Do Glimpse Policies See Like Humans? A Behavioral Audit Reveals Dissociated Viewing Priors in Classification-Trained Active Vision

Boheng Liu **Ziyu Li*** **Chenghua Duan**
Qing Li Xia Wu
Beijing Institute of Technology
boheng@bit.edu.cn, ziyuli@bit.edu.cn

Abstract

Classification-trained sequential glimpse policies have attained competitive performance and efficiency on large-scale recognition benchmarks, and are motivated by analogy to the saccadic architecture of biological active vision. Yet whether these policies reproduce the behavioral priors that organize human free viewing, or merely approximate their spatial distribution as a byproduct of classification reward, has not been systematically examined. We present a diagnostic behavioral audit of AdaptiveNN across three free-viewing datasets, evaluating five dimensions associated with known properties of biological gaze: spatial priors, face prioritization, local return, saccade amplitude structure, and semantic guidance. We show that AdaptiveNN exhibits: (i) spatial behavior that falls below a simple center-bias baseline; (ii) scanpath geometry unlike human scanpaths in amplitude and sequential structure; (iii) revisitation dynamics inconsistent with human free viewing; and (iv) selectively attenuated semantic guidance, with face-specific prioritization center-explained, social sampling attenuated, and object-level rates preserved. In AdaptiveNN, this dissociation suggests that classification reward alone cannot reproduce several behavioral priors organizing human free viewing, even when object-level sampling is preserved. These findings identify which behavioral priors require constraints beyond classification utility, offering targets for designing cognitively aligned glimpse policies.

1 Introduction

The intersection of visual neuroscience and machine learning has produced a new generation of active vision models that increasingly resemble the architecture of biological gaze. Sequential glimpse policies, trained with reinforcement learning to allocate visual sampling across image regions, have achieved strong classification performance by combining recurrent policy networks with powerful visual backbones [2, 31, 42]. Recent advances incorporating Vision Transformer backbones and adaptive computation frameworks have further extended both representational capacity and sampling flexibility [11, 14, 15]. These models are explicitly motivated by analogy to the saccadic architecture of the primate visual system, invoking sequential, resource-limited sampling as a hallmark of biological active vision [35, 47]. This biological motivation sets a clear expectation: if glimpse policies are to serve as computational models of biological vision, their behavior should be evaluated not only by task performance but by the behavioral signatures of biological gaze.

Human free viewing is not a passive sampling process but an active, structured allocation governed by multiple partially independent behavioral priors [18, 19]: a central viewing bias arising from

