
What Moves the Eyes: Doubling Mechanistic Model Performance Using Deep Networks to Discover and Test Cognitive Hypotheses

Federico D’Agostino

Tübingen AI Center, University of Tübingen
federico.dagostino@bethgelab.org

Lisa Schwetlick

EPFL
lisa.schwetlick@epfl.ch

Matthias Bethge*

Tübingen AI Center, University of Tübingen
matthias@bethgelab.org

Matthias Kümmerer*

Tübingen AI Center, University of Tübingen
matthias.kuemmerer@bethgelab.org

Abstract

Understanding how humans move their eyes to gather visual information is a central question in neuroscience, cognitive science, and vision research. While recent deep learning (DL) models achieve state-of-the-art performance in predicting human scanpaths, their underlying decision processes remain opaque. At an opposite end of the modeling spectrum, cognitively inspired mechanistic models aim to explain scanpath behavior through interpretable cognitive mechanisms but lag far behind in predictive accuracy. In this work, we bridge this gap by using a high-performing deep model—DeepGaze III—to discover and test mechanisms that improve a leading mechanistic model, SceneWalk. By identifying individual fixations where DeepGaze III succeeds and SceneWalk fails, we isolate behaviorally meaningful discrepancies and use them to motivate targeted extensions of the mechanistic framework. These include time-dependent temperature scaling, saccadic momentum and an adaptive cardinal attention bias: Simple, interpretable additions that substantially boost predictive performance. With these extensions, SceneWalk’s explained variance on the MIT1003 dataset doubles from 35% to 70%, setting a new state of the art in mechanistic scanpath prediction. Our findings show how performance-optimized neural networks can serve as tools for cognitive model discovery, offering a new path toward interpretable and high-performing models of visual behavior. Our code is available at <https://github.com/bethgelab/what-moves-the-eyes>.

1 Introduction

Every time we view a scene, our eyes produce a sequence of rapid movements (saccades) and brief pauses (fixations), selectively directing our high-resolution fovea to parts of the visual world. This active sampling process—captured in so-called scanpaths—reveals a great deal about how we perceive, attend to, and make sense of our surroundings. Understanding what governs fixation selection is therefore a key question in visual neuroscience and cognitive science and also receives substantial interest from computer vision due to many applications ranging from compression [31] to design and layouting [15, 16].

Indeed, fixation selection has been extensively studied in a variety of scenarios, from reading [65, 23],

*Shared senior authorship

to scene viewing [13, 93, 80, 45], to performing real-world tasks [33, 53, 64]. Here we focus specifically on modeling free viewing of natural scenes.

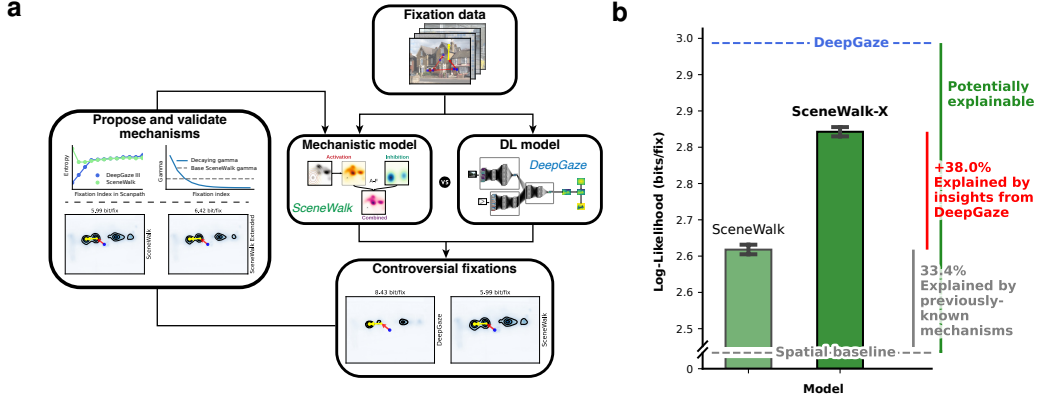


Figure 1: (a) We systematically compare prediction performances of a mechanistic scanpath model (SceneWalk) to a high-performing DNN model (DeepGaze III) to find situations that could be predicted very well but are not by the mechanistic model. Inspecting these extreme cases results in ideas for effects that can be confirmed through further analyses and give rise to mechanisms that are added to the mechanistic model. (b) This process yields three new mechanisms that double SceneWalk’s predictive performance on MIT1003, substantially narrowing the explainability gap.

Approaches to this question in the modeling literature can be viewed as occupying different points on a Pareto front trading off predictive performance with direct mechanistic interpretability. At one end, models like SceneWalk [24, 68, 70] are constructed from a small set of interpretable components that directly reflect cognitive hypotheses about attentional dynamics, oculomotor constraints, and perceptual processing. These approaches align with the view of scientific understanding as a form of compression [87]: the goal is to explain complex phenomena through a small set of generalizable principles, a critical aim even when performance is not maximized. At the other end of the spectrum, performance-oriented deep neural networks (DNNs) like DeepGaze III [51] take a data-driven approach, learning to predict human fixations from large-scale data without explicitly modeling the underlying mechanisms. Their strength lies in establishing a high mark for predictive accuracy, revealing the extent of predictable structure in the data.

As common for the prediction of complex behavior, deep scanpath models indeed outperform mechanistic ones in predicting where people look [44]. But this predictive success comes at a cost: While DNNs can embed mechanisms and offer scientific insights [91, 51], they often capture complex behavioral patterns in high-dimensional parameter spaces that are not immediately transparent [27]. Conversely, mechanistic models, while more transparent, typically under-perform due to their parsimonious parametrization and due to the same strong prior assumptions that make them great testbeds for formulating and evaluating theoretical principles. This leaves a gap: We know that certain behaviors are robustly predictable by data-driven models, but we lack a compact, mechanistic account of them. One flourishing line of research addresses this by starting from the predictive end of the modeling Pareto front, developing more transparent, hybrid, or task-constrained DNNs. In this work, however, we pursue a complementary strategy embracing the mechanistic modeling philosophy: starting from the interpretable end, we ask how much of the performance gap can be closed by systematically improving the fully interpretable model.

A principled way to close this gap, then, is to focus on the mechanistic model’s largest prediction errors. In the domain of static spatial saliency, model failures are often assessed by comparing model predictions to inter-observer consistency, which serves as an empirical upper bound on predictability [49, 14, 12, 45]. However, in high-dimensional distributions such as scanpaths, estimates of inter-observer consistency have been so far intractable: Each fixation is sampled from a conditional distribution $p(f_{i+1}|f_{\leq i}, I)$ that depends on the unique history of previous fixations. Because scanpaths are highly idiosyncratic, the same fixation history is almost never observed more than once, leaving us with a single empirical sample for each high-dimensional distribution. This sparse coverage makes it challenging to form stable, non-parametric estimates of inter-observer agreement.

Here is where a high-performing DNN becomes indispensable. We propose to use DeepGaze III not just as a benchmark, but as a dense, queryable approximation of the true data-generating process [41, 2]. By directly comparing the predictions of DeepGaze III and SceneWalk at the level of individual fixations, we identify *controversial fixations*: Cases where the models strongly disagree. These disagreements highlight specific behavioral patterns captured by the deep model but missed by the mechanistic one. We use them as entry points for analysis, confirming with empirical data whether the behavior is human-like, and then reformulating the missing tendencies as explicit mechanisms within the SceneWalk framework. A key question at the outset was whether the discrepancies between the models would reveal a few systematic, missing principles or a myriad of subtle, complex effects. Interestingly, as we will show, the disagreements clustered into clear, actionable categories, validating the approach. Subsequent model comparisons between versions of the mechanistic model provide a strong test for the newly added mechanisms. This iterative process repurposes the deep model as a tool for mechanistic discovery, helping quantify *how much* specific mechanisms contribute to scanpath behavior and bridging the gap between prediction and understanding. As such, our primary contribution is not the discovery of entirely novel phenomena—which would also have been possible—but a systematic, model-guided methodology for prioritizing which of the many plausible mechanisms most improve prediction on naturalistic data.

We evaluate our method on four free-viewing datasets and demonstrate that augmenting SceneWalk with mechanisms inspired by DeepGaze leads to consistent performance improvements, bringing interpretability closer to predictive performance by halving the explanatory gap on most datasets. In doing so, we offer a roadmap for principled model development that leverages the complementary strengths of hypothesis- and data-driven approaches.

2 Related Work

Scanpath Modeling While the field of eye movement prediction has been focused for a long time mainly on spatial fixation density prediction (“saliency prediction”, [36, 11, 38, 47, 56, 22], see [7, 45] for an overview), the problem of predicting the dependencies within scanpaths of fixations has also attracted substantial interest. The seminal model of Itti and Koch [36] already predicted scanpaths, and since then a range of different mechanistic models have been proposed. Many models take direct inspiration from neuroscience and biology (MASC [1], LATEST [81], Star-FC [89], [40]) or build on statistical ideas such as CLE [5], SaccadicFlow [19], WALD-EM [43] and others [10, 79, 55, 54, 21, 90, 94], and some models implement ideas from cognitive science and attention research, such as SceneWalk [24, 68, 70], Exploration-Exploitation [59] and ROI-LSTM [78]. Finally, the progress in machine learning has brought deep-learning based models that are either only partially interpretable or complete black boxes such as SaltiNet [4], PathGAN [4], DeepGaze III [51], GazeFormer [60], ScanDMM [77] and HAT [92]. Many scanpath models take as given a spatial saliency map of the image, however, there are also models that incorporate more complex interactions of scene content and scanpath dynamics. An extensive comparison [44, 51] showed that SceneWalk is the best mechanistic model in terms of explained information, which is why we are using it as the starting point of our project. While here we focus on free-viewing behaviour, other models also take into account task dependency, such as visual search [1, 34, 91, 71], or inter-subject differences [17, 37].

DNNs as scientific tools One conceptual foundation of our work is the idea that DNNs can serve as queryable empirical stand-ins for Bayesian ideal observers in domains where the analytical solution is intractable [41]. A direct extension of this idea is Scientific Regret Minimization (SRM [2]): using expressive machine learning models as powerful approximators of the data-generating process, leveraging their predictions to identify systematic failures in simpler, hypothesis-driven models. Crucially, this emphasizes cases where we know—because the DNN predicts accurately—that the behavior should have been predictable, but the mechanistic model fails. This reframes the role of DNNs not merely as predictive tools, but as guides for scientific inquiry.

Model comparisons for insight and hypothesis generation Our method also relates to work comparing model predictions to extract insights or derive interpretable mechanisms. Classic approaches include some early work in model distillation, where the goal is to extract simpler, interpretable models from complex DNNs—e.g., distilling neural networks into decision trees [26]. Other work frames model comparison as a way to generate maximally informative contrasts between hypotheses

or models. An example of this is the Maximum Differentiation (MAD) Competition for perceptual models [85], or the use of controversial stimuli to pit vision models against each other in regions of disagreement [30]. These approaches resonate with our method of using disagreement between models to uncover where new mechanisms may be required.

Finally, it is important to distinguish our approach from a different line of research in *post hoc* mechanistic interpretability methods, such as feature attribution [42, 95], which aim to explain black-box models’ internals. In contrast, we treat the DNN as an opaque but predictive reference, and focus instead on refining an interpretable model by learning from its disagreements with the DNN.

