Debiasing Central Fixation Confounds Reveals a Peripheral "Sweet Spot" for Human-like Scanpaths in Hard-Attention Vision

Pengcheng Pan[1], **Yonekura Shogo**[1] **& Yasuo Kuniyoshi**[1]
[1]Department of Mechano-Informatics, The University of Tokyo

Abstract

Human eye movements in visual recognition reflect a balance between foveal sampling and peripheral context, but evaluating whether artificial scanpaths are "human-like" is difficult on object-centric datasets with strong center bias. Using Gaze-CIFAR-10, we show that a trivial *center-fixation baseline* achieves surprisingly strong scores under common scanpath metrics, blurring the distinction between behavioral alignment and central tendency. We introduce **GCS** (Gaze Consistency Score), a practical center-debiased and movement-aware score that normalizes against human and corner references, subtracts the center baseline, and adds a small movement-similarity term. Applying GCS to a hard-attention classifier under varied fovea–periphery constraints identifies a restricted mid-range regime: a moderate foveal patch with peripheral context yields stronger center-debiased alignment than either narrower or broader alternatives. This regime is not identified by accuracy alone; the highest-accuracy setting differs from the best-GCS setting. These results highlight the need for bias-aware scanpath evaluation and suggest that, on Gaze-CIFAR-10 and under this hard-attention setting, perceptual constraints shape when task-trained policies appear relatively human-like.

Keywords: active perception; hard attention; scanpath similarity; center bias; gaze metrics; peripheral vision

Introduction

Human visual perception is an active process. Rather than processing the entire visual field uniformly, humans rely on a foveated visual system and sequential eye movements to gather task-relevant information. Classic work by Yarbus, 1967 demonstrated that eye movement patterns are systematically shaped by task demands, suggesting that scanpaths can provide a window into underlying cognitive strategies.

In recent years, this idea has motivated the development of *hard-attention* models in machine vision, which explicitly select where to look at each step while performing tasks such as image classification or visual search (Ba et al., 2014; Mnih et al., 2014; Pan et al., 2026). These models offer a promising computational framework for studying perception under resource constraints. However, an open question remains: *when do the scanpaths produced by such models trained purely on the task meaningfully resemble human eye movements?*

A major challenge in answering this question lies in how scanpaths are evaluated. A wide range of metrics have been proposed to compare fixation sequences, made available on FixaTons tools (Zanca et al., 2018), including Dynamic Time Warping ("Dynamic Time Warping," 2007), ScanMatch (Cristino et al., 2010), Normalized Scanpath Saliency (Peters et al., 2005), and AUC-based measures originally developed for saliency evaluation (Judd et al., 2012). These metrics capture complementary aspects of scanpaths, such as temporal alignment, spatial overlap, and distributional similarity.

However, a long-standing finding in human eye-movement research complicates the interpretation: *human fixations are strongly biased toward the image center*. This **central fixation bias** has been robustly documented across free viewing and task-driven settings (Bindemann, 2010; Tatler, 2007; Tatler et al., 2011). Proposed explanations include photographer bias, experimental framing, oculomotor constraints, and learned expectations about where informative content is likely to appear (Smith & Mital, 2013). Crucially, center bias is not merely noise, it is a systematic property of many gaze datasets.

Center bias as a confound in scanpath evaluation. When center bias is strong, scanpath similarity metrics can be dominated by marginal spatial distributions rather than by genuine strategy similarity. Indeed, prior work in saliency modeling has shown that center-biased predictors can achieve high AUC and NSS scores without modeling image content (Borji, 2021; Borji & Itti, 2013). Despite this, task-driven scanpath evaluation has rarely applied explicit chance correction or strong spatial baselines.

This issue is particularly acute for object-centric datasets, where objects are frequently centered by design. In such settings, a trivial strategy that repeatedly fixates the center may appear "human-like" under standard scanpath metrics, even if it lacks meaningful temporal structure or task-driven exploration. As a result, high scanpath similarity does not necessarily imply cognitively plausible attention control.

