

Beyond Scanpaths: Graph-Based Gaze Simulation in Dynamic Scenes

Supplementary Material

1. The Focus100 Dataset

Focus100 is a new dataset designed to facilitate research on dynamic human attention in driving scenarios, particularly for the development and evaluation of gaze estimation models. This dataset addresses critical limitations in existing driving gaze datasets, which often lack raw gaze data or sufficient scenario diversity [8, 9, 15, 43]. Unlike datasets that provide only aggregated saliency maps, Focus100 provides high-resolution, time-stamped gaze sequences from 30 participants viewing 100 egocentric driving videos. This rich data enables the study of fine-grained temporal attention patterns and scanpath dynamics, crucial for understanding human behaviour in complex driving environments.

Although the DR(eye)VE dataset [37] offers raw gaze sequences, it is hampered by limitations such as temporal misalignment, low scenario complexity, having only a single gaze sequence recorded per driving video, and lack of gaze data in the image plane (instead registered to moving driver-worn eye tracker glasses) [29]. While efforts have been made to enrich the dataset through in-lab gaze tracking [19], these have only addressed a small subset of the data. Focus100 overcomes these shortcomings by providing a diverse set of driving scenarios, several precise gaze recordings per driving video (in image coordinates), making it a valuable resource for advancing research in driver attention and automotive safety.

1.1. Data Collection

Driving Footage The driving videos incorporated in Focus100 were captured using a test vehicle equipped with a front-facing camera installed behind the windscreen to provide a point-of-view similar to that of the driver. The camera used for recording had a 52-degree horizontal visual angle and captured footage at 10 frames per second with a resolution of 1280×806 px. Comparable FoV and frame-rate settings are used in several driving datasets, *e.g.* Waymo Open [42]: 55° @10Hz; BDD100K [45]: 48° @30Hz; EuroPVI [3]: 52° @10Hz. All images were subjected to a calibration process to eliminate distortion and cropped to 1280×640 px resolution to remove calibration artefacts and ego-car bonnet pixels.

Driving sessions, which lasted up to 8 hours per day, were carried out over two weeks in and around the cities of Brussels and Leuven, Belgium. This geographical diversity allowed recording of a wide range of driving environments, including urban areas, suburban neighbourhoods, and highways. All video recordings were conducted during daylight hours to ensure good visibility.

From this collection of driving footage, a subset of 100 1-minute videos was selected to form the Focus100 dataset. The selection process aimed to maximise variance in driving complexity. To achieve this, we analysed randomly sampled 1-minute sections from the entire dataset and estimated the traffic density based on the total number of vehicle and pedestrian detections. We then selected 20 videos from each quintile of traffic density, ensuring a balanced representation of different traffic conditions within Focus100.

Gaze Data To study natural gaze behaviour in response to realistic driving scenarios, we designed an experiment that simulated the experience of driving while capturing participants' eye movements. This involved presenting participants with a series of engaging driving video clips and asking them to perform a hazard perception task, mirroring the hazard perception component of the UK driving test [12]. This task required participants to actively monitor the videos for potential hazards and respond by pressing the CTRL key whenever they perceived a developing hazard. This approach ensured that participants remained engaged and attentive while providing valuable insights into their natural gaze patterns in response to dynamic driving situations.

Thirty frequent drivers, 14 male and 16 female, with an age range between 21 and 60 years ($M=36.9$, $SD=6.7$), were recruited for this study. All participants had held a valid driver's license for at least three years, had normal vision, and confirmed that they had driven within the past week. Before commencing the study, each participant provided informed consent.

The study was conducted in a controlled laboratory setting. Participants were seated 57 cm from a 24 in Dell P2423 monitor, with the freedom to slightly adjust their position for comfort. A Tobii Pro Nano eye tracker, attached to the lower edge of the monitor, recorded their gaze data at 60 Hz. Participants used a standard Logitech K120 keyboard to provide responses during the hazard perception task.

Before each session, the eye tracker was calibrated to ensure accurate gaze capture for each participant. The participants were then briefed on the purpose and procedure of the study, given practice on the hazard perception task to familiarise themselves with the response mechanism, and asked about their driving history.

During data collection, participants viewed a series of 1-minute egocentric driving video clips. Each participant viewed 30 unique clips and each clip was shown to 7–12



Figure 1. Examples from Focus100. The top row shows diversity in pedestrian traffic, hazardousness, and road type. The bottom row shows the same video frame overlaid with the gaze samples of three separate subjects over the previous 2 s window, where higher alpha of the gaze position corresponds to a more recent sample. The example demonstrates the diversity of temporal gaze patterns across subjects for the same stimuli — information which is lost through averaging in traditional saliency map data representations.

randomly assigned participants, ensuring a balanced representation of individual viewing patterns and responses across the dataset. The order of presentation of the clips was balanced to maintain participant engagement and minimise fatigue. Regular breaks were also incorporated into the session to further combat fatigue and ensure data quality. Due to technical issues during gaze recording, we omit 10 recordings from the dataset, leaving 890 1-minute gaze recordings across 30 subjects.

Hazard Annotations Three annotators labelled and tracked the bounding boxes of objects in the scene that met the definition of a hazard from the UK driving theory test [12]; *A developing hazard is something that would cause you to take action, like changing speed or direction.* The objects were annotated using the CVAT [6] annotation tool. Each hazard was also assigned a *type*: *pedestrian, vehicle, other*; and a *severity level*: *low - preparing to act, or high - take evasive action, e.g. immediate application of the brakes.* On average, 4.08 ± 2.39 hazards were annotated per 60 s sequence (sequences are diverse in hazard counts) and tracked an average for 5.02 s. In total, 207 hazards were low severity and 201 severe; 201 hazards were pedestrians, 203 vehicles, and 4 ‘other’ (e.g., a dog). The labels were accepted by consensus among three annotators.

1.2. Ethics Statement

From the onset, privacy and ethics standards were critical to this data collection effort. The study was conducted in strict accordance with GlimpseML and Toyota Motor Europe institutional research policies. Participants in the gaze collection were fully informed about the purpose, procedures, and potential risks of the study, including the intention to pub-

lish anonymised data for academic research purposes. They were given the opportunity to ask questions and were free to withdraw at any time without consequence. Participants also retained the right to redact their own data at any point before or after publication.

To protect the privacy of individuals in driving videos, all personally identifiable information (PII) has been carefully removed. All detected faces and license plates in the videos were automatically blurred to ensure that individuals and vehicles could not be identified; this was then manually checked frame-by-frame by three annotators. The gaze data provided in the dataset has been processed to remove any information that could potentially identify individual participants. All personal identifiers associated with the gaze data, such as participant names or ID numbers, gender, age, recording locations, and times have been removed.

The Focus100 dataset is stored securely in a GDPR-compliant manner on MFA-protected servers with restricted access within the EU to prevent unauthorised access and ensure data confidentiality. The dataset is restricted to research or academic use only and requires institutional registration for access. Users of the dataset are expected to adhere to ethical research practices and comply with all relevant data privacy regulations, including GDPR. Commercial use is strictly prohibited.

By implementing these measures, we prioritise the privacy and anonymity of all individuals involved, while providing a valuable resource for the research community to advance the study of driver attention and automotive safety.

1.3. Characteristics

Focus100 comprises 100 egocentric driving videos, each 60 seconds in duration, captured at 10 frames per second with

a resolution of 1280×640 px and a 52° field of view. These videos encompass a diverse range of traffic conditions providing rich visual stimuli representative of real-world driving scenarios. See Table 1 for the relevant statistics of the dataset.