3 Methods and Experimental Setup

Modeling Framework and Evaluation Our models and evaluations are formulated in the framework of next-fixation-prediction [34, 45]: Eye movements can be considered a sequential decision process in which each decision is about where to look next. This means that given an image I and a sequence of previous fixations $f_0, \dots, f_i$, models predict the next fixation f_{i+1} as a conditional probability distribution $p(f_{i+1} | f_i, \dots, f_0, I)$, or for short $p(f_{i+1} | f_{\leq i}, I)$. Compared to predicting whole scanpaths, the framework of next-fixation-prediction comes with the advantage that it allows to inspect model predictions for each individual fixation in the dataset and is compatible with classic spatial fixation density prediction: it constitutes the special case $p(f_{i+1} | f_{\leq i}, I) = p(f_{i+1} | I)$ where each fixation is predicted independently of previously made fixations. Model predictions can naturally be scored via log-likelihood $\log p(f_{i+1} | f_i, \dots, f_0, I)$ which we report in bits and relative to the uniform baseline model as $\log p(f_{i+1} | f_i, \dots, f_0, I) - \log p_{\text{uniform}}(f_{i+1} | I)$ (sometimes also termed *information gain*, IG [48]). Information gain has been shown to be a preferable metric due to its sensitive nature [50] and we use it as our main metric of prediction quality. The traditionally often used scanpath similarity metrics [3] are not applicable in our case because they don’t allow single-fixation analyses and have additional conceptual problems [44].

Explainable Information Gain Because information gain constitutes a ratio scale [76], differences and relative changes are meaningful. This forms the basis of *explainable information gain* [48], which quantifies how much of the theoretically possible performance a model achieves. In spatial saliency, this is measured between a lower bound (i.e. center bias) and an upper bound (i.e. gold-standard empirical model), with the difference defining the explainable information gain $\log p(f | I) - p(f)$. In our setting, we are interested in modeling the dependency of fixations on scanpath history. Therefore, the static gold standard from spatial saliency becomes our lower baseline $p(f_{i+1} | I)$. The upper bound, a principled gold standard model of $p(f_{i+1} | f_{\leq i}, I)$, is intractable to estimate non-parametrically from empirical data due to the sparsity of unique scanpath histories. Since each trial generates a unique fixation history, we are typically left with only a single data point for each high-dimensional conditional distribution, precluding direct estimation of inter-observer consistency. We therefore approximate this ceiling with what is to our knowledge the best performance-optimized model for scanpath prediction, DeepGaze III [51], which serves as a queryable proxy for human predictability. As a model, it also enables controlled analyses of the data distribution that sparse empirical observations alone do not permit (see Fig. 5).

Controversial Fixations At the heart of our analysis technique is what we call *controversial fixations*, based on the concept of controversial stimuli [30]. A controversial fixation is a fixation for which the predictions from DeepGaze III and SceneWalk, given the previous scanpath history, differ substantially. We use two different measures to quantify these differences: First, we use the *log-likelihood difference* (LLD) $\log p_{\text{DG3}}(f_{i+1} | f_{\leq i}, I) - \log p_{\text{SW}}(f_{i+1} | f_{\leq i}, I)$. Large values mean that DeepGaze III assigns much more probability to a given fixation than SW, e.g. 8 times as much for a difference of 3 bit. Such differences suggest that while the fixation is predictable, the mechanisms currently implemented by SceneWalk do not capture it well. Additionally, we use the *weighted log-likelihood difference* (WLLD) $p_{\text{DG3}}(f_{i+1} | f_{\leq i}, I) (\log p_{\text{DG3}}(f_{i+1} | f_{\leq i}, I) - \log p_{\text{SW}}(f_{i+1} | f_{\leq i}, I))$. By weighting with DeepGaze’s probability, we select fixations with large LLD that are additionally considered very likely by DeepGaze. We chose these fixation-centric metrics over full-distribution measures like KL-divergence because they focus the analysis on the actually observed fixation and are substantially more computationally efficient. This means that we focus more on cases that DeepGaze considers easy and which might hence showcase effects more clearly.

These controversial fixations serve as entry points for further analysis. Rather than proposing mecha-

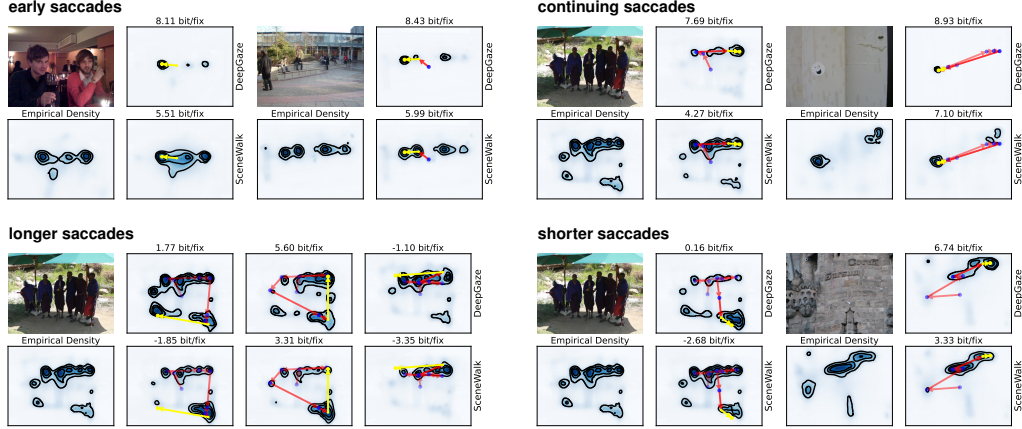


Figure 2: Controversial Fixations: Those fixations in the dataset where SceneWalk loses most performance compared to DeepGaze III in terms of LLD or WLLD. Numbers on top of predictions indicate the achieved log-likelihood relative to a uniform model for the fixation in question.

nisms *ad hoc* from single examples, we use these disagreements to guide targeted comparisons across models and empirical data. Patterns that generalize across fixations are formalized and integrated into the SceneWalk model. This provides a principled alternative to intuition-driven model development, using prediction disagreement to identify which candidate mechanisms are most impactful.

4 Experimental Setup

Datasets We use the MIT1003 dataset [38] for our controversial fixations experiments. It is well established in the gaze prediction community since it is the training set for the MIT300 benchmark dataset [39] of the MIT/Tübingen Saliency Benchmark [46]. It features a broad range of mostly photographic images, including indoor, urban, and natural scenes, which are viewed by 15 observers for 3 seconds each. All training and evaluation on this dataset employs a 10-fold cross-validation scheme across images (with separate training, validation and test splits, see Appendix D), ensuring that models are always tested on stimuli (and their associated fixations) held out during training.

We additionally use the DAEMONS [72], COCO Freeview [18] and Potsdam corpus [69] datasets for additional evaluations of the mechanisms added through our controversial fixations analyses on MIT1003.

Spatial Baseline Our spatial baseline model tries to predict the average spatial fixation density $p(f | I)$ on images, i.e. without taking scanpath dependencies into account. We require it for two reasons: Firstly we need it for putting scanpath model performances into perspective: performance gains compared to the spatial baseline show how well models exploit the scanpath dynamics. Secondly, as is the case for many scanpath models, the SceneWalk model requires as input a spatial saliency map: a map encoding which image areas are interesting independent of scanpath dynamics. Consistent with previous applications of the model [24, 70] we use our spatial baseline model to that end. Our spatial baseline model is a mixture model of a KDE of empirical fixations and predictions from a SOTA saliency model, where the parameters are fitted per image. For details see Appendix C.

SceneWalk The SceneWalk model [24, 68, 70] is a mechanistic scanpath model inspired by results from cognitive science. It implements an activation map and an inhibition map, which evolve over time depending on the fixated image locations. The activation map is updated for each fixation with a Gaussian window on the saliency map and decays over time, encoding what the model knows about relevant parts of the image. The inhibition map is updated for each fixation with a Gaussian at the fixation location and also decays over time, encoding which image areas the model has already explored. To predict the next saccade target, activation and inhibition maps are postprocessed with an exponential nonlinearity and combined to yield the fixation selection map. Here, we use the latest version of SceneWalk [70] which additionally includes a peri-saccadic attention shift mechanism:

Prior to a saccade, attention is already shifted towards the saccade target and hence, after the saccade, a small area in the direction of the saccade is uncovered. This mechanism results in higher prediction performance and explains the directional statistics of scanpaths better. SceneWalk has only about 20 parameters, which we fit using MAP estimation (for details see Appendix E).

DeepGaze III DeepGaze III [51] is a deep learning based scanpath model. For predicting the next fixation in a scanpath, it takes as input the viewed image and the last four fixation locations. The input image is encoded with a pretrained backbone and processed by a priority network that decodes a spatial priority map. This spatial priority map is then combined with the information about previous fixation locations in a fixation selection network to output a priority map that is conditioned on the previous scanpath history. This map is blurred, combined with a center bias and passed through a softmax to yield the probability distribution for the upcoming fixation location. For better comparability with SceneWalk, here we removed the priority network and instead used the same empirical densities that are also used by SceneWalk in the fixation selection network. This serves to make sure that all differences between the two models are necessarily due to how they combine the spatial priority map with the fixation history and avoids confounders. We fit DeepGaze III in this setting on the MIT1003 dataset using 10-fold crossvalidation across images (see Appendix E).

5 Experiments and Results

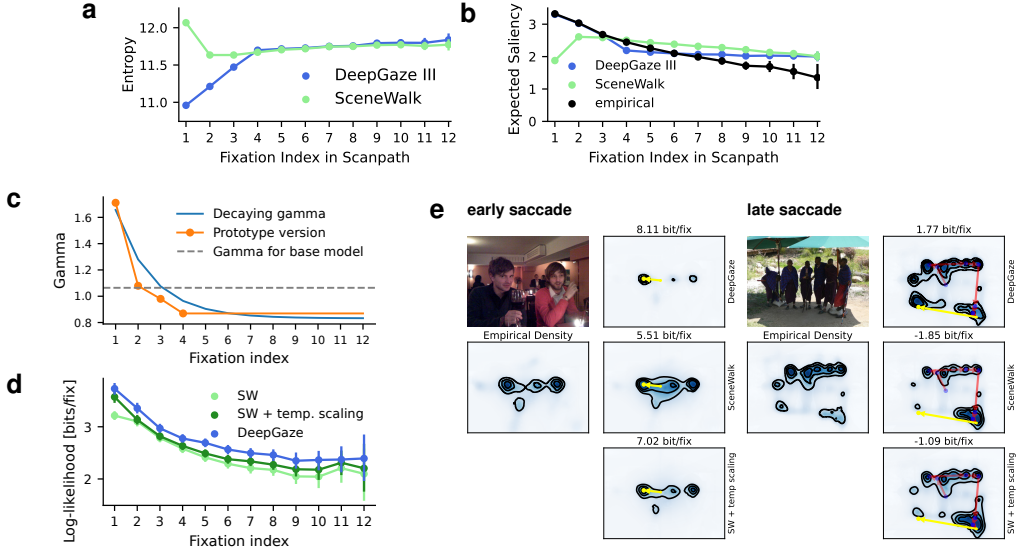


Figure 3: Time dependent temperature scaling. We found that DeepGaze shows higher confidence (less entropy) early on (a) and predicts fixations in more salient locations (b) in agreement with the empirical data. SceneWalk is not showing this effect. To address this, we introduced fixation-index-dependent temperature scaling (c), modeled with exponential decay (d), which improves predictions for both early and late fixations (e).