Active vision, perceptual constraints, and strategy regimes. Classic work in psychology and neuroscience emphasizes that scanpaths depend strongly on task demands and internal goals (Yarbus, 1967). In cognitive robotics and AI, the active perception framework formalizes perception as an action-conditioned process (Bajcsy, 1988; Ballard, 1991), motivating models that must decide *where to look* under limited sensing.

Normative models of active vision show that optimal fixation selection depends critically on sensory constraints such as acuity, noise, and field-of-view (Geisler, 2008; Najemnik & Geisler, 2005). Manipulating these constraints can induce

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

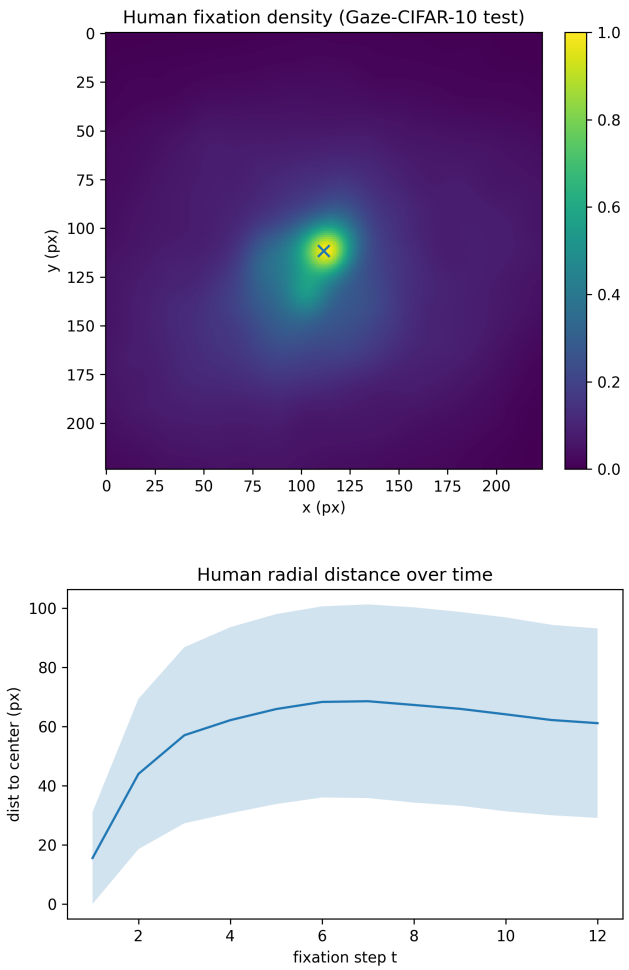

Figure 1: Human fixation density and radial distance over time (Gaze-CIFAR-10 test). The distribution is strongly center-biased, which can inflate scanpath similarity for trivial center policies.

qualitative shifts in behavior, from exploratory search to more conservative sampling.

Similar trade-offs arise in artificial agents. Hard-attention models with access to broader spatial context can reduce the need for movement, sometimes relying on shortcut strategies that bypass sequential evidence accumulation (Geirhos et al., 2020). Conversely, overly narrow views can force excessive exploration. These observations motivate a scoped hypothesis: under this hard-attention setting, relatively human-like scanpaths may appear most clearly within a restricted regime of perceptual constraints.

Goals and contributions. In this work, we study these issues using Gaze-CIFAR-10 (Li et al., 2025), a gaze-annotated object recognition dataset. We make three contributions:

1. **Quantifying strong center bias in Gaze-CIFAR-10:** We provide direct empirical evidence that Gaze-CIFAR-10 exhibits strong center bias, and show that a simple center-fixation baseline achieves surprisingly high scanpath similarity under DTW, ScanMatch, NSS, and AUC.
2. **A center-debiased, movement-aware evaluation score.** We introduce **GCS** (Gaze Consistency Score), which (i) normalizes similarity using human–human agreement as an upper bound and a corner baseline as a lower bound, and (ii) explicitly subtracts a center baseline to reduce dataset-induced bias, thereby emphasizing movement dynamics beyond mere spatial overlap.
3. **Peripheral sweet spot as a mechanistic regime:** Using GCS, we identify a restricted, setting-dependent regime in a hard-attention classifier: under Gaze-CIFAR-10, a moderate foveal patch with peripheral context yields the strongest center-debiased alignment. We interpret this as a mechanistic regime analysis rather than a strict human viewport matching.