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

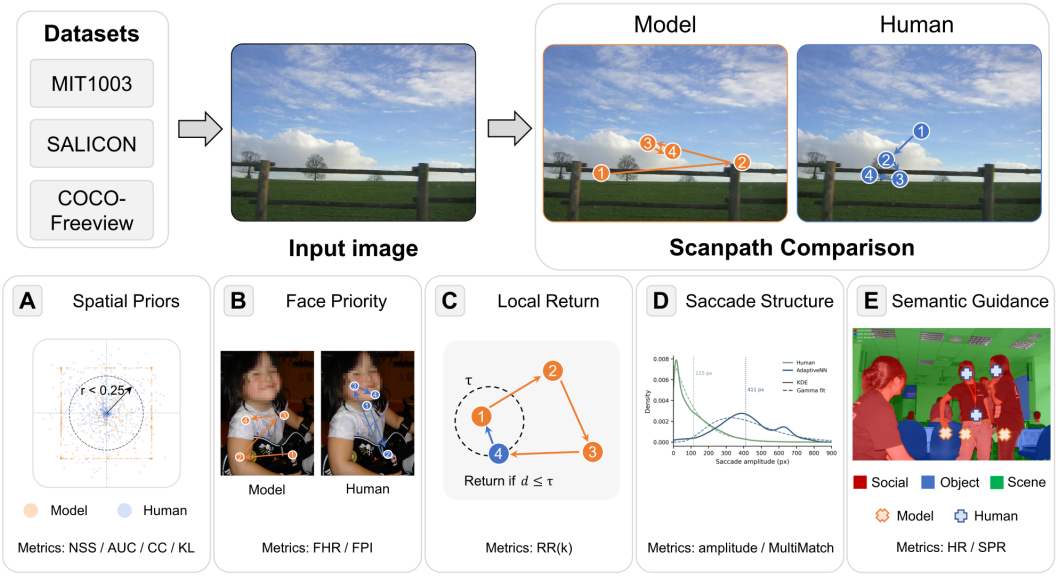

Figure 1: Diagnostic behavioral audit pipeline for a classification-trained glimpse policy. For each image, AdaptiveNN produces four sequential glimpse locations, which are compared with length-matched human free-viewing fixations. The audit decomposes gaze behavior into five dimensions behaviorally associated with known properties of biological vision: spatial priors, face prioritization, local return, saccade amplitude structure, and semantic guidance. Each dimension is evaluated with module-specific controls including uniform and center-Gaussian sampling, amplitude-matched human paths, model-amplitude random walks, distributional fits, and area-normalized semantic preference ratios. The displayed face example is mosaicked for privacy; all quantitative analyses are performed on the original unaltered stimuli.

oculomotor tendencies and natural scene statistics [34, 39]; preferential allocation toward faces and social content via cortical pathways for social cognition [5, 17]; spatially targeted suppression of recently visited locations through inhibition of return, with neural correlates in superior colliculus and parietal cortex [10, 27]; and top-down semantic guidance directing gaze toward meaningful image regions [20, 21]. These behavioral signatures are not incidental byproducts of task demands but are interpreted as functional expressions of the computational architecture of biological gaze. Whether classification-trained glimpse policies reproduce these signatures in the absence of fixation supervision has not been systematically examined.

This question has two independent motivations. First, a model that achieves high classification performance while failing to reproduce central viewing bias, inhibition of return, or social prioritization is not a credible model of biological active vision, regardless of its task utility. Evaluating glimpse policies against biological gaze signatures is therefore not a secondary concern but a prerequisite for their interpretation as models of biological vision. Second, the pattern of which priors are recovered and which are not provides indirect evidence about the computational separability of the mechanisms underlying human free viewing. Components that emerge from classification reward are consistent with being explained by object-recognition utility; those that do not implicate additional constraints such as social cognition, oculomotor history, or scene-level spatial statistics. This behavioral dissociation, a single rather than double dissociation, follows the cognitive neuroscience tradition of reasoning about partial independence of underlying mechanisms [4, 12, 38]. We therefore use *cognitive alignment* to refer specifically to agreement with the behavioral priors that organize human free viewing, and distinguish it explicitly from spatial overlap or task accuracy [36, 44].

We present a diagnostic behavioral audit of AdaptiveNN [42], a reinforcement-learning glimpse policy trained for classification without fixation supervision, evaluated across three free-viewing datasets using matched human baselines and module-specific diagnostic controls (Figure 1). Across five behavioral dimensions, AdaptiveNN showed systematic divergence from human free viewing. Spatial

alignment fell below a simple center-bias baseline, scanpath geometry was markedly unlike human scanpaths, and short-lag proximity was near-absent at $k = 1$, while $k \geq 2$ revisitation rates remained below humans but slightly exceeded a model-amplitude random walk. Face prioritization was indistinguishable from image-center statistics, providing no evidence for face-specific prioritization beyond center bias. Semantic guidance was selectively redistributed: object-level sampling was preserved while social and animate sampling was substantially attenuated. These findings reveal that classification reward recovers some, but not all, components of human free viewing, consistent with partially separable underlying mechanisms.

Our contributions are threefold. First, the dissociation between preserved object-level sampling and attenuated spatial, oculomotor, and social priors provides behavioral evidence consistent with the view that human free viewing reflects partially separable computational mechanisms, with direct implications for understanding the functional organization of biological gaze. Second, we demonstrate that spatial overlap is not cognitive alignment: above-chance saliency scores do not imply agreement with the behavioral priors that organize sequential viewing. Third, we introduce a modular diagnostic audit framework applicable to any sequential glimpse policy, enabling evaluations that assess not only where a model looks but which viewing mechanisms its policy expresses. To the best of our knowledge, this is the first multi-dimensional behavioral audit with module-specific diagnostic controls, conducted without fixation supervision, evaluated across multiple dimensions grounded in known properties of biological gaze. The identified dissociations reveal which behavioral priors require constraints beyond classification utility, offering concrete targets for designing more cognitively aligned glimpse policies. More broadly, we hope these findings contribute to a deeper understanding of the partially separable computational mechanisms underlying human free viewing.