5.1 Model Comparison

Explainable information By evaluating our spatial baseline, DeepGaze and SceneWalk on the held-out test cross-validation splits of the MIT1003 dataset, we find that DeepGaze explains 0.43 bit/fix more than the spatial baseline. From estimate of the explainable information gain, the SceneWalk model explains with 0.15 about one third, leaving an explanatory gap of two thirds.

Controversial Fixations After the full-dataset level analysis, we now go to the level of individual fixations and inspect those fixations where SceneWalk shows the largest failures. For both log-likelihood difference and weighted log-likelihood difference, we select the top six images with largest mean score and from them visualize the six fixations with highest individual scores. We find that

nearly all (92%, see Appendix J) of these controversial fixations can be classified into four categories (Figure 2): DeepGaze III seems to gain by predicting that (1) early saccades go exclusively to the most salient targets and that (2) often saccades tend to continue in the direction of the previous saccade. In addition, DeepGaze gains by having a better understanding of when saccades are (3) longer or (4) shorter.

5.2 New Mechanisms in SceneWalk

Time Dependent Temperature Scaling Controversial fixations revealed that DeepGaze favors high-saliency targets early in scanpaths, with this effect diminishing over time (Fig. 2). We confirmed that DeepGaze indeed shows higher confidence (less entropy) early on (Fig. 3a) and predicts fixations in more salient locations in agreement with the empirical data (Fig. 3b). SceneWalk partially captures this effect thanks to the buildup of the inhibitory stream over time which forces the model to increasingly attend to less salient locations. However, the effect in SceneWalk is clearly weaker than in DeepGaze and very off for the first fixation. The original SceneWalk model already uses exponents to shape the activation and inhibition maps to be more or less deterministic, i.e. emphasize highly activated areas. This essentially regulates the temperature of the distributions. To model the change in the focus on salient locations, we now made this exponent in SceneWalk dependent on the fixation index in the scanpath (Fig. 3c). We first used different independent exponents for earlier and later fixations (“prototype version” in Fig. 3c). These exponents closely resembled an exponential decay and hence for the final version we fitted the exponent as an exponential decay over the course of a scanpath (details in Appendix B.1). The result is a strong early preference for salient locations, with the exponent decaying below the baseline model’s value around the third fixation, allowing more exploratory behavior later in viewing. This improves prediction performance both for early and late fixations (Fig. 3d) and also visibly improves the behavior on the corresponding controversial fixations (Fig. 3e). Interestingly, it has been hypothesized before that the temperature of the fixation distribution might change over presentation time [69].

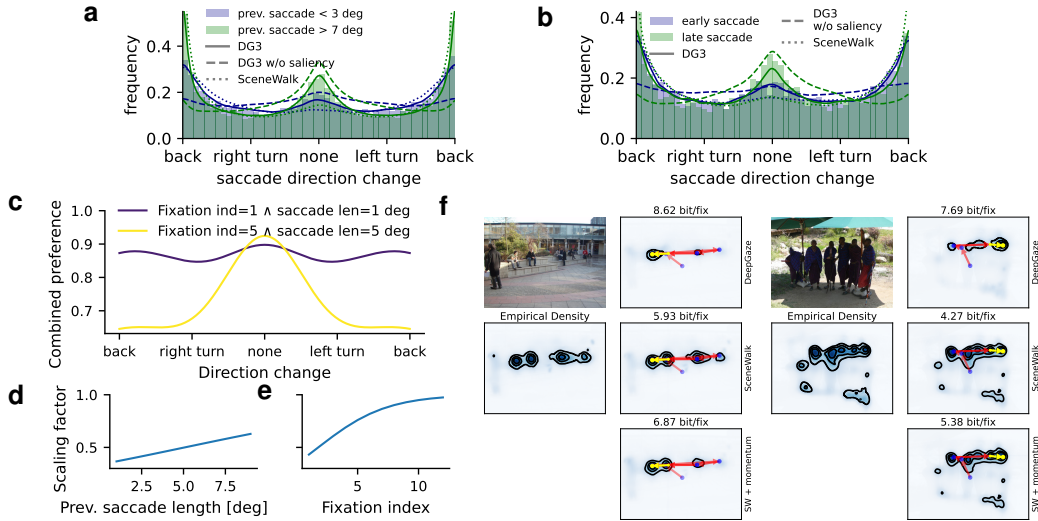


Figure 4: Saccadic momentum: Controversial fixations suggested that DeepGaze sometimes prefers saccade to continue in the same direction. We confirmed such a *saccadic momentum* effect especially after long saccades (a) and later in scanpaths (b). These effects remain even after controlling for the distribution of salient objects (dashed lines), indicating a genuine directional bias. We added a saccadic momentum and return mechanism to SceneWalk (c) modulated by previous saccade length (d) and fixation index (e), improving predictions for "ongoing saccades" controversial fixations (f).

Saccadic Momentum A second prominent cluster of controversial fixations clearly indicated that DeepGaze more accurately predicted short, continuing fixations mostly following long saccades. We found the effect to be consistent with known phenomena of saccadic momentum [74, 75, 88], which have also been previously used in scanpath models [40]. We noticed that this behavior is

particularly pronounced after long saccades [75] and during later stages of the scanpath (Fig.4a,b). While some of this structure could be explained by saliency alone, control analyses using DeepGaze evaluated on uniform empirical fixation densities (i.e., with saliency effectively removed) revealed that the momentum effect largely persists (Fig.4d, dashed lines), suggesting it reflects some form of oculomotor bias rather than image content. Besides momentum, both DeepGaze and the empirical data exhibited a balance between returning to previously attended regions and continuing in the same direction, especially during early viewing and following short saccades.

SceneWalk’s perisaccadic attention shift can partially explain such behavior, but much more weakly than in the data (see Appendix Fig 7 for a larger figure). To address this, we introduced a dynamic oculomotor bias map that captures both saccadic momentum and return tendencies. The mechanism modulates directional preferences based on the length of the previous saccade and the position within the scanpath and applies the resulting bias map to SceneWalk’s internal probability map (details in Appendix B.2). Following short saccades early in a scanpath, the model learns to balance return saccades and momentum, while after long saccades and later in viewing, it favors continued movement in the same direction (Fig. 4c). This extension significantly improves SceneWalk’s ability to capture directional patterns in controversial fixations, bringing its predictions closer to those of DeepGaze (Fig. 4f). Again, this mechanism aligns with prior literature suggesting that saccadic momentum and return effects may jointly reflect oculomotor constraints or facilitation-of-return mechanisms [74, 88, 67], and even more complex dynamics beyond muscle physics [62].

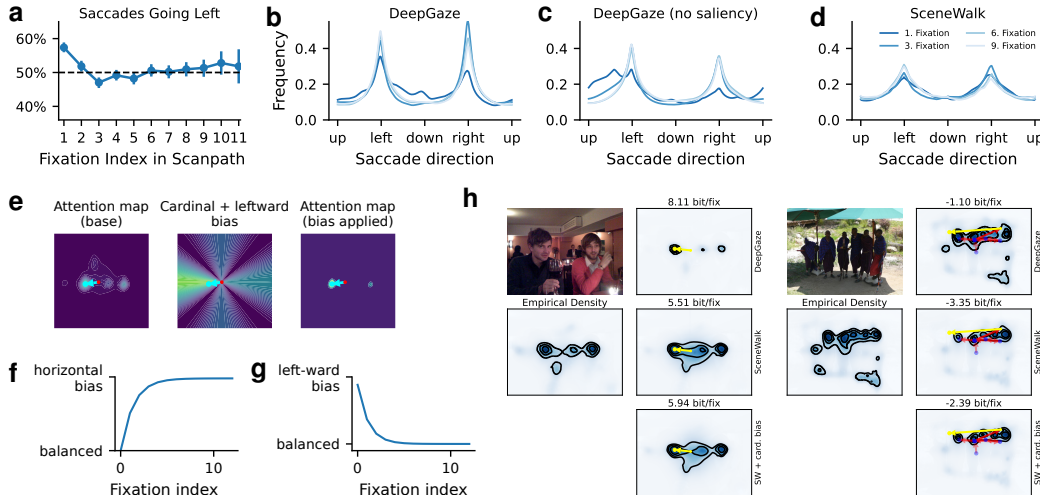


Figure 5: Horizontal and left-wards attention bias: Controversial fixations suggested that DeepGaze predicts early saccades to go to the left. This is indeed the case in the data (a) and the DeepGaze predictions (b). The effect persists when DeepGaze is run with uniform saliency maps (c) while SceneWalk (d) shows a too weak effect that increases over time. Therefore, we added a cardinal attention bias to SceneWalk, than can additionally have a left-asymmetry (e) and can adapt over time (f, g). This improves predictions on the relevant controversial fixations (f).

A horizontal and left-wards attentional bias Especially for early controversial fixations, DeepGaze seemed very confident about left-wards saccades (Figure 2). We found his effect indeed present in the data (Fig. 5a), as previously reported elsewhere [25]. We tested how well the models reflected the effect by computing, for each individual fixation, a histogram of expected saccade directions and averaged them for earlier and later fixations. This confirmed that DeepGaze learned a left-wards and more generally a horizontal bias (Fig. 5b). Importantly, the bias is still there when we evaluate DeepGaze with uniform saliency maps (Fig. 5c), confirming that the effect is not not purely saliency-driven. Interestingly, DeepGaze’s leftward bias fades over time, while its horizontal bias builds up across early fixations. SceneWalk, in contrast, shows only weak directionality and an incorrect increase in leftward bias over time (Fig. 5d), possibly reflecting artifacts from the empirical densities.

We decided to account for these directional biases by applying a variable cardinal prior to SceneWalk’s attention map before saccade selection. This modification is motivated by two complementary observations. First, a purely oculomotor bias—like the one implemented by [40]—is unlikely to explain the found preference for early saccades to the left or the weaker horizontal bias for early fixations even in the absence of saliency. Second, beyond known effects of cardinal oculomotor biases, empirical evidence suggests that attentional processes themselves exhibit anisotropies across the cardinal axes—particularly favoring horizontal over vertical directions [83, 57, 82]. Notably, attentional modulation has been observed as early as the lateral geniculate nucleus [57] lending biological plausibility to an attentional bias operating in concert with oculomotor constraints.

Concretely, our cardinal attention bias is implemented as a smooth function of saccade angle, with preference peaking along the cardinal axes and decaying with distance from the current fixation. Additionally, we model a dynamic asymmetry between horizontal and vertical directions, with the horizontal bias gradually increasing over the course of a trial (Fig. 5f). Finally, to account for the early leftward bias, we also introduced an asymmetry that favors leftward over rightward saccades during early fixations, with strength that decays over time (Fig. 5g). The resulting cardinal attention bias map is combined multiplicatively with the attention map (Fig. 5e). For details see Appendix B.3.

5.3 Results

A new SOTA for mechanistic scanpath prediction On the MIT1003 dataset, our full extended SceneWalk model (which we term SceneWalk-X for short), doubles the performance of the original SceneWalk perisaccadic model compared to the spatial baseline. It further closes 56% of the remaining gap in explainable performance as estimated by DeepGaze. Since SceneWalk was the previous SOTA on MIT1003 in mechanistic scanpath prediction [51], this sets a new SOTA in mechanistic scanpath prediction. Notably, we evaluated SceneWalk-X also on the DAEMONS dataset [72], which is specifically designed for scanpath prediction and includes longer scanpaths on high-resolution natural scenes. SceneWalk performs better when trained and evaluated on DAEMONS, where it already explains more than 60% of the explainable information relative to the baseline. SceneWalk-X closes 40% of the residual gap in explainable performance between DeepGaze and the base SceneWalk model. While this is smaller than the gains observed on MIT1003, it is worth noting that DAEMONS contains longer, more complex scanpaths where the spatial baseline alone is less predictive, and the original SceneWalk model already performs well—consistent with its design focus on extended viewing behavior. Nonetheless, SceneWalk-X still nearly halves the remaining gap to DeepGaze, demonstrating its robustness and leaving room for future improvements specific to this richer dataset.