Compared to the only comparable in-lab dataset, MAAD [19], a small subset of the DR(eye)VE dataset [37], Focus100 offers significant advantages in terms of scale and diversity. With nearly 15 hours of gaze recordings from 30 participants, Focus100 surpasses MAAD’s 4.83 hours of engaged gaze data, collected from 23 subjects across only 8 videos (all in urban downtown settings). This increased scale translates into a broader representation of driving situations. The distribution of traffic complexity in our new dataset in comparison with the DR(eye)VE and MAAD datasets is shown in Fig 4 of the paper; Focus100 surpasses both in vehicle and pedestrian diversity.

Following [37], we divide the manoeuvres of the ego-car into 4 classes: normal driving, turning left, turning right, and being still (defined as the vehicle being completely stationary or moving slowly). Each frame in the dataset was manually labelled with both the ego-car manoeuvre and the road type. The road types are divided into 5 categories: straight road, intersection, traffic lights, pedestrian crossing, and roundabout. These distributions are visualised in Figure 2.

1.4. Data Format Overview

The Focus100 dataset comprises anonymised driving videos with associated viewer gaze. We also provide object detections extracted from the original videos (non-anonymous):

Video: 100 60-second videos at 10 Hz frame rate, anonymised to remove PII, cropped and downsampled to 1280×640 px.

Gaze: 890 60-second gaze sequences sampled at 60 Hz, mean of left and right eye gaze positions in image space, for at least 7 subjects per video. Each sample is synchronised and associated to a video frame.

Detections: YOLOv8x [25] detections per frame for the following classes: pedestrian, traffic light, stop sign, car, bicycle, truck, motorcycle.

1.5. Discussion

Focus100 offers key advantages over existing driver attention datasets. It provides raw, temporally aligned gaze sequences for fine-grained visual attention analysis and covers diverse driving environments. Each of its 100 videos was viewed by 30 participants, yielding 7–12 gaze recordings per clip. All data collection followed strict privacy and ethical standards. We discuss potential limitations below.

Table 1. Key statistics of MAAD and Focus100 gaze datasets. MAAD collected gaze over a subset of the DR(eye)VE dataset (8 videos in downtown/urban settings). Detections are presented per frame, with mean and standard deviation across the whole dataset. Note that these detections were estimated on downsampled 448×224 and 398×224 image resolutions on Focus100 and MAAD, respectively, matching the resolutions used in our methods.

Category	Measure	MAAD [19]*	Focus100
Video	# Videos	8	100
	Video Length (s)	300	60
	Anonymised	no	yes
	Resolution (px)	1920×1080	1280×640
	FoV ($^\circ$)	unknown	52
	Frequency (Hz)	25	10
Detections	Pedestrian	0.38 ± 1.01	2.59 ± 3.26
	Traffic light	0.34 ± 0.87	0.12 ± 0.48
	Stop sign	0.02 ± 0.13	0.05 ± 0.22
	Car	5.57 ± 3.36	3.17 ± 3.03
	Bicycle	0.06 ± 0.32	1.26 ± 2.09
	Truck	0.26 ± 0.52	0.37 ± 0.71
	Bus	0.06 ± 0.24	0.18 ± 0.48
	Motorcycle	0.02 ± 0.17	0.08 ± 0.35
Gaze	# Subjects	23	30
	Recorded	in-lab	in-lab
	Frequency (Hz)	250	60
	Sub/video (range)	6–11	7–12
	Sub/video (avg)	7.25	8.9
	Subject age	20–55	21–60
	Subject gender	22M–6F	14M–16F
	Total dur (hours)	4.83	14.83
	Licensure (years)	>2	>3

* MAAD collected data across several conditions with distractions or reduced visibility, here we report the statistics for the control condition.

Sampling Frequency A potential concern is the video frame rate (10 fps) and eye-tracker sampling rate (60 Hz) used in Focus100. However, in practice, 10 frames per second is a standard in many autonomous driving datasets and perception stacks, for instance, the Waymo Open Dataset and Euro-PVI camera streams operate at 10 Hz [3, 42], while nuScenes imagery is captured at 12 Hz [4]. This frame rate is sufficient to capture the temporal dynamics of driving manoeuvres and hazard perception, especially since hazard events unfold over several seconds, and allows systems trained on Focus100 to be deployable in such stacks. Similarly, the 60 Hz gaze tracking in our setup provides a sampling rate that is robust for the analysis of fixations, which are the primary correlate of a driver’s perceptual information processing. Prior methodological work shows that fixation-based eye-tracking measures are accurate at 60 Hz, with non-significant difference in fixation detection when downsampling from high-rate data [1, 23]. In driving research, on-road studies often use 60 Hz eye

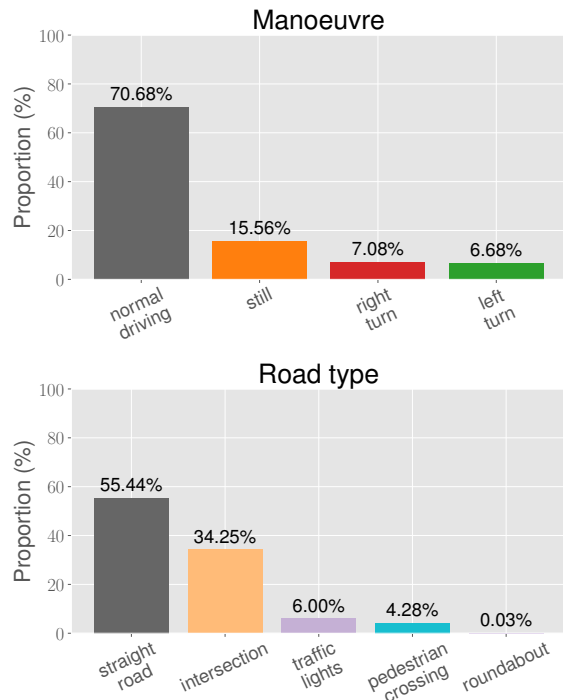


Figure 2. Frame-level distributions of ego-car manoeuvres and road types in the Focus100 dataset.

trackers [41]; with higher sampling rates mainly benefiting micro-saccade analyses [33]. Focus100’s 10 fps video and 60 Hz gaze recording can therefore be considered well-aligned with community norms and sufficient for capturing the phenomena of interest.

Lab-Collected Gaze The ecological validity of lab-based gaze data is a common concern. Differences between passive or semi-passive viewing and active vehicle control are documented; lab protocols remove visuomotor load and can broaden scanning relative to on-road or high-fidelity simulation settings [29, 43]. Controlled comparisons indicate that the magnitude of these differences is small, however, as statistical analyses in [35] report modest effect sizes (though statistically significant) for changes in gaze variance when moving from video-based hazard perception to simulator driving. Critically, lab-collected gaze remains highly informative: it reliably differentiates expert from novice drivers [36], while models trained solely on in-lab data generalise to on-road attention prediction, achieving competitive performance on real driving benchmarks [43]. Focus100 follows this paradigm while emphasising hazardous scenarios and releasing per-subject temporal gaze streams; capturing not only where drivers look but also *when*, not hitherto possible with datasets of this scale, enabling fine-grained tem-

poral analyses of human attention and situation awareness.