Together, these results highlight the importance of bias-aware evaluation when using scanpaths to evaluate relative human-likeness, and suggest that perceptual constraints can shape artificial eye movement strategies in this object-centric benchmark.

Method

Gaze-CIFAR-10

We use the Gaze-CIFAR-10 dataset, which provides human fixation sequences collected during CIFAR-10 image recognition. We evaluate on the test split (N = 9,116 images). We use 12 fixation steps for both human and model scanpaths so that all metrics compare trajectories under the same temporal budget.

Hard-Attention Classifier Under Constrained Vision

Our model is a hard-attention recurrent vision agent trained for image classification. To implement the above *constrained-vision* classifier as an *active perceiver*, we use a Multi-Level Recurrent Attention Model (MRAM), a hierarchical recurrent agent that couples *eye-movement control* with *evidence accumulation*. Conceptually, MRAM can be read as an idealized perception–action loop under partial observability: at each time step t, the agent (i) fixates, (ii) receives a limited retinal sample (a "glimpse") determined by a *foveal patch size*, (iii) updates an internal state, and (iv) chooses where to look next.

Retina-like sensing as momentary observation. Given the current fixation location l_{t-1}, a glimpse module extracts a high-resolution foveal crop and, when enabled in specific setting, a wider low-resolution peripheral crop. The peripheral crop is down-sampled to the same spatial size as the foveal crop, and the foveal and peripheral channels are concatenated before being embedded into a sensory vector g_t. Thus the peripheral channel carries *wider field-of-view but lower acuity* information, whereas the foveal channel carries *higher acuity*

but narrower information. This provides a simple computational analogue of the fovea–periphery trade-off familiar from human vision.

Two recurrent levels and training MRAM maintains two recurrent states updated sequentially at each time step: (1) a *lower-level* recurrent state $h_t^{(1)}$ that is driven directly by the current sensory sample g_t, and (2) a *higher-level* recurrent state $h_t^{(2)}$ that integrates information over time and supports the final decision. Intuitively, $h_t^{(1)}$ plays the role of a fast, action-oriented sensorimotor state (what to do next), whereas $h_t^{(2)}$ acts like a slower, more abstract state that accumulates evidence (what is it). This separation mirrors classic cognitive distinctions between *sampling and eye movement control* and *recognition in cortex*, and is consistent with hierarchical views of active perception in which lower circuitry drives saccade-like control while higher circuitry stabilizes task-relevant representations.

The higher-level state $h_t^{(2)}$ is mapped to a categorical prediction over labels, yielding a running belief that can be updated as new glimpses arrive. We take the final-step prediction as the agent's classification decision after a fixed sampling budget at 12 steps, which corresponds to a bounded evidence-accumulation process. From the lower-level state $h_t^{(1)}$, a stochastic location policy samples the next fixation l_t. In our implementation, l_t is a 2D coordinate (normalized to the image plane), so the observable output is a scanpath-like sequence of fixations.

Because fixation selection is a discrete action, MRAM is trained with a REINFORCE algorithm (a Reinforcement Learning method) so that successful classifications reinforce the sampling strategies that produced informative glimpses.

Parameters to test our hypothesis. We hypothesize that relatively human-like scanpaths are most likely when the model must solve the same classification task under constrained information-processing conditions. In particular, we expect human-like scanpaths to emerge most strongly under a moderately constrained field of view: if the visual input is too restricted, the agent lacks sufficient context to guide efficient sampling; if it is too broad, the task can be solved via passive, shortcut-like recognition with little need for active exploration.