More evaluations In Appendix A we confirm our results hold on two more datasets (COCO Freeview [18] and the Potsdam Scene Viewing Corpus [69]) and compare with other mechanistic models. We further report AUC scores for all datasets to show that our results do not depend on the chosen evaluation metric.

Relevance Assessment To quantitatively assess the overall relevance of each added effect, we evaluated different extensions of SceneWalk and also compared to the earlier “classic” SceneWalk [24, 68] without perisaccadic attention shifts (Fig. 6). We find that on MIT1003, each of the three mechanisms contributes about equally to the overall increase in explanatory performance. On the DAEMONS dataset, the performance gain is mostly due to momentum and cardinal biases, while temperature scaling contributes little. We hypothesize that this might be because DAEMONS features much longer scanpaths, and the temperature stays mostly constant after the first few fixations.

Reconciling controversial fixations In Appendix J, we show all examined controversial fixations together with the predictions from the fully extended SceneWalk model and find that many of them are now predicted substantially better.

6 Discussion

Data-driven (DNN) and hypothesis-driven (mechanistic) approaches to modelling cognitive and perceptual processes are usually painted as being in contrast with each other. One aims primarily to predict, the other primarily to understand and confirm hypotheses. In this work we demonstrate how

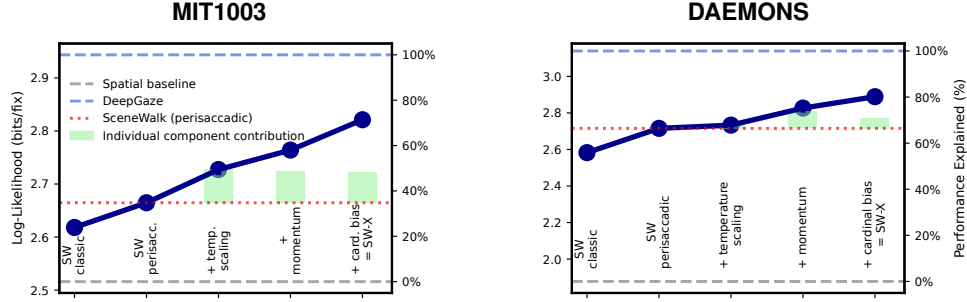


Figure 6: Contributions of the different added mechanisms on the MIT1003 and DAEMONS datasets.

both techniques can be used in concert to improve our understanding of the mechanisms underlying fixation selection. By focusing on fixations where the predictions of a deep model (DeepGaze III) and a mechanistic model (SceneWalk) diverged, we isolated specific mechanisms implicitly captured by the deep model but absent in SceneWalk. These mechanisms were then formalized and incrementally implemented into SceneWalk. Each addition was tested via its impact on predictive likelihood, thereby providing a strong test of its behavioral relevance.

The conceptual cornerstone of our method is to use a DNN model as a stand-in for predictability. This turns a black-box model into a scientific instrument: not something we trust blindly, but something we can probe, compare against, and learn from. This shift circumvents many of what are portrayed as the classic weaknesses of deep learning and focuses instead on the strenghts of such models. For example, DeepGaze III allows us to test whether certain spatial or temporal biases affect behavior beyond what is already explained by bottom-up saliency—a manipulation that would be extremely difficult, if not impossible, to achieve with empirical data alone. In this sense, deep models can offer novel ways to look at the underlying data distribution. When combined with direct comparisons to mechanistic models, this enables to more flexibly look at data and formulate hypotheses.

Importantly, the main contribution of our method is not necessarily in discovering novel mechanisms. Many of the effects we identify, have been observed before, and we mainly add nuance to the way they might be operating, which will require future work to confirm. Rather, our main contribution here lies in offering a principled procedure for identifying which among the many plausible mechanisms are most useful for improving behavioral predictions. By translating prediction errors into candidate explanations and rigorously testing them in a mechanistic framework, our method strengthens the link between predictive performance and theoretical insight.

Limitations and outlook A few potential limitations and subtleties of our approach warrant further discussion. First, our method assumes that DeepGaze provides a reasonable proxy for human fixation behavior, but, like any DNN, it may learn dataset-specific biases or shortcuts [28, 27]. Second, we acknowledge that adding a mechanism which improves the mechanistic model’s fit does not necessarily imply that the same mechanism is used by humans, as it could be a correlate or computational proxy. While mechanistic models aim for interpretability, they are still approximations and subject to their own inductive biases. Grounding mechanisms in empirical findings helps address this limitation. Finally, our current pipeline relies on manual identification and implementation of mechanisms, a deliberate choice to validate the methodology in a transparent, end-to-end fashion. We view this work as a foundational proof of concept. Automating this search over candidate model extensions, especially for more complex interactions across time and context, remains an open and exciting challenge [29] that points toward a generalizable, semi-automated modeling pipeline—a step towards what could be seen as an “automated scientist” [e.g., 58] for behavioral modeling. Similarly, our results are based on scanpaths over natural images; applying this framework to other domains such as reading, video, or real-world interactions is an exciting direction for future research.

Broader impact Accurate scanpath models have practical relevance in fields such as visual design, education, and assistive technology. However, there are ethical considerations: predictive models of attention could be used to manipulate user behavior, for example in advertising or UX design. It is therefore important that such models are developed transparently and deployed with caution.

Acknowledgments

The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Federico D’Agostino. This work was supported by the German Research Foundation (DFG): SFB 1233, Robust Vision: Inference Principles and Neural Mechanisms, project number: 276693517.

References

- [1] H. Adeli, F. Vitu, and G. J. Zelinsky. A Model of the Superior Colliculus Predicts Fixation Locations during Scene Viewing and Visual Search. *The Journal of Neuroscience*, 37(6):1453–1467, Feb. 2017. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.0825-16.2016.
- [2] M. Agrawal, J. C. Peterson, and T. L. Griffiths. Scaling up psychology via Scientific Regret Minimization. *Proceedings of the National Academy of Sciences*, 117(16):8825–8835, Apr. 2020. doi: 10.1073/pnas.1915841117.
- [3] N. C. Anderson, F. Anderson, A. Kingstone, and W. F. Bischof. A comparison of scanpath comparison methods. *Behavior Research Methods*, 47(4):1377–1392, Dec. 2015. ISSN 1554-3528. doi: 10.3758/s13428-014-0550-3.
- [4] M. Assens, X. Giro-i-Nieto, K. McGuinness, and N. E. O’Connor. PathGAN: Visual Scanpath Prediction with Generative Adversarial Networks. In L. Leal-Taixé and S. Roth, editors, *Computer Vision – ECCV 2018 Workshops*, Lecture Notes in Computer Science, pages 406–422, Cham, 2019. Springer International Publishing. ISBN 978-3-030-11021-5. doi: 10.1007/978-3-030-11021-5_25.
- [5] G. Boccignone and M. Ferraro. Modelling gaze shift as a constrained random walk. *Physica A: Statistical Mechanics and its Applications*, 331(1):207–218, Jan. 2004. ISSN 0378-4371. doi: 10.1016/j.physa.2003.09.011.
- [6] G. Boccignone and M. Ferraro. Ecological Sampling of Gaze Shifts. *IEEE Transactions on Cybernetics*, 44(2):266–279, Feb. 2014. ISSN 2168-2275. doi: 10.1109/TCYB.2013.2253460.
- [7] A. Borji and L. Itti. State-of-the-Art in Visual Attention Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207, Jan. 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2012.89.
- [8] A. Borji and L. Itti. CAT2000: A Large Scale Fixation Dataset for Boosting Saliency Research. *CVPR 2015 workshop on "Future of Datasets"*, 2015.
- [9] J. Bradbury, R. Frostic, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang. JAX: Composable transformations of Python+NumPy programs. jax-ml, May 2025.
- [10] D. Brockmann and T. Geisel. The ecology of gaze shifts. *Neurocomputing*, 32–33:643–650, June 2000. ISSN 0925-2312. doi: 10.1016/S0925-2312(00)00227-7.
- [11] N. D. B. Bruce and J. K. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3):5–5, Mar. 2009. ISSN 1534-7362. doi: 10.1167/9.3.5.
- [12] N. D. B. Bruce, C. Wloka, N. Frosst, S. Rahman, and J. K. Tsotsos. On computational modeling of visual saliency: Examining what’s right, and what’s left. *Vision Research*, 116, Part B:95–112, Nov. 2015. ISSN 0042-6989. doi: 10.1016/j.visres.2015.01.010.
- [13] G. T. Buswell. *How People Look at Pictures*. The University of Chicago Press, 1935.
- [14] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand. Where Should Saliency Models Look Next? In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 809–824. Springer, Cham, Oct. 2016. ISBN 978-3-319-46453-4 978-3-319-46454-1. doi: 10.1007/978-3-319-46454-1_49.
- [15] Z. Bylinskii, N. W. Kim, P. O’Donovan, S. Alsheikh, S. Madan, H. Pfister, F. Durand, B. Russell, and A. Hertzmann. Learning Visual Importance for Graphic Designs and Data Visualizations. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, UIST ’17, pages 57–69, New York, NY, USA, Oct. 2017. Association for Computing Machinery. ISBN 978-1-4503-4981-9. doi: 10.1145/3126594.3126653.
- [16] G. Cartella, M. Cornia, V. Cuculo, A. D’Amelio, D. Zanca, G. Boccignone, and R. Cucchiara. Trends, Applications, and Challenges in Human Attention Modelling. Apr. 2024.