2 Related Work

2.1 Sequential Glimpse Policies and Adaptive Visual Sampling

Unlike single-pass visual models, sequential glimpse policies model biological active vision as a resource-limited sequential sampling process, improving decision efficiency through iterative spatial allocation [35, 47]. The foundational reinforcement learning framework trains a recurrent network to select fixation locations that maximize classification reward, achieving strong performance with reduced computational cost [31]. Subsequent work extended this framework to multiple object recognition, hard attention, and saccade-like dynamics [2, 24]. More recent models incorporate stronger visual backbones and adaptive computation. AdaptiveNN combines a Vision Transformer backbone with a learned adaptive policy, achieving state-of-the-art performance and efficiency on ImageNet-1K, whose authors report that "the perceptual behaviours of AdaptiveNN are indistinguishable from people in many cases" [42]. This indistinguishability claim makes AdaptiveNN the theoretically obligated target for a cognitive alignment audit. Token-based dynamic vision models also perform adaptive computation [14, 15], but do not produce sequential glimpse locations and are therefore outside the scope of glimpse policy evaluation.

2.2 Alignment Between Computational Models and Human Visual Cognition

Early saliency models predicted aggregate human fixation distributions from low-level features [16, 23], and subsequent data-driven models trained on fixation maps substantially improved spatial prediction [28, 29, 32]. However, both lines of work characterize where humans look on average but do not model sequential fixation behavior or the mechanisms that organize it. To address sequential behavior, scanpath models incorporate inhibition-of-return mechanisms, scene-grammar priors, or learned policies [1, 6, 7, 13, 22], but are trained to predict human gaze rather than to audit classification-trained policies. A parallel line of work comparing CNN and transformer representations with human neural responses has found both correspondences and systematic differences [37, 45], yet focuses on representational rather than behavioral alignment. Neither line therefore directly addresses whether classification-trained policies reproduce the behavioral priors that organize human free viewing. Work on the cognitive plausibility of sequential policies has remained limited. Prior work has shown that recurrent attention models differ from human scanpaths in amplitude and revisitation structure [46], and auxiliary objectives have been proposed to encourage human-like saccade distributions [43]. Yet both studies either required direct gaze supervision or examined a single behavioral dimension without diagnostic controls to isolate candidate explanations. No prior

work has conducted a multi-dimensional behavioral audit of a classification-trained glimpse policy in the absence of fixation supervision, with module-specific controls across each dimension.

Table 1: Datasets used in the evaluation.

Dataset	Images	Human signal	Role
MIT1003	999	Eye-tracking (I-DT extracted)	Primary; all mechanistic analyses
SALICON	1000	Mouse-contingent fixation density	Spatial alignment generalization
COCO-FreeView	1000	Free-viewing scanpaths	Cross-scene validation; complexity

3 Method

3.1 Model, Datasets, and Fixation Sequences

We evaluated AdaptiveNN-DeiT-S [42] (hereafter AdaptiveNN), the best-performing variant of the AdaptiveNN family and the specific model its authors compare against human behavior, which combines a DeiT-S transformer backbone with a learned adaptive policy trained on ImageNet-1K classification without gaze supervision, making it the most theoretically motivated target for a cognitive alignment audit. RAM and DRAM are included as auxiliary recurrent-attention controls; as both perform substantially worse than AdaptiveNN on sequential visual tasks [42], results are reported separately in Appendix A.1. Contemporary classification-policy and gaze-supervised controls are reported in Appendix A.2. The policy produced $T{=}4$ sequential fixation coordinates mapped to original image resolution (Appendix A.3). We evaluated three free-viewing datasets, summarized in Table 1; dataset inclusion and exclusion criteria are reported in Appendix A. Across all datasets, the shared unit of comparison was a length-matched first-four fixation sequence, as early fixations are known to dominate initial scene orientation and salient region sampling [39–41], and thus best correspond to the functional scope of a four-step glimpse policy.