We study three sensory configurations that vary peripheral context:

- **fov_only:** observes only the foveal crop.
- **fov+per:** concatenates the foveal crop with a moderate low-resolution peripheral context.
- **big_per:** uses a wider peripheral context, increasing the effective field-of-view and reducing the need for active movement.

We sweep foveal patch size in $\{2, 4, 6, 8, 10, 12, 14, 16, 20, 24, 28, 32\}$ pixels, keeping the number of steps and the main sweep seed fixed. Intuitively, patch size corresponds to an *effective viewing condition hypothesis*: smaller patches require more active exploration, while larger patches allow more passive recognition.

Standard Scanpath Metrics

We report four commonly used scanpath similarity metrics: **DTW** (lower is better), **ScanMatch** (higher is better), **NSS** (higher is better), and **AUC** (higher is better).

To contextualize metric values under dataset bias, we compute three references: (i) an **upper bound** by comparing each human scanpath to itself (identical sequence), (ii) a **lower bound** using a **corner-fixation** policy, and (iii) a **center baseline** using an always-center policy. The center baseline is a fixed policy that repeats the image-center fixation for all 12 steps. The corner baseline repeats a corner fixation for all 12 steps; we use it as a conservative spatially mismatched lower reference, not as a theoretical worst possible policy.

GCS: Center-Debiased, Movement-Aware Score

We define **GCS** (*Gaze Consistency Score*) to reduce center-bias inflation. GCS is designed to answer a conservative question: does a learned policy exceed what can be explained by the dataset's center prior? First, each raw metric is normalized between a human reference and a spatially mismatched corner reference. Second, the normalized center baseline is subtracted, so policies receive credit only for performance above an always-center strategy. Third, we add a small movement-similarity term to discourage policies that match human fixation density only through central overlap while lacking human-like trajectory dynamics. For each metric M, we normalize scores to $[0, 1]$ using the upper/lower bounds and then subtract the center baseline:

$$\tilde{D} = \frac{D_{\min} - D}{D_{\min} - D_{\max}} \quad \text{(DTW)} \tag{1}$$

$$\tilde{M} = \frac{M - M_{\min}}{M_{\max} - M_{\min}} \quad \text{(SM,NSS,AUC)} \tag{2}$$

$$\tilde{M}_{db} = \tilde{M} - \tilde{M}_{center}. \tag{3}$$

We compute a relative-error distance: $d = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\frac{|f_k^{\text{model}} - f_k^{\text{human}}|}{|f_k^{\text{human}}| + \epsilon}\right)^2}$ where f_k denotes a run-level movement statistic including total path, mean saccade amplitude, mean distance-to-center, spatial coverage, direction entropy, and collapse rate, and map it to $\text{Sim}_{move} = \exp(-d/\tau)$. Finally, we add the movement term Sim_{move} to discourage policies that match only spatial center overlap but not temporal dynamics:

$$\text{GCS} = \frac{1}{4}\sum_{M \in \{D, SM, NSS, AUC\}} \tilde{M}_{db} \; + \; \lambda \, \text{Sim}_{move}, \tag{4}$$

with $\lambda = 0.1$. GCS can be negative. A negative value indicates that, after normalization and center-baseline subtraction, the policy falls below the center-calibrated reference overall. Sensitivity analysis shows that varying λ (0 to 0.5) does

Baseline	DTW ↓	SM ↑	NSS ↑	AUC ↑
Identical-Human (upper bound)	0.003	1.000	6.052	0.995
Corner-Human (lower bound)	2023.87	0.013	−0.053	0.541
Center-Human (baseline)	702.24	0.300	1.145	0.6515

Table 1: Calibration baselines on Gaze-CIFAR-10 test. The center baseline is unexpectedly strong under standard scanpath metrics.