Hazards vs. Crashes A reasonable question is whether Focus100’s lack of crash events constrain the scope of our conclusions. While most real-world driving is uneventful, drivers still face situations of varying risk, whereas actual crashes are rare [10, 27]. Crash-focused datasets are invaluable for analysing accident causation, but modelling driver attention and behaviour in the broader context of everyday driving requires coverage of non-crash yet hazardous situations. Attentional failures, such as prolonged off-road glances or mind-wandering, often precede both crashes and near-crashes [27, 39], suggesting shared cognitive mechanisms; near-miss and sub-critical hazardous events therefore serve as effective proxies for studying driver perception and attention in safety-critical contexts [28, 38]. Focus100 does not contain crashes but includes a wide range of situations with hazards of varying severity, capturing both routine and complex driving conditions where attentional demands naturally vary. This coverage complements existing datasets, including DADA-2000 [15], which focuses on crash prediction from in-lab attention data on crowd-sourced crash videos, and BDD-A [43], which uses hard braking events as hazard proxies. By spanning diverse hazards, Focus100 enables the study of driver attention in common critical conditions, complementing existing crash-centric datasets towards applications in automotive safety.

2. Implementation Details

We implemented our model using the *PyTorch* 2.2.1 [2], *PyTorch Geometric* 2.5.0 [16, 17], *PyTorch Lightning* 2.1.3 [13], and *ClearML* 1.14.4 [5] frameworks. Here we report some implementation specifics.

2.1. Gaze Processing

Our method learns using minimally preprocessed gaze data in our gaze-centric scene graphs. Here we report that process, as well as that for converting gaze sequences into fixation and saliency representations.

Preprocessing Our minimal preprocessing stage consisted of linearly interpolating across samples deemed as blinks (as detected by the Tobii tracker), ensuring the temporal continuity of the gaze signal. More specifically, gaze positions during blinks were set to (*NaN*, *NaN*), and the *interp* function from *numpy* [21] was applied independently for the *x* and *y* coordinates to replace all invalid values. The same procedure was applied to the gaze data from the MAAD dataset. This process was carried out in the data’s native gaze sampling frequency (*e.g.* 60 Hz for Focus100), before downsampling by linear interpolation to align with

the desired temporal scene graph frequency (as described in Sec 5.1 of the main paper).

Temporal sampling at 20Hz vs 10Hz Focus100 videos are recorded at 10 fps, while gaze is acquired at a higher native rate with multiple gaze samples per displayed video frame. For scene-graph construction, we represent 1 s windows using 20 timesteps (20 Hz). To align modalities, we upsample the video stream from 10 fps to 20 fps by frame duplication (each video frame is repeated once), and we downsample gaze to 20 Hz so that every timestep contains synchronised traffic-object, road-structure, and gaze nodes.

We choose 20 Hz to support reliable fixation-based evaluation: using a minimum fixation duration of 100 ms, low sampling rates can under-sample short fixations and distort estimated fixation statistics. We therefore ablate the effective gaze sampling rate across Focus100 and observe fixation rates of 1.70 s^{-1} at 60 Hz, 1.64 s^{-1} at 20 Hz, 1.23 s^{-1} at 10 Hz, and 0.59 s^{-1} at 5 Hz, confirming that 10 Hz is inadequate for fixation analysis in our setting. Importantly, the video upsampling is used only for temporal synchronisation with the 20 Hz graph; it does not introduce new visual content beyond the original 10 fps frames.

Postprocessing While our method learns from this minimally processed data, we also implement training-free post-processing to generate fixations and saliency map estimates. An identical process is also used to turn raw ground-truth human gaze sequences into saliency maps for training several baseline saliency estimation approaches.

To detect fixations in gaze sequences we apply the EyeMMV algorithm [31]. EyeMMV is a two-stage, dispersion-based fixation detector (I-DT). Subsets of samples are preliminarily classified as fixations where spatial dispersion remains below a coarse threshold; when this bound is exceeded, the segment is refined with a stricter dispersion threshold to trim edge samples. The candidate is then accepted as a fixation if its duration surpasses a minimum, with inter-fixation intervals labelled as saccades and fixation position defined by the centroid of accepted samples. In our setup we use thresholds $t_0 = 0.08$ and $t_1 = 0.05$ (in normalised image space), and enforce a minimum fixation duration of 0.1 s; detected fixations were additionally manually spot-checked on a subset of trials.

Saliency maps are generated by first accumulating fixations onto a 2D grid matching the spatial resolution of the input frame, where each pixel value represents the number of fixation samples falling at that location (after rounding coordinates to the nearest integer), aggregated across all subjects or generated sequences corresponding to that frame. The resulting discrete fixation map is then smoothed with a Gaussian filter using a standard deviation of $\sigma = 19 \times (w/640)$, where w is the frame width, following [11].

Finally, the saliency map is normalised by its maximum value, yielding intensity values in the range $[0, 1]$.

2.2. Scene Graph Construction

As mentioned in Sec 5.1 of the paper, we used the *YOLOv8x* [25] detector to obtain the object bounding boxes. The appearance features were extracted from the 12th layer of a pretrained *vgg16_bn* network [40] using *ROIAlign* [22], yielding a 128-D appearance vector. The *structure* node is obtained by estimating the drivable-area mask with *YOLOPv2* [20], resizing the mask to 16×8 px, and flattening. Each object’s depth was estimated using *monodepth2* [18] as the mean of inverse disparity within the object’s bounding box.

Input node vectors: The dimensionality of each input node vector used in our experiments is 144: the object’s x and y coordinates (2), its bounding box shape (2), the detector detection score (1), the appearance vector (128), depth estimate (1), and the label one-hot encoding (10; ‘car’, ‘person’, ‘bicycle’, ‘motorcycle’, ‘bus’, ‘truck’, ‘traffic light’, ‘stop sign’, ‘gaze node’, ‘structure node’).

Input edge vectors: The dimensionality of input edge vectors used in the experiments is 5: 3D positional difference between the connecting nodes (3; x , y , depth), timestep difference (1), and cosine similarity between the node appearance vectors (1).

Temporal connectivity: Nodes are connected temporally if the timestep difference between the nodes is included in the set $\mathcal{T}_d = \{1, 2, 4, 8, 16\}$.

2.3. Graph Processor

The Graph Processor processes the input scene graph as described in Sec 3.2 of the paper. Here we provide additional details.

Node embeddings: The dimensionality used for the node-type-specific linear embeddings of the node vectors is $d = 128$. Each node’s timestep is encoded as alternating sine and cosine waves and added to the embedding.

ART block: We use $L = 2$ ART blocks in the Graph Processor in our experiments. An illustration of an ART block is shown in Figure 4. The edge vectors in ART (relative affinity $\mathbf{a}_{i,j}$ in Fig. 3 in the main paper) are embedded into key and value embeddings ($\mathbf{p}_{i,j}^K$ and $\mathbf{p}_{i,j}^V$ in Eqs. (6) and (7)) using two independent MLPs, each implemented as a node-type dependant linear layer followed by *BatchNorm*, a *ReLU*, and another node-type dependant linear layer. Both linear projections output d -dimensional vectors, with $d = 128$. The query, key and value vectors, \mathbf{Q}_i , \mathbf{K}_j , and \mathbf{V}_j , are calculated using a node-type-specific linear layer with a bias, outputting a 3×128 -dimensional vector which is then split into three 128-dimensional vectors.