For MIT1003 [26] , raw eye-tracking files provided gaze samples rather than fixation sequences, extracted via I-DT. Because the raw gaze files contained ordered samples but no timestamps, we assigned sample-index timestamps using the dataset eye tracker's reported 240 Hz sampling rate. These timestamps were used only for I-DT fixation extraction, not for fixation-duration analyses; sanity checks confirm plausible fixation counts and durations (Appendix B). This procedure yielded 14,861 valid subject-image fixation sequences, with a mean of 14.82 subjects per image. For COCO-FreeView [8] , fixation sequences were read directly from the released fixation JSON.

3.2 Aggregate Spatial and Scanpath Alignment

Spatial metrics and baselines. For each image, model fixation points were converted into a fixation density map by placing impulses at fixation locations, smoothing with a Gaussian kernel ($\sigma = 25$ pixels), and normalizing to unit mass. We evaluated alignment with human fixation maps using four standard saliency metrics: normalized scanpath saliency (NSS), AUC-Judd, Pearson correlation coefficient (CC), and KL divergence [3].

Three diagnostic references were used across all spatial analyses. The uniform baseline sampled points uniformly over the image. The center-Gaussian baseline was a fixed two-dimensional Gaussian centered at the image center, with $\sigma_x = W/6$ and $\sigma_y = H/6$, where W and H are image width and height, and was not fit to test images. The human ceiling was computed by leave-one-subject-out evaluation, using the remaining subjects to predict the held-out subject and averaging across images.

Scanpath alignment. Scanpath structure was measured using MultiMatch [9] across four spatial dimensions: vector, direction, length, and position. The duration dimension was excluded because AdaptiveNN does not produce fixation durations. Spatial alignment maps used all available human fixations; sequential metrics (MultiMatch, local return) used first-four fixation paths only, with human-human references computed from matched path pairs. Saccade amplitudes were computed as Euclidean distances between consecutive fixation coordinates.

Image complexity. As a robustness check, we tested whether alignment varied with image complexity on COCO-FreeView. Complexity was measured as grayscale entropy, and images were split into

three strata using entropy tertiles. Four planned metrics covering spatial alignment, center-baseline gap, scanpath similarity, and saccade scale were evaluated with Bonferroni correction; details are reported in Appendix C.

3.3 Mechanistic Analysis

We decomposed gaze behavior into five diagnostic dimensions, each corresponding to a known functional property of human free viewing. MIT1003 served as the primary dataset for all mechanistic analyses. SALICON [25] was used for spatial-alignment generalization, and COCO-FreeView was used for cross-scene validation, including spatial alignment, scanpath geometry, and image-complexity stratification. Across all modules, AdaptiveNN fixations were compared against uniform sampling, a fixed center-Gaussian prior, and human baselines. Module-specific controls were added to isolate candidate explanations: amplitude-matched human paths and model-amplitude random walks for local return, and area-normalized preference ratios for face and semantic analyses.

Spatial priors. We quantified central viewing bias using normalized radial eccentricity:

$$r(x, y) = \sqrt{\left(\frac{x}{W} - 0.5\right)^2 + \left(\frac{y}{H} - 0.5\right)^2}.$$

Fixations were assigned to one of five spatial regions: center ($r < 0.25$) or left, right, top, and bottom sectors by their dominant displacement from image center. Regional distributions were compared using chi-square tests with Cramér's V as the effect size, and radial distributions were compared using two-sample Kolmogorov-Smirnov tests.

Face prioritization. Faces were detected with MTCNN; detection parameters are reported in Appendix D.2. The primary metrics were face hit rate:

$$\mathrm{FHR} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{(x_i, y_i) \in \mathcal{B}_{\mathrm{face}}\},$$

and the area-normalized face preference index $\mathrm{FPI} = \mathrm{FHR}/a_{\mathrm{face}}$, where a_{face} is union face area divided by image area. The center-Gaussian control tests whether apparent face hits are explained by image-center statistics rather than face-specific prioritization.

Local return. For lag $k \in \{1, 2, 3\}$, return rate was defined as:

$$\mathrm{RR}(k) = \frac{1}{T - k} \sum_{t=k+1}^{T} \mathbf{1}[\|f_t - f_{t-k}\|_2 \leq \tau],$$

with $\tau = 0.05$ in normalized image coordinates. At $k = 1$, this is a short-saccade statistic rather than a return; return interpretations therefore use $k \geq 2$. We compared AdaptiveNN and human return rates against two diagnostic controls for saccade-amplitude effects: amplitude-matched human paths and model-amplitude random walks. Implementation details are in Appendix E.

