction categorisations at participant level, each participant contributing the mean of their available valid sessions, so that the Friedman blocks are independent and the unit matches Table 4 ($N = 376$). Both give a significant Friedman effect of task with quadrants the more consistent ($W = 0.239$ [0.194, 0.293] against 0.113 [0.083, 0.151]). Four of the six pairwise contrasts are significant with the same sign under both schemes: VRG and RAN each exceed VID and TEX. VID versus TEX is significant only under quadrants, and VRG versus RAN is significant only under quadrants and with the opposite sign to the octant point estimate, which is why the rank order of the two highest tasks differs between schemes. The interpretation in Section 4.5 rests on the four contrasts that hold under both.

Table A8. Direction channel under both categorisations at participant level: one value per participant and task, the mean over that participant's available valid sessions, as in Table 4. $N = 376$ participants with valid data in all four tasks. Pairwise p-values are Wilcoxon signed-rank with Holm adjustment within scheme; the interval on W is a percentile bootstrap over 2000 resamples of participants. The transitions row counts transitions per occupied cell, unlike the nominal per-cell figure in Table A1. **: significant after Holm adjustment at $\alpha = 0.05$; n.s.: not significant.

	8 Octants	4 Quadrants
Friedman χ^2_F	127.1	269.9
Kendall's W	0.113 [0.083, 0.151]	0.239 [0.194, 0.293]
Task order	RAN > VRG > TEX > VID	VRG > RAN > TEX > VID
Transitions/occupied cell, VRG	0.87	3.3
VID vs VRG	VRG > VID **	VRG > VID **
VID vs. RAN	RAN > VID **	RAN > VID **
VRG vs. TEX	VRG > TEX **	VRG > TEX **
TEX vs. RAN	RAN > TEX **	RAN > TEX **
VID vs. TEX	n.s.	TEX > VID **
VRG vs. RAN	n.s.	VRG > RAN **

References

1. Ma, L.; Sbert, M.; Feixas, M. Gaze information channel. In *Proceedings of the Pacific Rim Conference on Multimedia (PCM 2018)*; Springer: Cham, Switzerland, 2018; pp. 575–585. [CrossRef]
2. Hao, Q.; Sbert, M.; Feng, Q.; Zhang, J. Entropy based analysis of gaze fixation duration. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2022)*; IEEE: New York, NY, USA, 2022; pp. 1–9.
3. Hao, Q.; Sbert, M.; Feixas, M.; Zhang, Y.; Vila, M.; Zhang, J. Saccade direction information channel. In *Proceedings of ICONIP 2022*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2023; Volume 13623, pp. 37–50. [CrossRef]
4. Vila, M.; Hao, Q.; Feixas, M.; Martin, M.Y.; Sbert, M. Saccade amplitude and pupil diameter information channels: Extending the gaze information channel framework and assessing cross-channel association in eye tracking of Van Gogh paintings. *Entropy* **2026**, *28*, 767. [CrossRef]
5. Lohr, D.; Aziz, S.; Friedman, L.; Komogortsev, O.V. GazeBaseVR, a large-scale, longitudinal, binocular eye-tracking dataset collected in virtual reality. *Sci. Data* **2023**, *10*, 177. [CrossRef]
6. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
7. Ellis, S.R.; Stark, L. Statistical dependency in visual scanning. *Hum. Factors* **1986**, *28*, 421–438. [CrossRef]
8. Krejtz, K.; Szmidt, T.; Duchowski, A.T.; Krejtz, I. Entropy-based statistical analysis of eye movement transitions. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '14)*; ACM: New York, NY, USA, 2014; pp. 159–166.
9. Krejtz, K.; Duchowski, A.; Szmidt, T.; Krejtz, I.; González Perilli, F.; Pires, A.; Vilaro, A.; Villalobos, N. Gaze transition entropy. *ACM Trans. Appl. Percept.* **2015**, *13*, 4. [CrossRef]
10. Shiferaw, B.; Downey, L.; Crewther, D. A review of gaze entropy as a measure of visual scanning efficiency. *Neurosci. Biobehav. Rev.* **2019**, *96*, 353–366. [CrossRef]
11. Moreno-Arjonilla, E.; López-Ruiz, A.; Jiménez-Pérez, J.R.; Callejas-Aguilera, J.E.; Jurado, J.M. Eye-tracking on virtual reality: A survey. In *Virtual Reality*; Springer Nature: London, UK, 2024. [CrossRef]
12. Cheng, J.; Chukoskie, L. Eye tracking in virtual reality for neurorehabilitation. *Front. Virtual Real.* **2026**, *7*, 1757251. [CrossRef]
13. Lohr, D.; Komogortsev, O.V. Eye Know You Too: Toward viable end-to-end eye movement biometrics for user authentication. *IEEE Trans. Inf. Forensics Secur.* **2022**, *17*, 3151–3164. [CrossRef]
14. Lohr, D.; Proulx, M.J.; Komogortsev, O. Establishing a baseline for gaze-driven authentication performance in VR: A breadth-first investigation on a very large dataset. In Proceedings of the 2024 IEEE International Joint Conference on Biometrics (IJCB), Buffalo, NY, USA, 15–18 September 2024; pp. 1–10.
15. Raju, M.H.; Lohr, D.J.; Komogortsev, O.V. Evaluating eye movement biometrics in virtual reality: A comparative analysis of VR headset and high-end eye-tracker collected dataset. *arXiv* **2024**, arXiv:2405.03287. [CrossRef]
16. Grandhi, S.G.; Samet, S. Evaluating the long-term viability of eye-tracking for continuous authentication in virtual reality. In Proceedings of the AIMLA, CCNET, NLTM—2025, CS & IT (CSCP), Chennai, India, 15–16 February 2025; pp. 11–22. [CrossRef]
17. Patergianakis, A.; Lambrinoudakis, C. Through the looking glass: Eye tracking biometrics and the loss of anonymity in extended reality. *Int. J. Inf. Secur.* **2026**, *25*, 76. [CrossRef]
18. Ren, X.; Fan, J.; Xu, N.; Wang, S.; Dong, C.; Wen, Z. DPGazeSynth: Enhancing eye-tracking virtual reality privacy with differentially private data synthesis. *Inf. Sci.* **2024**, *675*, 120720. [CrossRef]