FFN: The feed-forward network is implemented as two

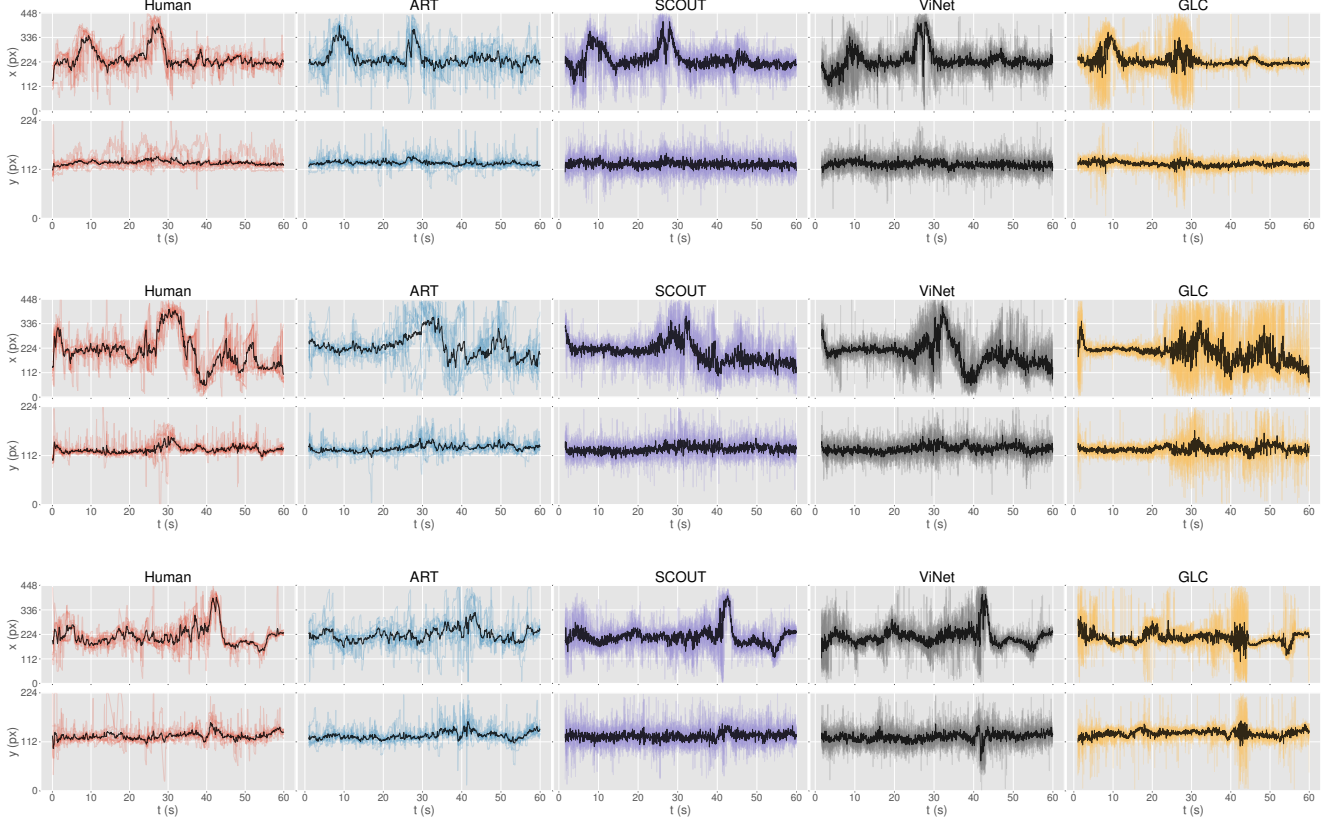


Figure 3. Human gaze sequences compared to generated sequences by ART, SCOUT, ViNet and GLC on 3 videos from the test set. We plot the x and y positions of gaze over time separately, including the y -axis for completeness as it was not shown in the main text. Each line represents a sampled gaze sequence, with the mean gaze sequence shown in black.

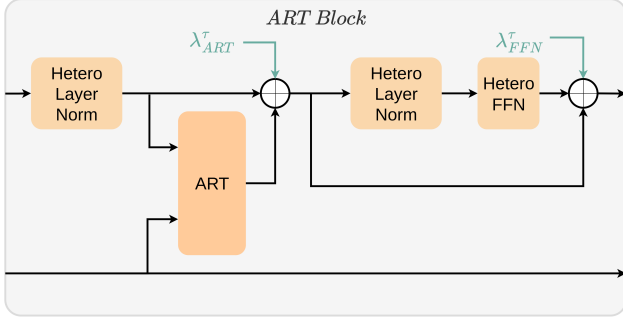


Figure 4. Illustration of the ART block. The block applies LayerNorm, ART attention with a residual connection, another LayerNorm, followed by a two-layer feed-forward network (FFN) with a second residual connection. λ_{ART}^{τ} and λ_{FFN}^{τ} denote the node-type-specific learnable parameters controlling the strengths of the residual connections, $0 \leq \lambda_{ART}^{\tau}, \lambda_{FFN}^{\tau} \leq 1$.

node-type-specific linear layers with biases; the first outputs a 256-dimensional vector, which is passed through a ReLU, and the second linear layer outputs a 128-dimensional vec-

tor. A residual connection with a node-type-specific learnable parameter λ_{FFN}^{τ} is used. See the ART Block illustration in Figure 4.

2.4. Object Density Network

The updated node features belonging to the last timestep in the input spatio-temporal scene graph are fed into the ODN to estimate the parameters of a GMM modelling the future gaze position probability distribution. The parameters predicted by the ODN are listed in Sec 3.3 of the paper and are the output of a node-type-specific linear layer.

3. Qualitative Results

Gaze Sequences The visualisations of sampled gaze in the main paper only show the horizontal position of gaze plotted against time. In Fig. 3 we include a number of plots in which both the x and y dimensions of sampled gaze over time are shown.

2D Sequences Figs. 5 and 6 display more sampled gaze sequences and saliency maps generated by ART and the

baseline models on additional unseen test videos, along with human gaze sequences for comparison. We consistently show that gaze sequences generated by ART closely mimic the human gaze behaviour. We show a failure case of our method in Fig. 7.

Object Saliency Here we explore whether our model produces reasonable estimates of the saliency of objects within images, a task considered in [7] for example. To estimate the saliency ranking of objects in a given frame, we perform 60 gaze sequence simulation runs using our proposed method for a specified video sequence. For each frame, we store the mixing weights estimated by the ODN for each graph node (*i.e.* object detection). We ignore the *structure* and *gaze* nodes as we are only interested in the saliency of individual objects. The average mixing weight for each node in a frame is estimated by summing the mixing weights across all runs for each node, and renormalising them using softmax to account for the removed nodes. As we are interested in ranking the objects within a specific frame, we further divide the mixing weights of nodes in the frame by the maximum mixing weight in that frame. We use YOLOv8x-seg [25] to estimate the segmentation masks for all objects contained as nodes in the graph for a given frame. We overlay the segmentation masks over the input image, assigning them a colour based on the estimated normalised mixing weight. Objects of low saliency rank within an image are shown in blue, and the most salient object(s) is highlighted in red. Example saliency rankings can be seen in Figure 8.