- [17] X. Chen, M. Jiang, and Q. Zhao. Beyond Average: Individualized Visual Scanpath Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25420–25431, 2024.
- [18] Y. Chen, Z. Yang, S. Chakraborty, S. Mondal, S. Ahn, D. Samaras, M. Hoai, and G. Zelinsky. Characterizing Target-Absent Human Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5031–5040, 2022.
- [19] A. D. F. Clarke, M. J. Stainer, B. W. Tatler, and A. R. Hunt. The saccadic flow baseline: Accounting for image-independent biases in fixation behavior. *Journal of Vision*, 17(11):12–12, Sept. 2017. ISSN 1534-7362. doi: 10.1167/17.11.12.
- [20] D. Cousineau. Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson’s method. *Tutorials in Quantitative Methods for Psychology*, 1(1):42–45, Sept. 2005. ISSN 1913-4126. doi: 10.20982/tqmp.01.1.p042.
- [21] A. Coutrot, J. H. Hsiao, and A. B. Chan. Scanpath modeling and classification with hidden Markov models. *Behavior Research Methods*, 50(1):362–379, Feb. 2018. ISSN 1554-3528. doi: 10.3758/s13428-017-0876-8.
- [22] R. Droste, J. Jiao, and J. A. Noble. Unified Image and Video Saliency Modeling. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 419–435, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58558-7. doi: 10.1007/978-3-030-58558-7_25.
- [23] R. Engbert, A. Longtin, and R. Kliegl. A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision Research*, 42(5):621–636, Mar. 2002. ISSN 0042-6989. doi: 10.1016/S0042-6989(01)00301-7.
- [24] R. Engbert, H. A. Trukenbrod, S. Barthelmé, and F. A. Wichmann. Spatial statistics and attentional dynamics in scene viewing. *Journal of Vision*, 15(1):14–14, Jan. 2015. ISSN 1534-7362. doi: 10.1167/15.1.14.
- [25] T. Foulsham, A. Gray, E. Nasiopoulos, and A. Kingstone. Leftward biases in picture scanning and line bisection: A gaze-contingent window study. *Vision Research*, 78:14–25, Jan. 2013. ISSN 0042-6989. doi: 10.1016/j.visres.2012.12.001.
- [26] N. Frosst and G. Hinton. Distilling a Neural Network Into a Soft Decision Tree, Nov. 2017.
- [27] C. Funke, J. Borowski, T. Wallis, W. Brendel, A. Ecker, and M. Bethge. Comparing the ability of humans and DNNs to recognise closed contours in cluttered images. *Journal of Vision*, 18(10):800–800, Sept. 2018. ISSN 1534-7362. doi: 10.1167/18.10.800.
- [28] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, Nov. 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z.
- [29] A. Ghosh, S. Dziadzio, A. Prabhu, V. Udandara, S. Albanie, and M. Bethge. ONEBench to Test Them All: Sample-Level Benchmarking Over Open-Ended Capabilities, Dec. 2024.
- [30] T. Golan, P. C. Raju, and N. Kriegeskorte. Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, 117(47):29330–29337, Nov. 2020. doi: 10.1073/pnas.1912334117.
- [31] R. Gupta, M. Thapar Khanna, and S. Chaudhury. Visual saliency guided video compression algorithm. *Signal Processing: Image Communication*, 28(9):1006–1022, Oct. 2013. ISSN 0923-5965. doi: 10.1016/j.image.2013.07.003.
- [32] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, Sept. 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2649-2.
- [33] M. Hayhoe and D. Ballard. Eye movements in natural behavior. *Trends in Cognitive Sciences*, 9(4): 188–194, Apr. 2005. ISSN 1364-6613. doi: 10.1016/j.tics.2005.02.009.
- [34] D. Hoppe and C. A. Rothkopf. Multi-step planning of eye movements in visual search. *Scientific Reports*, 9(1):1–12, Jan. 2019. ISSN 2045-2322. doi: 10.1038/s41598-018-37536-0.

- [35] J. D. Hunter. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. ISSN 1521-9615. doi: 10.1109/MCSE.2007.55.
- [36] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.
- [37] Y. Jiang, Z. Guo, H. Rezazadegan Tavakoli, L. A. Leiva, and A. Oulasvirta. EyeFormer: Predicting Personalized Scanpaths with Transformer-Guided Reinforcement Learning. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, UIST ’24, pages 1–15, New York, NY, USA, Oct. 2024. Association for Computing Machinery. ISBN 979-8-4007-0628-8. doi: 10.1145/3654777.3676436.
- [38] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th International Conference On*, pages 2106–2113. IEEE, 2009.
- [39] T. Judd, F. Durand, and A. Torralba. A Benchmark of Computational Models of Saliency to Predict Human Fixations. *MIT Tech Report*, Jan. 2012.
- [40] F. Kadner, T. Thomas, D. Hoppe, and C. A. Rothkopf. Improving Saliency Models’ Predictions of the Next Fixation With Humans’ Intrinsic Cost of Gaze Shifts. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2104–2114, 2023.
- [41] A. J. Kell and J. H. McDermott. Deep neural network models of sensory systems: Windows onto the role of task constraints. *Current Opinion in Neurobiology*, 55:121–132, Apr. 2019. ISSN 0959-4388. doi: 10.1016/j.conb.2019.02.003.
- [42] P.-J. Kindermans, K. T. Schütt, M. Alber, K.-R. Müller, D. Erhan, B. Kim, and S. Dähne. Learning how to explain neural networks: PatternNet and PatternAttribution. In *ICLR 2018*, Feb. 2018.
- [43] Š. Kucharský, D. van Renswoude, M. Raijmakers, and I. Visser. WALD-EM: Wald accumulation for locations and durations of eye movements. *Psychological Review*, 128(4):667–689, 2021. ISSN 1939-1471. doi: 10.1037/rev0000292.
- [44] M. Kümmerer and M. Bethge. State-of-the-Art in Human Scanpath Prediction. *arXiv:2102.12239 [cs]*, Feb. 2021.
- [45] M. Kümmerer and M. Bethge. Predicting Visual Fixations. *Annual Review of Vision Science*, 9(1):269–291, Sept. 2023. ISSN 2374-4642, 2374-4650. doi: 10.1146/annurev-vision-120822-072528.
- [46] M. Kümmerer, Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durant, A. Oliva, A. Torralba, and M. Bethge. MIT/Tuebingen Saliency Benchmark. saliency.tuebingen.ai.
- [47] M. Kümmerer, L. Theis, and M. Bethge. Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet. *ICLR Workshop Track*, 2015.
- [48] M. Kümmerer, T. S. A. Wallis, and M. Bethge. Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences*, 112(52):16054–16059, Dec. 2015. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1510393112.
- [49] M. Kümmerer, T. S. A. Wallis, L. A. Gatys, and M. Bethge. Understanding Low- and High-Level Contributions to Fixation Prediction. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 4789–4798, 2017.
- [50] M. Kümmerer, T. S. A. Wallis, and M. Bethge. Saliency Benchmarking Made Easy: Separating Models, Maps and Metrics. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, pages 798–814. Springer International Publishing, 2018. ISBN 978-3-030-01270-0.
- [51] M. Kümmerer, M. Bethge, and T. S. A. Wallis. DeepGaze III: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision*, 22(5):7, Apr. 2022. ISSN 1534-7362. doi: 10.1167/jov.22.5.7.
- [52] M. Kümmerer, H. Khanuja, and M. Bethge. Modeling Saliency Dataset Bias, May 2025.
- [53] M. F. Land and B. W. Tatler. Steering with the head: The visual strategy of a racing driver. *Current Biology*, 11(15):1215–1220, Aug. 2001. ISSN 0960-9822. doi: 10.1016/S0960-9822(01)00351-7.
- [54] O. Le Meur and A. Coutrot. Introducing context-dependent and spatially-variant viewing biases in saccadic models. *Vision Research*, 121:72–84, Apr. 2016. ISSN 0042-6989. doi: 10.1016/j.visres.2016.01.005.

- [55] O. Le Meur and Z. Liu. Saccadic model of eye movements for free-viewing condition. *Vision Research*, 116:152–164, Nov. 2015. ISSN 0042-6989. doi: 10.1016/j.visres.2014.12.026.
- [56] A. Linardos, M. Kümmerer, O. Press, and M. Bethge. DeepGaze IIE: Calibrated Prediction in and Out-of-Domain for State-of-the-Art Saliency Modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12919–12928, 2021.
- [57] S. Ling, M. S. Pratte, and F. Tong. Attention alters orientation processing in the human lateral geniculate nucleus. *Nature Neuroscience*, 18(4):496–498, Apr. 2015. ISSN 1546-1726. doi: 10.1038/nn.3967.
- [58] C. Lu, C. Lu, R. T. Lange, J. Foerster, J. Clune, and D. Ha. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery, Sept. 2024.
- [59] N. Malem-Shinitzki, M. Oppen, S. Reich, L. Schwetlick, S. A. Seelig, and R. Engbert. A mathematical model of local and global attention in natural scene viewing. *PLOS Computational Biology*, 16(12): e1007880, Dec. 2020. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007880.
- [60] S. Mondal, Z. Yang, S. Ahn, D. Samaras, G. Zelinsky, and M. Hoai. Gazeformer: Scalable, Effective and Fast Prediction of Goal-Directed Human Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1441–1450, 2023.
- [61] R. D. Morey. Confidence Intervals from Normalized Data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, 4(2):61–64, Sept. 2008. ISSN 1913-4126. doi: 10.20982/tqmp.04.2.p061.
- [62] B. C. Motter. Saccadic momentum and attentive control in V4 neurons during visual search. *Journal of Vision*, 18(11):16, Oct. 2018. ISSN 1534-7362. doi: 10.1167/18.11.16.
- [63] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library, Dec. 2019.
- [64] J. B. Pelz and R. Canosa. Oculomotor behavior and perceptual strategies in complex tasks. *Vision Research*, 41(25):3587–3596, Nov. 2001. ISSN 0042-6989. doi: 10.1016/S0042-6989(01)00245-0.
- [65] K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422, 1998. ISSN 1939-1455. doi: 10.1037/0033-2909.124.3.372.
- [66] J. Reback, jbrockmendl, W. McKinney, J. V. den Bossche, T. Augspurger, P. Cloud, S. Hawkins, M. Roeschke, gyoung, Sinhrks, A. Klein, P. Hoyer, T. Petersen, J. Tratner, C. She, W. Ayd, S. Naveh, J. H. M. Darbyshire, M. Garcia, R. Shadrach, J. Schendel, A. Hayden, D. Saxton, M. E. Gorelli, F. Li, M. Zeitlin, V. Jancauskas, A. McMaster, P. Battiston, and S. Seabold. Pandas-dev/pandas: Pandas 1.4.1. Zenodo, Feb. 2022.
- [67] L. O. M. Rothkegel, H. H. Schütt, H. A. Trukenbrod, F. A. Wichmann, and R. Engbert. Searchers adjust their eye-movement dynamics to target characteristics in natural scenes. *Scientific Reports*, 9(1):1635, Feb. 2019. ISSN 2045-2322. doi: 10.1038/s41598-018-37548-w.
- [68] H. H. Schütt, L. O. M. Rothkegel, H. A. Trukenbrod, S. Reich, F. A. Wichmann, and R. Engbert. Likelihood-based parameter estimation and comparison of dynamical cognitive models. *Psychological Review*, 124(4): 505–524, 2017. ISSN 1939-1471, 0033-295X. doi: 10.1037/rev0000068.
- [69] H. H. Schütt, L. O. M. Rothkegel, H. A. Trukenbrod, R. Engbert, and F. A. Wichmann. Disentangling bottom-up versus top-down and low-level versus high-level influences on eye movements over time. *Journal of Vision*, 19(3):1–1, Mar. 2019. ISSN 1534-7362. doi: 10.1167/19.3.1.
- [70] L. Schwetlick, L. O. M. Rothkegel, H. A. Trukenbrod, and R. Engbert. Modeling the effects of perisaccadic attention on gaze statistics during scene viewing. *Communications Biology*, 3(1):1–11, Dec. 2020. ISSN 2399-3642. doi: 10.1038/s42003-020-01429-8.
- [71] L. Schwetlick, D. Backhaus, and R. Engbert. A dynamical scan-path model for task-dependence during scene viewing. *Psychological Review*, 130(3):807–840, Apr. 2023. ISSN 1939-1471, 0033-295X. doi: 10.1037/rev0000379.
- [72] L. Schwetlick, M. Kümmerer, M. Bethge, and R. Engbert. Potsdam data set of eye movement on natural scenes (DAEMONS). *Frontiers in Psychology*, 15, May 2024. ISSN 1664-1078. doi: 10.3389/fpsyg.2024.1389609.