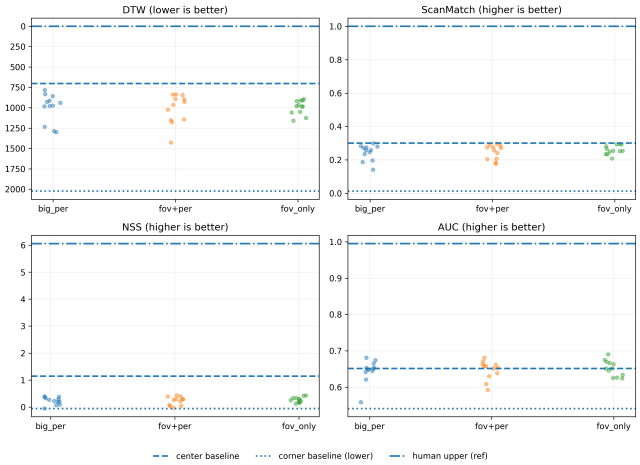

Figure 2: Raw scanpath metric distributions for three sensory settings. Dashed lines: center baseline; dotted lines: corner lower bound; dash-dot lines: human upper bound. Standard metrics can be optimistic due to center bias.

not change the qualitative setting-dependent pattern, indicating that our conclusions are driven by debiased alignment rather than a particular weighting of movement similarity. To make this confound explicit for model evaluation, we compare learned policies against the three calibration references under each metric (Figure 2). Across DTW/ScanMatch/AUC, many learned policies lie close to the center baseline, supporting the need for debiasing.

Results

Center Bias Inflates Standard Scanpath Similarity on Gaze-CIFAR-10

We first report a confound that is critical for interpreting behavioral alignment on object-centric datasets. Figure 1 shows that human fixations are strongly center-weighted over time. Consistent with this bias, Table 1 shows that a trivial center-fixation policy already attains high similarity under common metrics. As a consequence, raw scanpath metrics can substantially overestimate alignment for policies that exploit the dataset's center prior.

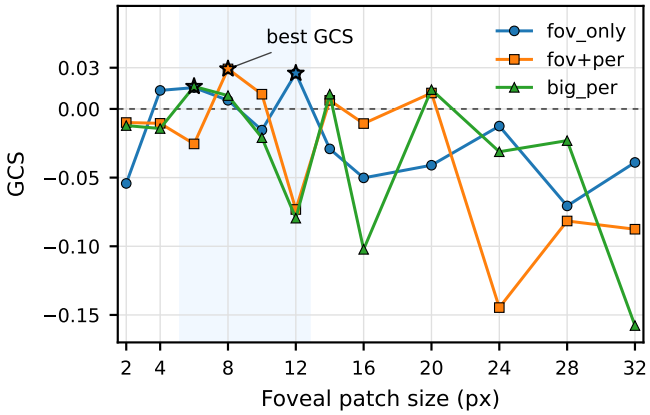

Figure 3: GCS as a function of foveal patch size under three sensory settings. The horizontal line marks the center-calibrated reference (GCS = 0). The sweet spot is setting-dependent: moderate fovea-plus-periphery reaches the strongest overall debiased alignment, while fov_only shifts its optimum toward a larger patch size. This suggests that peripheral context changes the movement–evidence trade-off under constrained vision.

Peripheral Sweet Spot Revealed by GCS

Figure 3 plots GCS across patch sizes and sensory settings. The strongest center-debiased alignment occurs in a restricted mid-range regime: **fov+per with patch size 8** achieves the highest GCS. This point is not the highest-accuracy model. The highest accuracy is obtained by **fov_only with patch size 8**, whose GCS is substantially smaller. Notably, the peak is not identical across sensory settings: fov+per reaches its strongest debiased alignment at a smaller/intermediate patch size, fov_only shifts its optimum toward a larger patch size, and big_per shows a different profile under broader peripheral context. This shift is consistent with the interpretation that peripheral context changes the effective information available per glimpse, but it is an interpretation of the observed regime shift rather than a direct measurement of information content. Thus, the best recognition policy is not necessarily the most human-like under center-debiased scanpath evaluation.