4. Gaze State Dynamics

In this section we analyse the dynamics of gaze state transitions between *saccades* and *fixations*. We identify all timesteps marking the onset of a fixation and observe a time window of 0.5s before and after this point. Each timestep within this window is labelled with a 1 if the gaze at that timestep corresponds to a fixation, or 0 otherwise (*i.e.*, if it was a part of a saccade). We calculate the differences between consecutive elements of the described array, *i.e.* $\mathbf{d}[t] = \mathbf{v}[t+1] - \mathbf{v}[t]$, where we use \mathbf{v} to denote the initial vector of fixations and saccades, \mathbf{d} to mark the vector of differences, and t to index the elements. Each value of the resulting vector will be either -1, 1 or 0, where -1 denotes a change from a fixation to a saccade, 1 marks a change from a saccade to a fixation, and 0 means no state change. Calculating the mean value of all the vectors \mathbf{d}_i , constructed for each fixation in the test set, will give us an empirical expected value of the state change direction for each timestep in the observed window centered around the start of a fixation, $\mathbb{E}(\mathbf{d})$. A more positive value means a higher probability of a saccade-to-fixation state change,

and a more negative value means a higher probability of a fixation-to-saccade state change.

In Figures 9a and 9b we plot this expected value $\mathbb{E}(\mathbf{d})$ as a function of time centered at the beginning of a fixation, estimated using ground truth human gaze samples and samples generated by ART, respectively. We can see that the plots for both the human gaze and our method closely resemble each other. The initial dip preceding the start of a fixation denotes an increase of probability of a state change from a fixation into a saccade; first a saccade needs to happen for a fixation to start, *i.e.* the probability of a saccade needs to increase. The probability of a saccade-to-fixation state change is the highest when the fixation is actually starting, shown at $\Delta t = 0$. This is then followed by another drop, denoting a slightly increased probability of another saccade occurring.

ART/ODN Fixation Mechanism In Figure 9c we plot the gaze node mixing weight as a function of time since the start of a fixation. A high gaze node mixing weight implies a higher probability that the gaze in the next timestep stays at the same location (a part of a fixation), while a lower gaze node mixing weight means an increased probability of a saccade occurring at the next timestep. Notice the resemblance of the plotted shape and the shape of the signal in Figs. 9a and 9b, suggesting the gaze node mechanism for producing fixations worked as intended. As the gaze node mixing weight affects the gaze state in the *next timestep*, the signal appears to be shifted one step to the right compared to the gaze state change dynamics plots in Figs. 9a and 9b.

5. Spectral Analysis of Gaze Variance

To assess whether models reproduce the temporal structure of human inter-observer variability, we analyse the power spectral density (PSD) of residual gaze trajectories relative to each group’s own mean trajectory. For each test sequence s and group $g \in \{\text{Human, ART, SCOUT, VINET, GLC}\}$, we compute the group mean trajectory

$$\bar{\mathbf{p}}_g^{(s)}(t) = \frac{1}{N_g^{(s)}} \sum_{i=1}^{N_g^{(s)}} \mathbf{p}_{g,i}^{(s)}(t), \quad (1)$$

and define residual trajectories

$$\mathbf{r}_{g,i}^{(s)}(t) = \mathbf{p}_{g,i}^{(s)}(t) - \bar{\mathbf{p}}_g^{(s)}(t). \quad (2)$$

We compute the scalar residual magnitude $r(t) = \|\mathbf{r}(t)\|$ and estimate its PSD using Welch’s method. The integral of the PSD corresponds to total within-group residual variance, while its distribution over frequency reflects the temporal organisation of that variance.

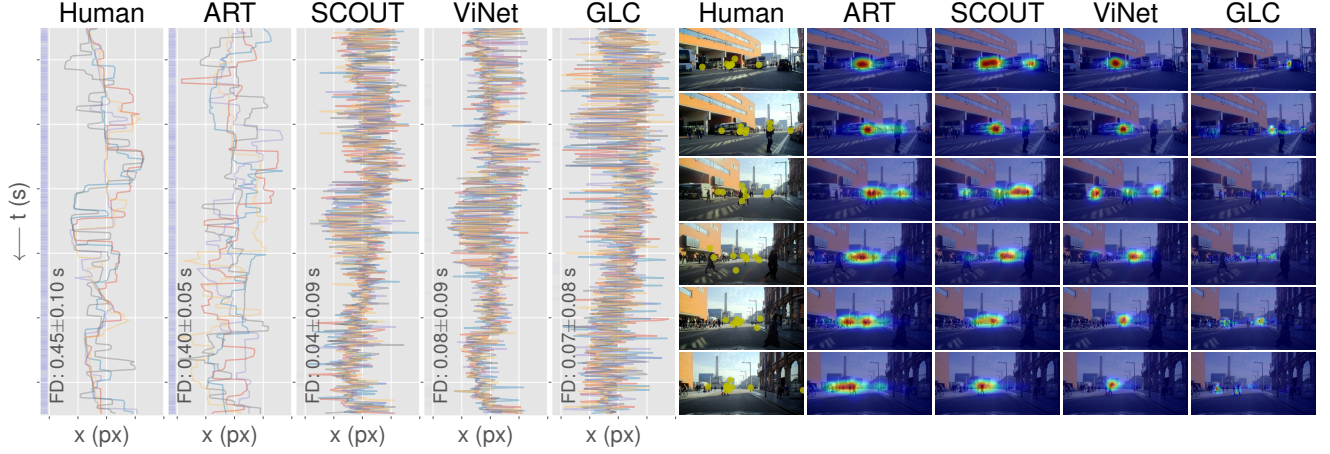


Figure 5. Preceding frames from the same test subsequence show that the largest values in the ART saliency maps are concentrated at positions corresponding to the ground truth gaze points. As shown in the left part of the figure, the gaze dynamics generated by ART closely follow human gaze patterns, whereas the SCOUT, ViNet, and GLC sequences exhibit notably more volatile behaviour.

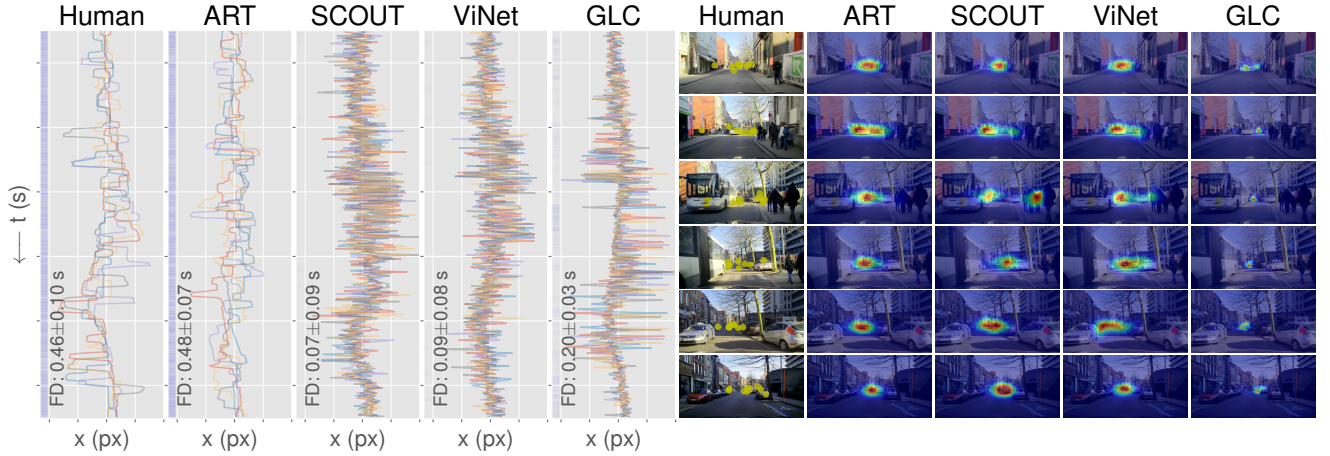


Figure 6. Another example sequence from the unseen test set is presented. Consistent with the previous examples, the samples generated by ART closely follow the ground-truth human gaze dynamics, while the other methods exhibit unsteady and less realistic gaze behaviour throughout the sequence. The saliency maps produced by ART closely reflect the distribution of ground-truth human gaze points. In contrast, GLC produces low-variance saliency maps concentrated near the centre of the road. SCOUT and ViNet generate saliency maps that are qualitatively similar to those of ART, except in the third row, where they highlight the pedestrians on the pavement on the right.