19. Zandi, S.; Luhan, G. Exploring gaze dynamics in virtual reality through multiscale entropy analysis. *Sensors* **2024**, *24*, 1781. [CrossRef]
20. Zafar, N.; Locke, J.; Chaudhry, S.A. Deep learning-based visual fatigue detection using eye gaze patterns in VR. In *2025 IEEE International Symposium on Emerging Metaverse (ISEMV)*; IEEE: New York, NY, USA, 2025; pp. 69–76. [CrossRef]
21. Lohr, D.; Aziz, S.; Friedman, L.; Komogortsev, O.V. GazeBaseVR Data Repository. Figshare 2022. Available online: https://figshare.com/articles/dataset/GazeBaseVR_Data_Repository/21308391?file=38844024 (accessed on 15 November 2026).
22. Krakowczyk, D.G.; Reich, D.R.; Chwastek, J.; Jakobi, D.N.; Prasse, P.; Süss, A.; Turuta, O.; Kasprowski, P.; Jäger, L.A. pymovements: A Python Package for Processing Eye Movement Data. In *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications (ETRA '23)*; ACM: New York, NY, USA, 2023; p. 53. [CrossRef]
23. Salvucci, D.D.; Goldberg, J.H. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '00)*; ACM: New York, NY, USA, 2000; pp. 71–78.
24. Panzeri, S.; Senatore, R.; Montemurro, M.A.; Petersen, R.S. Correcting for the sampling bias problem in spike train information measures. *J. Neurophysiol.* **2007**, *98*, 1064–1072. [CrossRef]
25. Montemurro, M.A.; Senatore, R.; Panzeri, S. Tight data-robust bounds to mutual information combining shuffling and model selection techniques. *Neural Comput.* **2007**, *19*, 2913–2957. [CrossRef]
26. Miller, G.A. Note on the bias of information estimates. In *Information Theory in Psychology: Problems and Methods*; Quastler, H., Ed.; Free Press: Glencoe, IL, USA, 1955; pp. 95–100.
27. Panzeri, S.; Treves, A. Analytical estimates of limited sampling biases in different information measures. *Netw. Comput. Neural Syst.* **1996**, *7*, 87–107. [CrossRef]
28. Koo, T.K.; Li, M.Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* **2016**, *15*, 155–163. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.