The best overall alignment is achieved by **fov+per with patch size 8**:

$$\text{GCS} = 0.0291, \quad \text{Acc} = 58.50\%, \quad \text{DTW} = 835.5,$$
$$\text{SM} = 0.298, \quad \text{NSS} = 0.401 \quad \text{AUC} = 0.681.$$

In contrast, the highest accuracy configuration is **fov_only with patch size 8** (Acc = 65.31%), but its GCS is much smaller (GCS = 0.0060), illustrating a non-trivial trade-off between task performance and behavioral alignment.

Table 2 summarizes key regimes, including best-GCS and best-accuracy configurations for each sensory setting, as well as illustrative regimes that appear strong under raw metrics. Figure 4 shows a modest correlation between recognition accuracy and scanpath similarity. Because the hard-attention agent is trained only for classification, this suggests that task

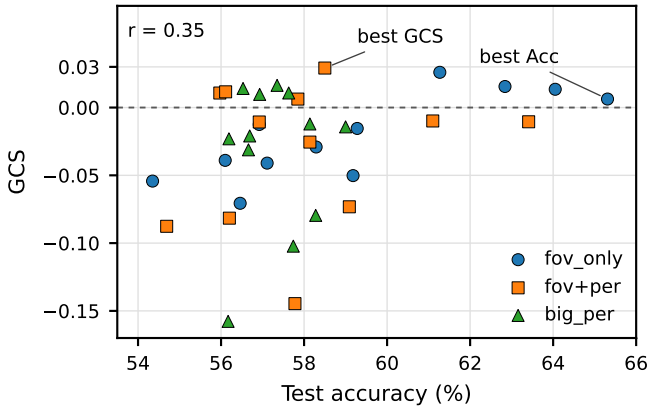

Figure 4: Accuracy versus GCS across sensory settings and patch sizes. Recognition performance and center-debiased scanpath alignment are only modestly coupled: the highest-accuracy configuration is not the best-GCS configuration.

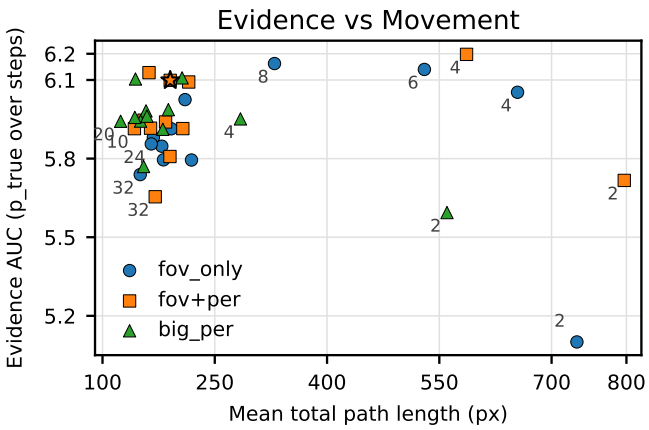

Figure 5: Evidence accumulation versus movement. Each point denotes one sensory configuration and foveal patch size, with numbers indicating patch size. Evidence is measured as the AUC of the true-class probability over glimpse steps. Very small windows induce large movements without proportional evidence gain, whereas overly broad views reduce the need for active sampling. The sweet-spot regime balances movement and evidence accumulation.

learning can encourage the model to acquire some human-relevant evidence, but recognition success alone does not determine whether the scanpath is human-like after center-bias correction.

Movement Similarity at the Sweet Spot

A possible alternative explanation is that the sweet spot simply reflects stronger central tendency. The movement statistics argue against this interpretation. The best-GCS run (fov+per, ps=8) still differs substantially from human scanpaths in absolute terms: total path is 190.5 px vs. human 429.9 px, and coverage is 7.37 vs. 10.15, indicating under-exploration. However, compared with other regimes, it better balances center distance, spatial coverage, direction entropy, and collapse behavior. Thus, the sweet spot should be interpreted as relative human-likeness under center-debiased evaluation, not as a full match to human scanpaths.