Figure 10 shows the residual PSD averaged across test sequences. Human residuals exhibit a low-frequency-dominated spectrum, indicating that inter-observer variability is dominated by lower-frequency components rather than short-timescale fluctuations. ART closely matches the human spectral profile across frequencies. In contrast, SCOUT, VINET, and GLC show comparatively reduced low-frequency power and flatter spectra, indicating less temporally structured variability.

To summarise spectral structure per sequence, we com-

pute the ratio

$$r_{g,i}^{(s)} = \frac{\int_1^5 \text{PSD}_{g,i}^{(s)}(f) df}{\int_{0.1}^1 \text{PSD}_{g,i}^{(s)}(f) df}, \quad (3)$$

where the lower band (0.1–1 Hz) captures slower residual dynamics and the higher band (1–5 Hz) captures faster fluctuations. The ratio therefore reflects the relative contribution of fast versus slow components of within-group variability. Sequence-level group statistics are obtained by averaging across samples.

Across the 20 test sequences, ART consistently exhibited the smallest deviation from the human spectral ratio.

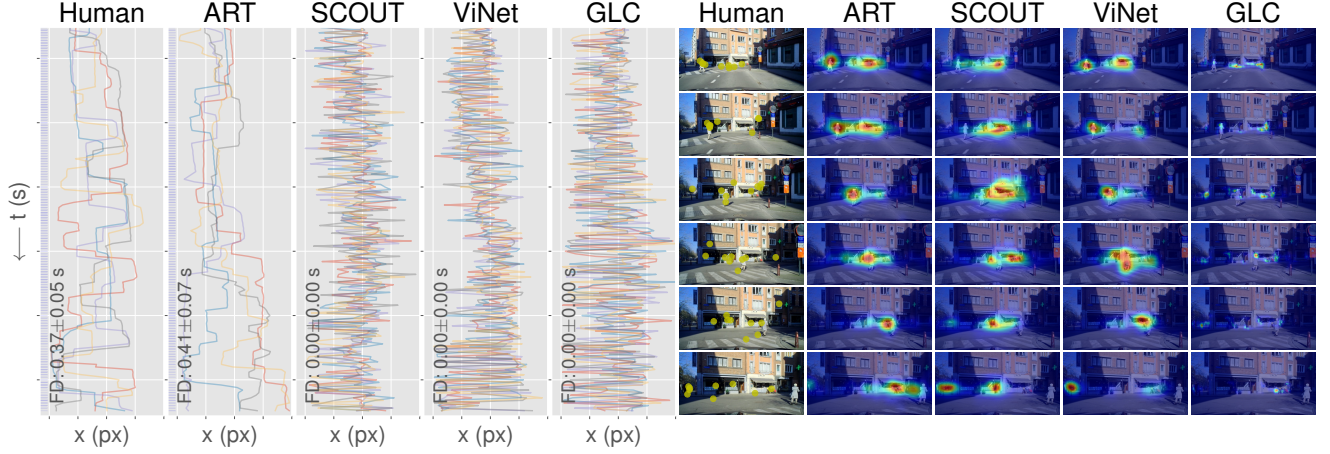


Figure 7. An example of a failure case for our method. The ART gaze sequence samples shown on the left indicate that the simulated gaze tends to follow the pedestrian crossing the road, in contrast to the ground truth human gaze sequences.



Figure 8. Object saliency ranking estimated as the average mixing node weight calculated over 60 ART simulation runs. The node mixing weights are further normalised by the maximum mixing weight in the given frame. Low saliency objects are shown in blue, and the most salient object is shown in red. Note that the colormaps represent rank order and are not consistent across images. See under *Object Saliency* in Section 3 for details.

We performed paired one-sided Wilcoxon signed-rank tests comparing ART against each alternative model under the hypothesis $d_{\text{ART}} < d_{\text{other}}$, with Holm correction for multiple comparisons. All comparisons were significant after correction ($p < 0.001$), with complete directional consistency across sequences.

These results indicate that human inter-observer variability is temporally structured and dominated by slower components. Among evaluated models, ART closely reproduces the spectral organisation of this variance across simulated gaze trajectories on the same sequence, whereas alternative models exhibit significantly different spectral signatures.

6. Latency-Accuracy Trade-Off

Simulations using the method presented in the paper run at an average of 68 ms per frame (15 fps), using input data of resolution of 448×224 px, on a single L40S GPU, including the perception stack, graph construction, ART and ODN. Profiling shows runtime split between perception (43%) and ART+ODN (57%).

We conducted a latency-accuracy trade-off analysis by varying the input resolution of the method, results of which can be seen in Figure 11. The curve shows a consistent accuracy-latency trade-off: increasing input resolution improves predicted gaze likelihood (lower \mathcal{L}_{NLL}) but the gains

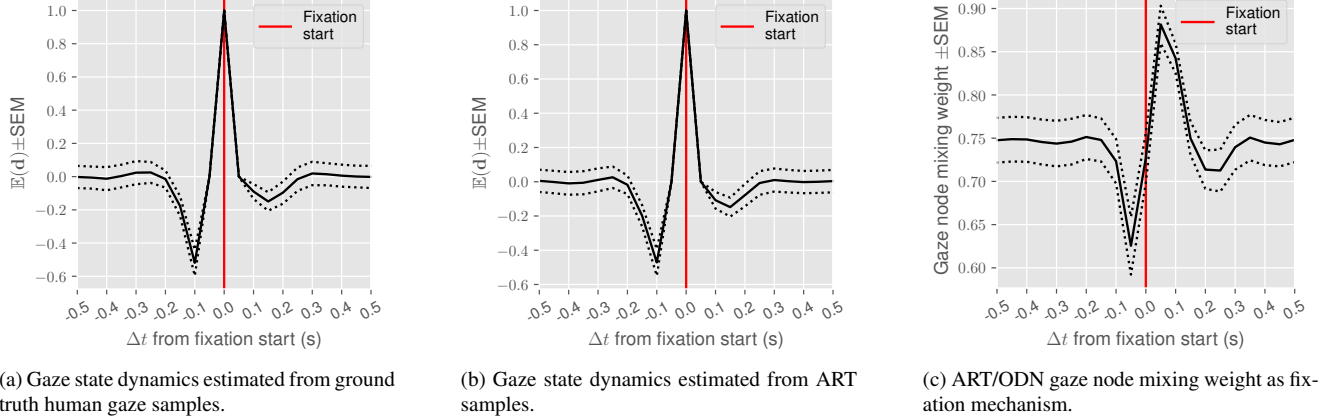


Figure 9. Evolution of the expected value of the gaze state change direction as a function of time relative to the start of a fixation. In Figs. 9a and 9b positive values on the y axis denote a higher probability of the gaze state changing from a *saccade* to a *fixation*, while negative values indicate a higher probability of the reverse. The start of a fixation (at $\Delta t = 0$) is preceded by an increased probability of a saccade (the dip on the left), and followed by another slight increase of the saccade probability. The gaze state change probability dynamics are very similar when estimated for the ground truth **human** gaze samples (Fig. 9a) and using the samples from **ART** (Fig. 9b). In Fig. 9c we show the **model mechanism**, *i.e.* gaze node mixing weight as a function of time relative to a fixation start. High gaze node mixing weight means a high probability of the gaze maintaining the same location in the next timestep, while a lower mixing weight means a higher probability of the gaze changing its location in the next timestep. Note the similarity of this plot to the ones shown in Figs. 9a and 9b. The shift on the x -axis towards the right is due to the gaze node mixing weight affecting the gaze state at the *next timestep*.