- [73] L. N. Smith and N. Topin. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates, May 2018.
- [74] T. J. Smith and J. M. Henderson. Facilitation of return during scene viewing. *Visual Cognition*, 17(6-7): 1083–1108, Aug. 2009. ISSN 1350-6285. doi: 10.1080/13506280802678557.
- [75] T. J. Smith and J. M. Henderson. Looking back at Waldo: Oculomotor inhibition of return does not prevent return fixations. *Journal of Vision*, 11(1):3, Jan. 2011. ISSN 1534-7362. doi: 10.1167/11.1.3.
- [76] S. S. Stevens. On the Theory of Scales of Measurement. *Science*, 103(2684):677–680, June 1946. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.103.2684.677.
- [77] X. Sui, Y. Fang, H. Zhu, S. Wang, and Z. Wang. ScanDMM: A Deep Markov Model of Scanpath Prediction for 360deg Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6989–6999, 2023.
- [78] W. Sun, Z. Chen, and F. Wu. Visual Scanpath Prediction using IOR-ROI Recurrent Mixture Density Network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019. ISSN 1939-3539. doi: 10.1109/TPAMI.2019.2956930.
- [79] X. Sun, H. Yao, R. Ji, and X.-M. Liu. Toward Statistical Modeling of Saccadic Eye-Movement and Visual Saliency. *IEEE Transactions on Image Processing*, 23(11):4649–4662, Nov. 2014. ISSN 1941-0042. doi: 10.1109/TIP.2014.2337758.
- [80] B. W. Tatler, M. M. Hayhoe, M. F. Land, and D. H. Ballard. Eye guidance in natural vision: Reinterpreting saliency. *Journal of Vision*, 11(5):5–5, May 2011. ISSN 1534-7362. doi: 10.1167/11.5.5.
- [81] B. W. Tatler, J. R. Brockmole, and R. H. S. Carpenter. LATEST: A model of saccadic decisions in space and time. *Psychological Review*, 124(3):267–300, 2017. ISSN 1939-1471, 0033-295X. doi: 10.1037/rev0000054.
- [82] I. M. Thornton, Q. C. Vuong, and K. S. Pilz. A Search Advantage for Horizontal Targets in Dynamic Displays. *i-Perception*, 12(2):20416695211004616, Apr. 2021. ISSN 2041-6695. doi: 10.1177/20416695211004616.
- [83] Y. Tsal. Attending to horizontal, diagonal, and vertical positions in space. *Bulletin of the Psychonomic Society*, 27(2):133–134, Feb. 1989. ISSN 0090-5054. doi: 10.3758/BF03329920.
- [84] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, Í. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, and P. van Mulbregt. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, pages 1–12, Feb. 2020. ISSN 1548-7105. doi: 10.1038/s41592-019-0686-2.
- [85] Z. Wang and E. P. Simoncelli. Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities. *Journal of Vision*, 8(12):8, Sept. 2008. ISSN 1534-7362. doi: 10.1167/8.12.8.
- [86] M. Waskom. Seaborn: Statistical data visualization. *Journal of Open Source Software*, 6(60):3021, Apr. 2021. ISSN 2475-9066. doi: 10.21105/joss.03021.
- [87] D. A. Wilkenfeld. Understanding as compression. *Philosophical Studies*, 176(10):2807–2831, Oct. 2019. ISSN 1573-0883. doi: 10.1007/s11098-018-1152-1.
- [88] N. Wilming, S. Harst, N. Schmidt, and P. König. Saccadic Momentum and Facilitation of Return Saccades Contribute to an Optimal Foraging Strategy. *PLOS Computational Biology*, 9(1):e1002871, Jan. 2013. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002871.
- [89] C. Wloka, I. Kotseruba, and J. K. Tsotsos. Active Fixation Control to Predict Saccade Sequences. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3184–3193, Salt Lake City, UT, USA, June 2018. IEEE. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00336.
- [90] Y. Xu, Y. Dong, J. Wu, Z. Sun, Z. Shi, J. Yu, and S. Gao. Gaze Prediction in Dynamic 360° Immersive Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5333–5342, 2018.
- [91] Z. Yang, L. Huang, Y. Chen, Z. Wei, S. Ahn, G. Zelinsky, D. Samarasinghe, and M. Hoai. Predicting Goal-Directed Human Attention Using Inverse Reinforcement Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 193–202, 2020.

- [92] Z. Yang, S. Mondal, S. Ahn, R. Xue, G. Zelinsky, M. Hoai, and D. Samaras. Unifying Top-down and Bottom-up Scanpath Prediction Using Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1683–1693, 2024.
- [93] A. L. Yarbus. *Eye Movements and Vision*. Plenum Press, 1967.
- [94] D. Zanca, S. Melacci, and M. Gori. Gravitational Laws of Focus of Attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019. ISSN 0162-8828. doi: 10.1109/TPAMI.2019.2920636.
- [95] R. S. Zimmermann, J. Borowski, R. Geirhos, M. Bethge, T. Wallis, and W. Brendel. How Well do Feature Visualizations Support Causal Understanding of CNN Activations? In *Advances in Neural Information Processing Systems*, volume 34, pages 11730–11744. Curran Associates, Inc., 2021.

A Comparative evaluation on more models and datasets

We evaluated our models on four datasets: MIT1003 [38], DAEMONS [72], COCO Freeview [18] and the Potsdam Corpus [69]. On all four datasets, we construct a common spatial baseline model and then train DeepGaze as well as the different SceneWalk variants using the spatial baseline model. For better comparison, we also evaluate the Constrained Levy Exploration model [CLE; 6] and the model of Kadner et al., 2023 [40]. For all models we evaluate information gain relative to the uniform baseline (LL) and AUC [44]. We find (Table 1) that on all datasets, SceneWalk-X closes a substantial part of the performance gap between the perisaccadic SceneWalk and DeepGaze, both in LL and in AUC.

As a note, temperature scaling has only a very small effect in AUC: this is to be expected, as AUC is invariant to temperature changes. The marginal score increase is likely an indirect effect, as having temperature as a free parameter during training gives the optimizer more flexibility, allowing the other model parameters to converge to a slightly better solution.

Model	MIT1003		DAEMONS		COCO Freeview		Potsdam Corpus	
	LL	AUC	LL	AUC	LL	AUC	LL	AUC
Spatial Baseline	2.52	0.9223	1.88	0.8743	2.08	0.9034	1.81	0.8768
Kadner et al. 2023 [40]	-	0.9237	-	0.8806	-	0.9113	-	-
CLE (Boccignone et al. 2004, [6])	2.18	0.9240	2.42	0.9118	2.34	0.9212	-	-
SceneWalk (perisaccadic)	2.66	0.9304	2.71	0.9207	2.53	0.9240	2.32	0.9129
+ temp. scaling	2.72	0.9305	2.73	0.9209	2.54	0.9241	2.33	0.9130
+ momentum	2.77	0.9321	2.82	0.9235	2.58	0.9254	2.38	0.9147
+ card. bias = SceneWalk-X	2.82	0.9343	2.88	0.9267	2.64	0.9275	2.42	0.9164
DeepGaze	2.94	0.9386	3.18	0.9365	2.87	0.9343	2.54	0.9194

Table 1: Performance of our models and other previous scanpath models on MIT1003, DAEMONS, COCO-Freeview and the Potsdam Scene Viewing Corpus

B Implementation details of the new mechanisms

B.1 Time dependent temperature scaling

To capture the observation that early fixations tend to be more strongly biased toward salient image regions, we introduce a time-dependent temperature scaling mechanism into the SceneWalk model. This mechanism dynamically adjusts the exponents used to shape the attention and inhibition maps during scanpath generation based on the fixation index. The original SceneWalk model already applies a fixed exponent to these maps to regulate their sharpness, acting similarly to a softmax temperature. Our extension generalizes this by making the exponents decay over time, allowing for sharper attention selection early in the trial and broader selection later.

Concretely, in the original SceneWalk model, the priority map $u_{ij}(t)$ at each location is computed as a subtractive combination of normalized, exponentiated attention $A_{ij}(t)$ and inhibition $F_{ij}(t)$ maps:

$$u_{ij}(t) = \frac{(A_{ij}(t))^\gamma}{\sum_{kl} (A_{kl}(t))^\gamma} - C_F \cdot \frac{(F_{ij}(t))^\gamma}{\sum_{kl} (F_{kl}(t))^\gamma}. \quad (1)$$

Here, C_F is a weight parameter for inhibition, and, importantly, γ controls the sharpness of both the attention and inhibition distributions and was constant across fixations.

We replace the constant γ with a fixation-index-dependent value. Specifically, the exponent γ_n for the n -th fixation is given by an exponential decay from an initial high value γ toward a base value γ_{base} :

$$\gamma_n = \gamma_{\text{base}} + (\gamma - \gamma_{\text{base}}) \cdot e^{-\lambda_\gamma \cdot (n-1)}. \quad (2)$$

Note that here the fixation index in the scanpath n follows a 1-based indexing.

The new trainable parameters we introduce with this formulation are:

1. γ_{base} : minimum exponent for inhibition
2. λ_γ : decay rate for γ

This enables the model to express a strong preference for high-saliency targets early on (large γ_n) and gradually reduce this preference (lower γ_n) over time, converging to γ_{base} .

B.2 Saccadic momentum

To better capture directional tendencies that DeepGaze can capture in human scanpaths—specifically the tendency to continue in the same direction (momentum) and to return to previously attended regions (return saccades)—we augment the SceneWalk model with a dynamic oculomotor bias map M_{ij}^{dir} . The map is computed over all possible gaze targets on each fixation, and combines two angular Gaussian components: A momentum component favoring forward continuation and a return component favoring 180° reversals. These components are blended using a dynamic weighting function, which depends both on the current fixation index and the length of the preceding saccade. Specifically, let

- n be the current fixation index (with 1-based indexing),
- s be the length of the previous saccade, in degrees of visual angle,
- θ_{ij} be the angular difference between the current fixation and candidate saccade directions, normalized between $-\pi$ and π .

Then we define the fixation index weight as:

$$w_{\text{fix}}(n) = \frac{1}{1 + \exp\left(-\frac{n - \mu_{\text{fix}}}{\sigma_{\text{fix}}}\right)}, \quad (3)$$

and the saccade length weight as:

$$w_{\text{len}}(s) = \frac{1}{1 + \exp\left(-\frac{s - s_{\text{short}}}{s_{\text{scale}}}\right)}. \quad (4)$$

The fixation index weight and the saccade length weight are then multiplied to get the combined direction weight:

$$w_{\text{dir}} = w_{\text{fix}} \cdot w_{\text{len}}. \quad (5)$$

We can compute the angular preference fields as:

$$\begin{aligned} G_{\text{momentum}}(\theta_{ij}) &= \exp\left(-\frac{\theta_{ij}^2}{2\sigma_{\text{mom}}^2}\right), \\ G_{\text{return}}(\theta_{ij}) &= \exp\left(-\frac{(\theta_{ij} - \pi)^2}{2\sigma_{\text{return}}^2}\right) + \exp\left(-\frac{(\theta_{ij} + \pi)^2}{2\sigma_{\text{return}}^2}\right). \end{aligned} \quad (6)$$

These are then combined as follows, using the combined direction weight:

$$M_{ij}^{\text{dir}} = (1 - w_{\text{dir}}) \cdot G_{\text{return}}(\theta_{ij}) + w_{\text{dir}} \cdot G_{\text{momentum}}(\theta_{ij}). \quad (7)$$

Finally, the map is combined with SceneWalk’s priority map u_{ij}^{final} before saccade selection:

$$u_{ij}^{\text{final}} = \frac{u_{ij}^{\text{final}} \cdot (\omega_{\text{bias}} \cdot M_{ij}^{\text{dir}})}{Z}. \quad (8)$$

The new trainable parameters we introduce with this mechanism are:

1. σ_{mom} : width of the angular Gaussian for saccadic momentum (same-direction preference)
2. σ_{return} : width of the angular Gaussian for return saccades (opposite-direction preference)
3. μ_{fix} : fixation index midpoint for the sigmoid controlling the momentum-return tradeoff
4. σ_{fix} : steepness of the sigmoid over fixation index
5. s_{short} : midpoint of the sigmoid controlling sensitivity to saccade length
6. s_{scale} : scale parameter of the sigmoid over saccade length
7. ω_{bias} : scaling factor controlling the influence of the directional prior in the final priority map.