Evidence Accumulation vs Movement

To connect scanpaths to decision dynamics, we compare total movement with evidence accumulation, measured as the AUC of the true-class probability over glimpse steps. Very small foveal windows often induce large movements, but this movement does not necessarily translate into efficient evidence accumulation. Conversely, very large fields-of-view can reduce the need for movement, suggesting a shortcut-like regime under this setting. The best-GCS regime lies between these extremes, consistent with a balance between active sampling and evidence gain.

Discussion

Does "Human-Like" Reduce to Matching Human Viewport?

A reasonable concern is that changing patch size alters the agent's effective viewing geometry, whereas the human scanpaths were recorded under a fixed display setup. Accordingly, our analysis varies the model's sensory constraint while the human recording geometry is fixed. The result should not be read as estimating the human viewport. Instead, it tests how active-vision policies change when sensory constraints are manipulated.

Crucially, our claim is not that a particular patch size *matches* humans, but that three coupled observations hold: (i) widely used scanpath metrics are sufficiently center-biased that trivial center-fixation policies can obtain high scores; (ii) once we explicitly calibrate against such trivial baselines, only a narrow range of sensory configurations remains meaningfully above the center baseline; and (iii) within this regime, movement statistics become more human-like displacement patterns in ways that cannot be explained by "being centered" alone.

This perspective yields a concrete, testable prediction for future data collection. If the human viewing geometry is manipulated (e.g., viewing distance or effective image scale), human scanpaths should shift systematically across regimes that mirror our patch-size manipulation. In particular, one may observe *multiple* "sweet spots" across viewing conditions, rather than a single universal setting that defines human-likeness.

Implications for Gaze Benchmarks and Cognitive Modeling

For cognitive modeling, the results highlight a computational interpretation of gaze as *policy under sensory constraints*. For benchmark design, they caution that object-centric datasets can make scanpath evaluation deceptively easy. We advocate reporting center and corner baselines by default, and using debiased metrics such as GCS when scanpaths are used as a

Setting	Viewport	ps	Acc(%)	DTW↓	SM↑	NSS↑	AUC↑	GCS↑	Path(px)
Best raw DTW/SM	big_per	20	56.53	783.6	0.299	0.396	0.653	0.011	123.7
Best raw AUC	fov_only	12	61.27	917.2	0.293	0.435	0.690	0.026	210.2
Best Acc	fov_only	8	65.31	986.9	0.281	0.427	0.669	0.006	329.8
Best Acc (big_per)	big_per	4	59.00	964.2	0.257	0.301	0.653	-0.015	284.3
Best Acc (fov+per)	fov+per	4	63.41	964.2	0.272	0.291	0.651	-0.010	586.4
Best GCS	fov+per	8	58.50	835.5	0.298	0.401	0.681	0.029	190.5
Best GCS (big_per)	big_per	6	57.35	857.5	0.283	0.378	0.673	0.016	158.4
Best GCS (fov_only; = Best AUC)	fov_only	12	61.27	917.2	0.293	0.435	0.690	0.026	210.2

Table 2: Key regimes illustrating the center-bias pitfall of standard scanpath metrics and the debiased composite metric (GCS). ps: foveal patch size. Path: mean total path length over 12 glimpses.

claim of human-likeness.

Limitations

Our experiments focus on an object-centric, low-resolution classification benchmark, a single hard-attention architecture family, and a fixed 12-step horizon. The main sweep uses a fixed seed; selected additional-seed runs provide only a limited sanity check, not a systematic multi-seed robustness analysis. Therefore, the sweet-spot result should be interpreted as a mechanistic hypothesis in this setting, not as a general law of human gaze control. Future work should test richer datasets, visual search tasks, additional architectures, and systematic human viewing manipulations.