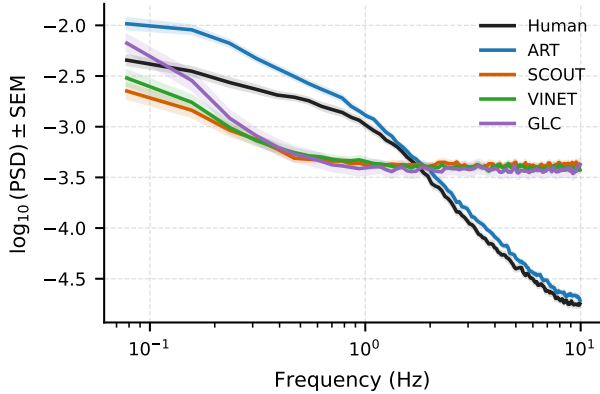


Figure 10. Residual PSD across all shared test sequences. Curves show mean \log_{10} PSD with $\pm \text{SEM}$ across all sequences for Human, ART, SCOUT, VINET, and GLC. ART most closely tracks the human spectral variance profile, consistent with the dynamics results presented in the main text.

taper off as resolution increases. Most of the improvement is achieved when moving from very low to mid resolutions, while further increasing resolution produces relatively smaller additional accuracy benefit. The 448×224 px setting is a good mid-point for real-time analysis, whereas full resolution 1280×640 px, running at 6 fps, is better suited to offline use when latency is less critical.

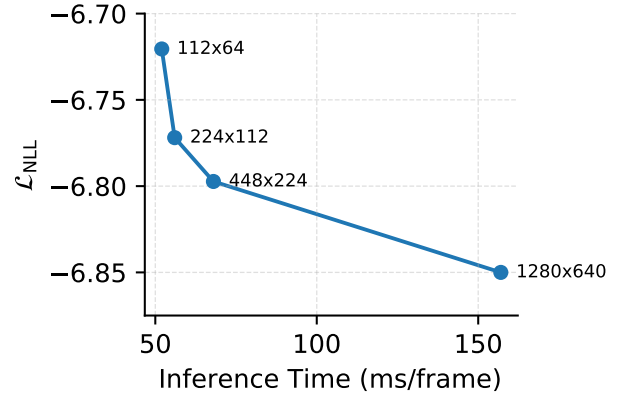


Figure 11. Negative log-likelihood loss across the test set against the time taken to process each frame and generate a next-step gaze position. The used input video resolution is shown next to each data point.

7. Baseline Details

We used publicly available official implementations¹ of GLC [32], SCOUT [30], ViNet [24] and DReyeVENet [37] in our experiments. We give the training and inference details for these methods below.

¹GLC: <https://github.com/BolinLai/GLC/>, SCOUT: <https://github.com/ykotseruba/SCOUT>, ViNet: <https://github.com/samyak0210/ViNet>, DReyeVENet: <https://github.com/ndrplz/dreyeve>.

7.1. Global-Local Correlation (GLC)

Training The *GLC* model [32] is the current state-of-the-art model in egocentric gaze estimation. It is trained on sequences of 8 temporally equidistant square (256×256 px) crops from the input RGB video, which is resized to a height of 256 px while maintaining the aspect ratio. We follow the training procedure from [32] with the *MViT* [14] architecture as the backbone network, initialised with weights pretrained on the *Kinetics-400* [26] dataset. We use a temporal sampling rate of 3 in our experiments, *i.e.* we sample 8 frames from a 22-frame window with equal spacing, and take the last frame’s predicted gaze map as the model output. We train the model for 25 epochs, with a base learning rate set to 5×10^{-5} . We use batch size 16 and run the training on 2 *NVIDIA GeForce RTX 3090* GPUs.

Inference At inference, the original approach only produces gaze probability maps for the central crop of the input; to create our rectangular maps we slide the 256×256 px cropping region horizontally over the rectangular input with a 16 px stride and average the results. To produce sequences using GLC we sample the output gaze probability map for each frame of our sequences.

7.2. SCOUT

Training We train the SCOUT model using the task-free configuration, with an input clip length of 16 frames and image resolution of 224×224 px. The encoder consists of 4 layers with a pretrained and trainable Video Swim Transformer [34] backbone. Training is performed for up to 10 epochs using the Adam optimiser with a learning rate of 1×10^{-4} , a batch size of 4, and early stopping enabled. Learning rate scheduling is applied, and the model achieving the lowest validation loss is used for inference.

Inference Inference also follows the official SCOUT implementation. Given an input sequence of 16 frames from the test set, the predicted saliency map produced by the trained SCOUT model is reshaped to match the size of the ground truth saliency map (448×224 px), it is blurred using a Gaussian kernel of size 11×11 px. We normalise the predicted saliency map by dividing it by its maximum value. This saliency map is used as the prediction for the last frame in the input sequence.

7.3. ViNet

Training We train ViNet without the audio modality with the clip size fixed to 16 frames. The optimiser is Adam with learning rate 1×10^{-4} , batch size 8, and training for up to 40 epochs. Learning-rate scheduling was disabled. The architecture uses the S3D network [44] as the video encoder, pretrained on the *Kinetics-400* [26] action-recognition dataset.

The best model is selected as the one with the lowest validation loss.

Inference We evaluate ViNet the same way as SCOUT; given an input clip of 16 frames, the model predicts a saliency map corresponding to the last frame. This predicted map is blurred by a 11×11 px Gaussian kernel and normalised by dividing it by its maximum value. To obtain saliency predictions across the entire sequence, a sliding window approach is used, generating overlapping 16-frame input clips. This is repeated on the whole test set.

7.4. DReyeVENet

Training We used the official implementation from the DReyeVENet repository in our experiments. Only the image saliency branch was trained, while the optical flow and semantic segmentation branches were excluded, as the pre-trained segmentation model weights were not publicly available. Focusing on the image branch enabled efficient experimentation and ensured a stable, reproducible setup while maintaining a representative subset of the original architecture. The batch size used was set to 4, ‘train samples per epoch’ was set to 8192, and the learning rate was set to 5×10^{-5} . Clips of 16 frames were used as input, normalised by subtracting the mean frame value estimated from the training set. All the frames were resized to 448×448 px to match the original DReyeVENet training setup.

Inference The model with the lowest validation loss was used for evaluation. We follow the official testing code to generate saliency maps for entire sequences in the test set. As with the other baseline methods, a sliding-window approach with a clip size of 16 frames was employed to produce predictions for all frames within each sequence.

8. MAAD Dataset Splits

The MAAD dataset [19] defines only a training and testing split (80% / 20%). In our experiments, we further divide the training set into training and validation subsets (87.5% / 12.5%), ensuring no overlap between any of the splits.

Acknowledgements

This work was funded by Toyota Motor Europe. We thank Catriona Rutter for her assistance with the collection and annotation of the Focus100 dataset.