B.3 Cardinal attention bias

Following comparisons with DeepGaze and consulting the literature, to capture anisotropies in visual attention and oculomotor behavior we introduce a mechanism that biases SceneWalk’s attention map toward cardinal (horizontal and vertical) saccade directions. Additionally, we model a leftward bias for early fixations. These components are combined into a single potential map that modulates the attention map before it is normalized and combined with inhibition. As for saccadic momentum and temperature, the mechanism is dependent on fixation index.

Let:

- n be the current fixation index (with 1-based indexing),
- θ_{ij} be the angular difference between the current fixation and candidate saccade target,
- d_{ij} be the Euclidean distance of a candidate saccade target to the current fixation.

Then we formulate the cardinal bias as:

$$Q_{ij}^{\text{cardinal}} = \cos(4\theta_{ij}) \cdot \exp\left(-\frac{d_{ij}}{\tau}\right), \quad (9)$$

where $\tau \propto \frac{1}{\chi}$ is a decay constant.

We further define the vertical/horizontal strength modulation as a scalar that increases over fixations:

$$\alpha_{\text{dir}}(n) = 1 + (\alpha_{\text{vert}} - 1) \cdot (1 - e^{-\rho n}), \quad (10)$$

and the leftward bias as:

$$\alpha_{\text{left}}(n) = 1 + (\alpha_{\text{left}}^0 - 1) \cdot e^{-\lambda(n-1)}. \quad (11)$$

The biases are then applied as follows. First, for the horizontal/vertical bias, let $m_{ij}^{\text{vert}} = 1$ if location (i, j) lies in the vertical direction relative to the current fixation (i.e., $|dx| < |dy|$), otherwise $m_{ij}^{\text{vert}} = 0$. The vertical bias is applied as:

$$Q_{ij}^{\text{cardinal}} \leftarrow Q_{ij}^{\text{cardinal}} \cdot [m_{ij}^{\text{vert}} \cdot \alpha_{\text{dir}}(n) + (1 - m_{ij}^{\text{vert}}) \cdot 1]. \quad (12)$$

Similarly, for the leftward bias, we define binary masks for horizontal directions:

- $m_{ij}^{\text{left}} = 1$ if $dx < 0$ and $|dx| > |dy|$, 0 otherwise;
- $m_{ij}^{\text{right}} = 1$ if $dx > 0$ and $|dx| > |dy|$, 0 otherwise;

and then apply:

$$Q_{ij}^{\text{cardinal}} \leftarrow Q_{ij}^{\text{cardinal}} \cdot \left[m_{ij}^{\text{left}} \cdot \alpha_{\text{left}}(n) + m_{ij}^{\text{right}} \cdot \frac{1}{\alpha_{\text{left}}(n)} + (1 - m_{ij}^{\text{left}} - m_{ij}^{\text{right}}) \cdot 1 \right]. \quad (13)$$

Finally, the resulting potential map is normalized to sum to 1, and applied multiplicatively to the attention map:

$$A_{ij}^{\text{mod}} = \frac{A_{ij} \cdot (1 + \omega_{\text{mix}} \cdot Q_{ij}^{\text{cardinal}})}{Z} \quad (14)$$

The new trainable parameters we introduce with this mechanism are:

1. χ : controls the spatial decay rate of the cardinal potential with distance
2. α_{vert} : maximum vertical vs horizontal bias ratio
3. ρ : growth rate of vertical/horizontal anisotropy over fixations
4. α_{left}^0 : initial strength of the leftward bias
5. λ : decay rate of the leftward bias over fixations
6. ω_{mix} : global scaling factor for the oculomotor potential’s influence on the attention map.

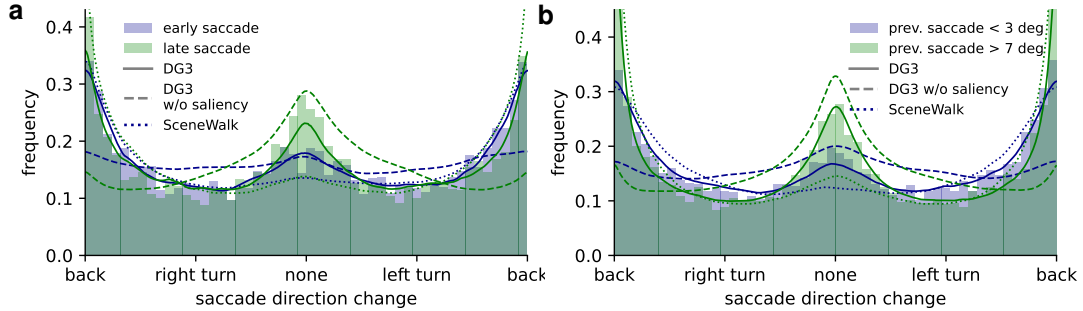


Figure 7: relative saccade direction distribution as predicted by DeepGaze, DeepGaze without saliency maps and SceneWalk for early and late saccades (a) as well as after short or long previous saccades (b).

C Spatial baseline

Commonly, empirical fixation densities are estimated with Gaussian kernel density estimates. However, recent progress in spatial saliency modeling [52] has produced models which predict fixations from new observers better than the empirical density. While this feels counterintuitive, it is possible because saliency datasets typically have data from only 10–20 subjects per image [18, 8]). Suboptimal saliency maps could lead to confounders in our study, e.g., models might use previous fixation locations simply for estimating a better saliency map. Hence we extend the approach by [52] and estimate the average fixation density for each subject on an image as a mixture of four distributions: first, a KDE over the fixations from all other subjects, secondly the centerbias, i.e. a KDE over the fixations from all other images, thirdly a uniform component, and in addition as fourth component we add the predicted fixation density from the state-of-the-art saliency model [52]. The last component guarantees that the spatial baseline model is at least as good as the used saliency model. For a discussion of the other three components, see [45].

The parameters of the model, i.e. bandwidth and mixture weights, are optimized individually on each image for maximum likelihood. This results in a prediction for each subject, using only other subject’s fixations. To compute the final fixation density for an image, we average over the predicted fixation densities for all subjects with weights that are location dependent: for each true fixation location, only the components not using this fixation in their KDE are switched on. This defines weighting coefficients for all fixated pixels such that the prediction for that pixel only comes from subjects which didn’t fixate the pixel. For all other pixels, the weights are interpolated linearly, and the resulting spatial average is re-normalized. This approach removes visual artifacts due to single, potentially random fixations and gives more interpretable results.

D Details about the datasets

The MIT1003 dataset is accessed via the pysaliency python library <https://github.com/matthias-k/pysaliency>. DAEMONS and COCO Freeview come with an official training/validation split which we are using. MIT1003 and the Potsdam Corpus do not have official validation splits. Instead, we use 10-fold crossvalidation using `pysaliency.filter_datasets.train_validation_test_fold(stimuli, fixations, crossval_folds=10, test_folds=1, val_folds=1)`. On MIT1003, the validation split has been used for tuning model hyperparameters and learning rate schedules, which have then also been used on the other datasets. All reported evaluations use the test split on MIT1003 and the Potsdam Corpus, and the official validation split for DAEMONS and COCO Freeview.

E Details about models and training

DeepGaze DeepGaze III was originally trained using a four stage training paradigm including spatial pretraining on a different dataset, spatial training on the target dataset, scanpath training on

the target dataset and a final joint training phase. This procedure was necessary for not overfitting the spatial priority network while training the scanpath parts. Since we removed the spatial priority network, we only need the scanpath training phase. Since DeepGaze III was originally trained on MIT1003 dataset, which we are also using, we can apply exactly the same hyperparameters: DeepGaze is trained using the Adam optimizer with a batchsize of 4 and an initial learning rate of 0.001. In addition, we use a learning rate schedule consisting of decays of the learning rate by a factor of 10 after 10, 20, 30 and 31 epochs. After the 32nd epoch, training is stopped. Consistent with the original training, we also use 10-fold crossvalidation. For all results in this paper use for each image the model weights which have not seen this image in training.

SceneWalk Originally, SceneWalk was trained fully Bayesian using MCMC. Here, we instead re-implemented the model in Jax and train it using MAP estimation, resulting in substantially faster fitting times. SceneWalk is trained via gradient descent with the Adam optimizer on a MAP objective. Parameter prior ranges are chosen to be broad, and taken mostly from Schwetlick et al. [70]. The hyperparameter configuration for the base perisaccadic model is also taken from Schwetlick et al. [70]. We use a “OneCycle” learning rate schedule [73] over 30 epochs for the MIT1003 dataset, and 50 epochs for DAEMONS, with an initial learning rate 0.001, a peak learning rate of 0.01 and a batch size of 32.

F Error bars

All reported error bars are bootstrapped 95% confidence intervals for the mean log-likelihood per image using the normalization method of Cousineau [20] for paired comparisons with the correction of Morey [61].

G Assets

DeepGaze is implemented in python using pytorch [63] using the official implementation from github.com/matthias-k/DeepGaze. SceneWalk was reimplemented in jax [9] (Apache 2 license) based on the official implementation from github.com/lswetlick/SceneWalk_Model. Model evaluations and saliency metrics were using the public pysaliency toolbox (github.com/matthias-k/pysaliency, MIT license). Also used were scipy [84] and numpy [32] for computations, pandas [66] for statistics and data handling as well as matplotlib [35] and seaborn [86] for plotting.

H Code Availability

The SceneWalk-X implementation in jax, along with code to reproduce our main findings can be found at <https://github.com/bethgelab/what-moves-the-eyes>.

I Compute Resources

All main experiments were conducted on 2080Ti GPUs. Training the DeepGaze model took about 3 days for all 10 splits and was done four times until the final model setup evolved. Training SceneWalk took about 30 minutes, while developing the mechanisms we trained about 50 model iterations. On DAEMONS, we used A100 GPUs and DeepGaze took about 4 days to train, and each SceneWalk model about two hours. Overall, we used about 18 days of compute.

J Controversial Fixations

Here we show the full set of controversial fixations that we inspected and classified as basis for our mechanistic extensions. We examined a total of 72 fixations, defined as the top 6 fixations from the top 6 images in both conditions (LLD and WLLD). For all controversial fixations, we show predictions from DeepGaze, SceneWalk and all mechanistic extensions of SceneWalk, including the fully extended SceneWalk-X model.

Of the four categories we detail in Figure 2, we found:

- Long saccades: 17 clear, 3 plausible
- Short saccades: 8 clear, 1 plausible
- Continuing: 13 clear, 2 plausible
- Early fixations: 30 clear

66 out of 72 fixations (92%) could be clearly assigned to one of the four categories, and the remaining 6 (8%) still plausibly belonged to a known type. We see this as a strong indicator that the analysis indeed revealed failure cases that are not idiosyncratic, but rather fall into meaningful, repeatable patterns that are interpretable and mechanistically actionable.



Figure 8: Controversial Fixations selected for maximum difference in log-likelihood between DeepGaze and SceneWalk, Part 1. We first select the top images with mean log-likelihood difference and then show for each image the 6 fixations with maximum log-likelihood difference. We visualize the image, the empirical densities and the predictions of DeepGaze, SceneWalk and all extension versions that we introduce. On top of each prediction, we show the log-likelihood of how well the fixation is predicted. For the DeepGaze prediction, we additionally show the score difference between DeepGaze and SceneWalk that resulted in this fixation being picked as controversial. On top of the image we show the mean score difference that resulted in this image being picked.