Conclusion

We showed that Gaze-CIFAR-10 exhibits strong center bias, such that a trivial center-fixation baseline achieves high scores under common scanpath metrics. We introduced GCS, a practical center-debiased and movement-aware evaluation score that calibrates model scanpaths against human, corner, and center references. Under this metric, a hard-attention classifier shows a restricted mid-range regime in which foveal and peripheral information jointly yield stronger center-debiased alignment than either overly narrow or overly broad sensory settings. The highest-accuracy setting differs from the best-GCS setting, showing that recognition performance alone is insufficient for evaluating human-like scanpaths. These findings motivate bias-aware evaluation practices for active vision and gaze-alignment benchmarks.

Acknowledgments

We acknowledge the use of AI tools for language editing and draft refinement; all analyses, results, and claims were verified by the authors. This work was supported by the JST SPRING GX Program (Grant Number JPMJSP2108). We thank Prof. Hirokazu Takahashi for insightful discussions and helpful comments that inspired this work.

References

Ba, J., Mnih, V., & Kavukcuoglu, K. (2014). Multiple object recognition with visual attention. *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Bajcsy, R. (1988). Active perception. *Proceedings of the IEEE*, *76*(8), 966–1005. https://doi.org/10.1109/5.5968

Ballard, D. H. (1991). Animate vision. *Artificial Intelligence*, *48*(1), 57–86. https://doi.org/https://doi.org/10.1016/0004-3702(91)90080-4

Bindemann, M. (2010). Scene and screen center bias early eye movements in scene viewing [Vision Research Reviews]. *Vision Research*, *50*(23), 2577–2587. https://doi.org/https://doi.org/10.1016/j.visres.2010.08.016

Borji, A. (2021). Saliency prediction in the deep learning era: Successes and limitations. *IEEE Trans. Pattern Anal. Mach. Intell.*, *43*(2), 679–700. https://doi.org/10.1109/TPAMI.2019.2935715

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*, 185–207. https://api.semanticscholar.org/CorpusID:641747

Cristino, F., Mathôt, S., Theeuwes, J., & Gilchrist, I. (2010). Scanmatch: A novel method for comparing fixation sequences. *Behavior Research Methods*, *42*, 692–700.

Dynamic time warping. (2007). In *Information retrieval for music and motion* (pp. 69–84). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-74048-3_4

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R. S., Brendel, W., Bethge, M., & Wichmann, F. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*, 665–673. https://api.semanticscholar.org/CorpusID:215786368

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual review of psychology*, *59*, 167–92. https://api.semanticscholar.org/CorpusID:523591

Judd, T., Durand, F., & Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations.

Li, J., Xue, S., & Su, Y. (2025). Gaze-guided learning: Avoiding shortcut bias in visual classification. *arXiv preprint arXiv:2504.05583*.

Mnih, V., Heess, N., Graves, A., & Kavukcuoglu, K. (2014). Recurrent models of visual attention. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Nips* (pp. 2204–2212).

Najemnik, J., & Geisler, W. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*, 387–91. https://doi.org/10.1038/nature03390

Pan, P., Yonekura, S., & Kuniyoshi, Y. (2026). Emergence of fixational and saccadic movements in a multi-level recurrent attention model for vision. In T. Taniguchi, C. S. A. Leung, T. Kozuno, J. Yoshimoto, M. Mahmud, M. Doborjeh, & K. Doya (Eds.), *Neural information processing* (pp. 299–313). Springer Nature Singapore.

Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*, *45*(18), 2397–2416. https://doi.org/https://doi.org/10.1016/j.visres.2005.03.019

Smith, T. J., & Mital, P. K. (2013). Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes. *Journal of Vision*, *13*(8), 16–16. https://doi.org/10.1167/13.8.16

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*(14), 4–4. https://doi.org/10.1167/7.14.4

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, *11*(5), 5–5. https://doi.org/10.1167/11.5.5

Yarbus, A. L. (1967). Eye movements during perception of moving objects. *Eye Movements and Vision*, 159–170.

Zanca, D., Serchi, V., Piu, P., Rosini, F., & Rufa, A. (2018). Fixatons: A collection of human fixations datasets and metrics for scanpath similarity. https://arxiv.org/abs/1802.02534