References

- [1] Richard Andersson, Marcus Nyström, and Kenneth Holmqvist. Sampling frequency and eye-tracking measures: how speed affects durations, latencies, and more. *Journal of Eye Movement Research*, 3(3), 2010. 3
- [2] Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian Puhersch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou, Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM, 2024. 4
- [3] Apratim Bhattacharyya, Daniel Olmeda Reino, Mario Fritz, and Bernt Schiele. Euro-PVI: Pedestrian vehicle interactions in dense urban centers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6408–6417, 2021. 1, 3
- [4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multi-modal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 3
- [5] ClearML. Clearml - your entire mlops stack in one open-source tool, 2024. Software available from <http://github.com/clearml/clearml>. 4
- [6] CVAT.ai Corporation. Computer Vision Annotation Tool (CVAT), 2023. 2
- [7] Bowen Deng, Siyang Song, Andrew P French, Denis Schluppeck, and Michael P Pound. Advancing saliency ranking with human fixations: Dataset models and benchmarks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28348–28357, 2024. 7
- [8] Tao Deng, Kaifu Yang, Yongjie Li, and Hongmei Yan. Where does the driver look? Top-down-based saliency detection in a traffic driving environment. *IEEE Transactions on Intelligent Transportation Systems*, 17(7):2051–2062, 2016. 1
- [9] Tao Deng, Hongmei Yan, Long Qin, Thuyen Ngo, and BS Manjunath. How do drivers allocate their potential attention? Driving fixation prediction via convolutional neural networks. *IEEE Transactions on Intelligent Transportation Systems*, 21(5):2146–2154, 2019. 1
- [10] Thomas A Dingus, Feng Guo, Suzie Lee, Jonathan F Antin, Miguel Perez, Mindy Buchanan-King, and Jonathan Hankey. Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences*, 113(10):2636–2641, 2016. 4
- [11] Yasser Abdelaziz Dahou Djilali, Kevin McGuinness, and Noel O'Connor. Learning saliency from fixations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 383–393, 2024. 5
- [12] Driver and Vehicle Standards Agency. Hazard perception test. <https://www.gov.uk/theory-test/hazard-perception-test>, 2024. Accessed: 2024-08-21. 1, 2
- [13] William Falcon and The PyTorch Lightning team. PyTorch Lightning, 2019. 4
- [14] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835, 2021. 11
- [15] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, and Hongkai Yu. DADA: Driver attention prediction in driving accident scenarios. *IEEE transactions on intelligent transportation systems*, 23(6):4959–4971, 2021. 1, 4
- [16] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019. 4
- [17] Matthias Fey, Jinu Sunil, Akihiro Nitta, Rishi Puri, Manan Shah, Blaž Stojanovič, Ramona Bendias, Barghi Alexandria, Vid Kocijan, Zecheng Zhang, Xinwei He, Jan E. Lenssen, and Jure Leskovec. PyG 2.0: Scalable learning on real world graphs. In *Temporal Graph Learning Workshop @ KDD*, 2025. 4
- [18] Clément Godard, Oisín Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838, 2019. 5
- [19] Deepak Gopinath, Guy Rosman, Simon Stent, Katsuya Terahata, Luke Fletcher, Brenna Argall, and John Leonard. MAAD: A model and dataset for “attended awareness” in driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3426–3436, 2021. 1, 3, 11
- [20] Cheng Han, Qichao Zhao, Shuyi Zhang, Yinzi Chen, Zhenlin Zhang, and Jinwei Yuan. YOLOPv2: Better, faster, stronger for panoptic driving perception. *arXiv preprint arXiv:2208.11434*, 2022. 5
- [21] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser,

- Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020. 4
- [22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 5
- [23] Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. *Eye tracking: A comprehensive guide to methods and measures*. oup Oxford, 2011. 3
- [24] Samyak Jain, Pradeep Yarlagadda, Shreyank Jyoti, Shyampal Karthik, Ramanathan Subramanian, and Vineet Gandhi. ViNet: Pushing the limits of visual modality for audio-visual saliency prediction. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3520–3527. IEEE, 2021. 10
- [25] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO, 2023. 3, 5, 7
- [26] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 11
- [27] Sheila G Klauer, Thomas A Dingus, Vicki L Neale, Jeremy D Sudweeks, David J Ramsey, et al. The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data. Technical report, United States. Department of Transportation. National Highway Traffic Safety . . . , 2006. 4
- [28] Xiaoqiang Kong, Subasish Das, Yunlong Zhang, et al. Patterns of near-crash events in a naturalistic driving dataset: applying rules mining. *Accident Analysis & Prevention*, 161: 106346, 2021. 4
- [29] Iuliia Kotseruba and John K Tsotsos. Data limitations for modeling top-down effects on drivers’ attention. *arXiv preprint arXiv:2404.08749*, 2024. 1, 4
- [30] Iuliia Kotseruba and John K Tsotsos. Understanding and modeling the effects of task and context on drivers’ gaze allocation. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pages 1337–1344. IEEE, 2024. 10
- [31] Vassilios Krassanakis, Vassiliki Filippakopoulou, and Byron Nakos. EyeMMV toolbox: An eye movement post-analysis tool based on a two-step spatial dispersion threshold for fixation identification. *Journal of Eye Movement Research*, 7(1), 2014. 5
- [32] Bolin Lai, Miao Liu, Fiona Ryan, and James M Rehg. In the eye of transformer: Global–local correlation for egocentric gaze estimation and beyond. *International Journal of Computer Vision*, 132(3):854–871, 2024. 10, 11
- [33] Alexander Leube, Katharina Rifai, and Siegfried Wahl. Sampling rate influences saccade detection in mobile eye tracking of a reading task. *Journal of eye movement research*, 10(3): 16, 2017. 4
- [34] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 11
- [35] Andrew K Mackenzie and Julie M Harris. Eye movements and hazard perception in active and passive driving. *Visual cognition*, 23(6), 2015. 4
- [36] Yasaman Habibzadeh Omran, Homayoun Sadeghi-Bazargani, Mohammad Hossein Yarmohammadian, and Golrokh Atighechian. Driving hazard perception tests: a systematic review. *Bulletin of Emergency & Trauma*, 11(2): 51, 2023. 4
- [37] Andrea Palazzi, Davide Abati, Francesco Solera, Rita Cucchiara, et al. Predicting the driver’s focus of attention: the DR(eye)VE project. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1720–1733, 2018. 1, 3, 10
- [38] Thomas Seacrist, Ethan C Douglas, Chloe Hannan, Rachel Rogers, Aditya Belwadi, and Helen Loeb. Near crash characteristics among risky drivers using the SHRP2 naturalistic driving study. *Journal of safety research*, 73:263–269, 2020. 4
- [39] Bobbie D Seppelt, Sean Seaman, Joonbum Lee, Linda S Angell, Bruce Mehler, and Bryan Reimer. Glass half-full: On-road glance metrics differentiate crashes from near-crashes in the 100-Car data. *Accident Analysis & Prevention*, 107: 48–62, 2017. 4
- [40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5
- [41] Jork Stapel, Mounir El Hassnaoui, and Riender Happee. Measuring driver perception: Combining eye-tracking and automated road scene perception. *Human factors*, 64(4): 714–731, 2022. 4
- [42] Pei Sun, Henrik Kretschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 1, 3
- [43] Ye Xia, Danqing Zhang, Jinkyu Kim, Ken Nakayama, Karl Zipser, and David Whitney. Predicting driver attention in critical situations. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14*, pages 658–674. Springer, 2019. 1, 4
- [44] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018. 11
- [45] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 1