Figure 9: Controversial Fixations selected for maximum difference in log-likelihood between DeepGaze and SceneWalk, Part 2.



Figure 10: Controversial Fixations selected for maximum difference in log-likelihood between DeepGaze and SceneWalk, Part 3.



Figure 11: Controversial Fixations selected for maximum difference in log-likelihood between DeepGaze and SceneWalk, Part 4.

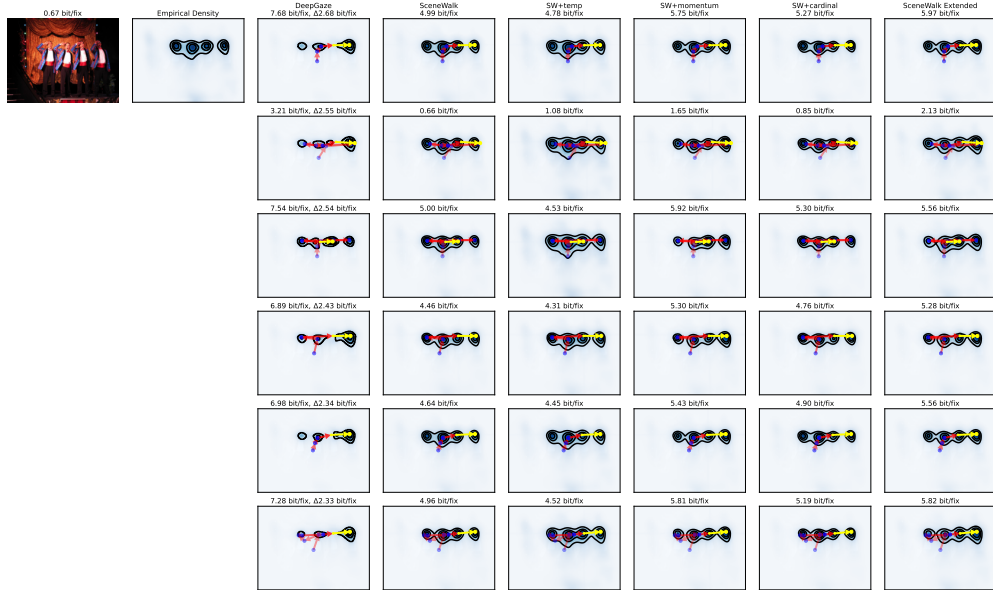


Figure 12: Controversial Fixations selected for maximum difference in log-likelihood between DeepGaze and SceneWalk, Part 5.



Figure 13: Controversial Fixations selected for maximum difference in log-likelihood between DeepGaze and SceneWalk, Part 6.

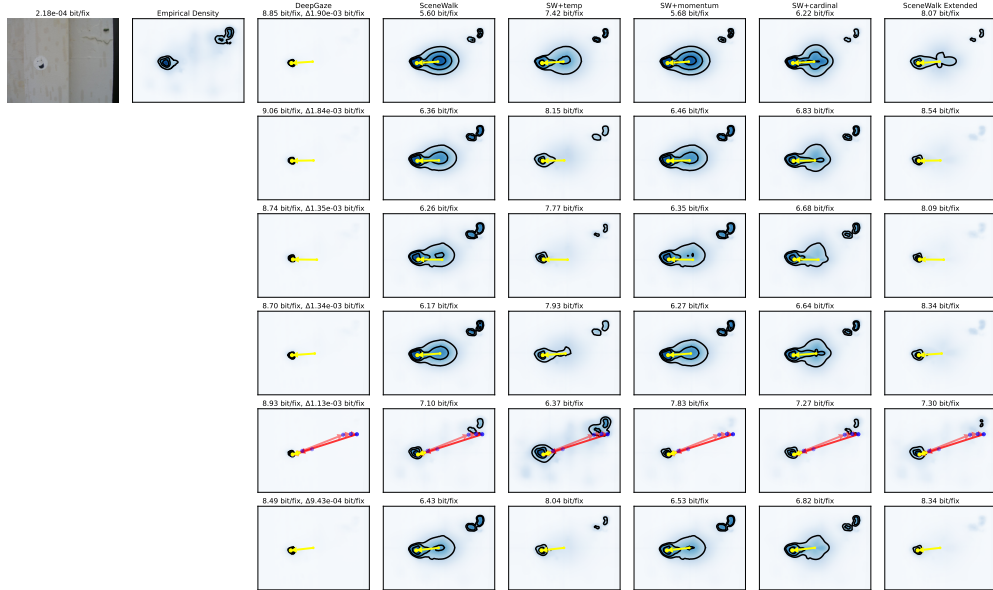


Figure 14: Controversial Fixations selected for maximum difference in log-likelihood between DeepGaze and SceneWalk weighted by the probability assigned by DeepGaze, Part 1.



Figure 15: Controversial Fixations selected for maximum difference in log-likelihood between DeepGaze and SceneWalk weighted by the probability assigned by DeepGaze, Part 2.



Figure 16: Controversial Fixations selected for maximum difference in log-likelihood between DeepGaze and SceneWalk weighted by the probability assigned by DeepGaze, Part 3.

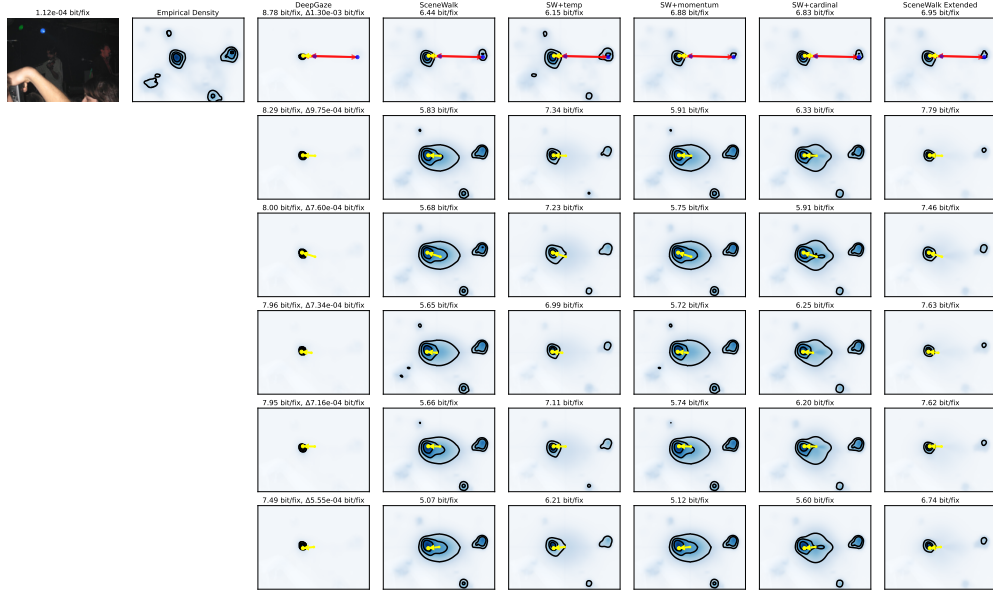


Figure 17: Controversial Fixations selected for maximum difference in log-likelihood between DeepGaze and SceneWalk weighted by the probability assigned by DeepGaze, Part 4.

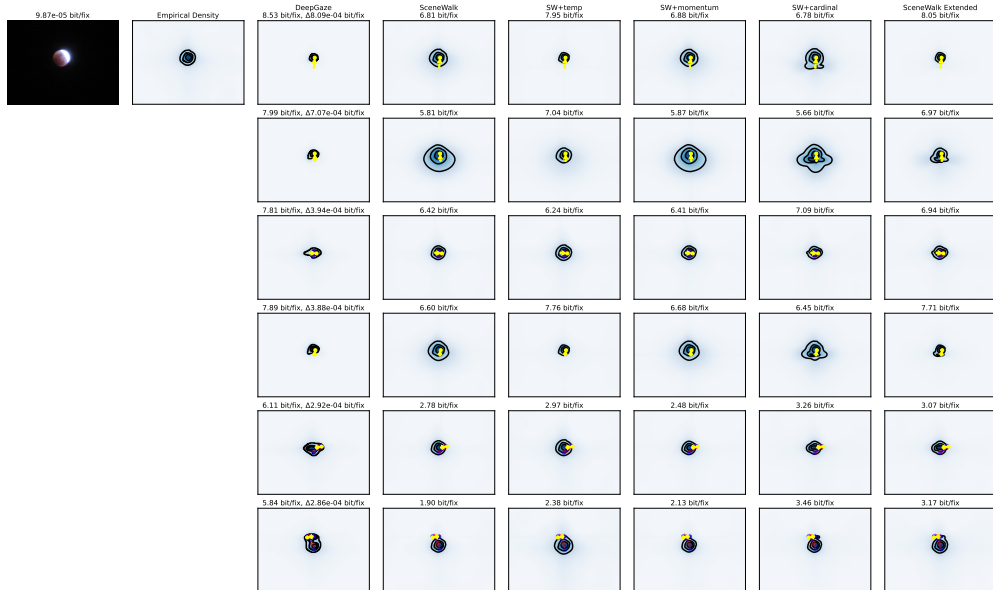


Figure 18: Controversial Fixations selected for maximum difference in log-likelihood between DeepGaze and SceneWalk weighted by the probability assigned by DeepGaze, Part 5.



Figure 19: Controversial Fixations selected for maximum difference in log-likelihood between DeepGaze and SceneWalk weighted by the probability assigned by DeepGaze, Part 6.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: All claims in the abstract and the introduction are supported by the results in the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discuss the limitations of our work as far as applicable already in the section on the new mechanisms, and more broadly in the discussion including a specific section on limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper doesn’t contain any theoretical results.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We specify details about used dataset splits and training methods for our models in the appendix, as well as the used metrics and selection methods. The base models we’re using are public models, so there is no need to specify details about the architecture. We specify details for the new mechanisms we have implemented in SceneWalk-X in the appendix.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: <https://github.com/bethgelab/what-moves-the-eyes>

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: We specify details about used dataset splits and training methods for our models in the Appendix, as well as the used metrics and selection methods. The models we're using are public models, so there is no need to specify details about the architecture.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[Yes\]](#)

Justification: We display error bars for all our results and give details about their computation in the Appendix.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: We detail the used resources in the Appendix and include an estimate of the total compute needed.

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [\[Yes\]](#)

Justification: We made sure to adhere to the NeurIPS code of ethics in each step of our experiments.

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [\[Yes\]](#)

Justification: We discuss positive and negative broader impact of our research in the discussion

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [\[Yes\]](#)

Justification: We do not release any data on our own and our models will be released under a Responsible AI Licence (RAIL, licenses.ai) to prohibit misuse.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [\[Yes\]](#)

Justification: All authors of used assets are referenced in the paper and licenses, as far as specified, are mentioned and respected. However, for some software libraries we couldn't find license details.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[NA\]](#)

Justification: This paper does not introduce new assets. We will make model and checkpoints available in the future.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[NA\]](#)

Justification: The paper does not involve crowdsourcing or research with humans subjects. We only use previously published datasets.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [\[NA\]](#)

Justification: The paper does not involve crowdsourcing or research with humans subjects.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [\[NA\]](#)

Justification: This paper does not use LLMs as core part of the research.