Saccade amplitude structure. We fit four candidate distributions to pixel saccade amplitudes: truncated power law, lognormal, exponential, and Gamma, covering heavy-tailed finite-bound, multiplicative, light-tailed memoryless, and flexible unimodal shape hypotheses. Fits were compared primarily by Akaike information criterion (AIC), with fitted parameters used to assess distributional shape differences; fitting details are in Appendix F.

Semantic guidance. Images were segmented into three semantic groups: social or animate (persons, animals), object or manipulable (furniture, vehicles, common objects), and scene or background (sky, ground, vegetation, architectural surfaces); segmentation details are in Appendix G, and the keyword-based ADE20K grouping rules are in Appendix G.2. For each group g, the primary outcome was fixation hit rate HR_g. To control for differences in region size, we used an area-normalized semantic preference ratio,

$$\mathrm{SPR}_g = \frac{\mathrm{HR}_g}{a_g},$$

where a_g is the group area ratio. Values above one indicate above-area preference and values below one indicate avoidance. Because raw ratios can be unstable for small semantic regions, Appendix G reports a weakly smoothed SPR variant; primary paired comparisons were based on hit rates. Groups with $a_g < 0.005$ were excluded.

All paired comparisons report Cohen's d_z; for the four planned image-complexity tests we applied Bonferroni correction ($\alpha = 0.0125$). Full statistical procedures are in Appendix H.

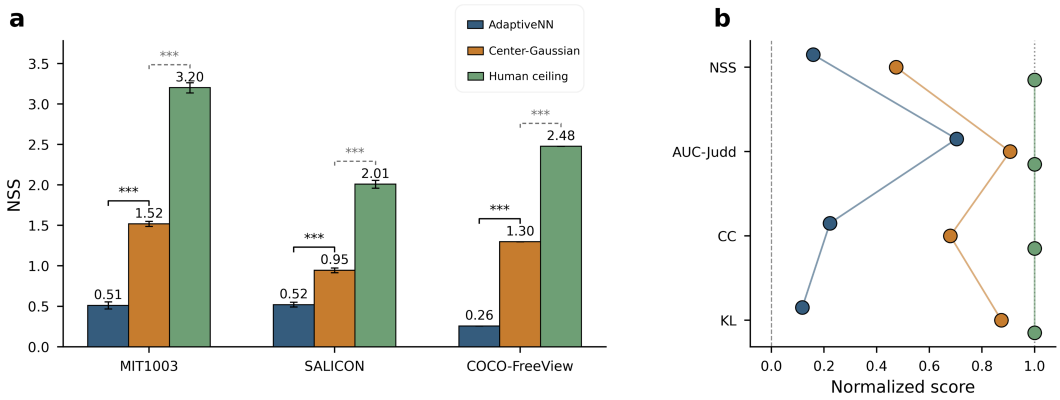

Figure 2: Spatial alignment of AdaptiveNN first-four fixations against human free-viewing maps. **a**, NSS across datasets for AdaptiveNN, center-Gaussian, and leave-one-subject-out human ceiling. Error bars show 95% bootstrap CIs; solid brackets compare AdaptiveNN with center-Gaussian, dashed gray brackets compare center-Gaussian with human ceiling (*** denotes $p < 0.001$). **b**, MIT1003 metrics normalized to the human ceiling; inverted KL ($-$KL, normalized by human ceiling) ensures higher values indicate better alignment.

4 Results

4.1 Spatial Alignment Falls Below a Simple Center-Bias Baseline

AdaptiveNN produced non-random spatial structure across all three datasets, with AUC-Judd reliably above chance. Qualitative scanpath examples across three dimensions are shown in Appendix I. However, non-random spatial structure is not the same as human-like spatial structure: across all datasets and metrics, AdaptiveNN fell consistently and substantially below a fixed center-Gaussian baseline that requires no learning and no image content (Figure 2). On MIT1003, AdaptiveNN achieved NSS $= 0.510$, compared with 1.518 for the center-Gaussian baseline and 3.202 for the leave-one-subject-out human ceiling ($d_z = -1.39$, $p < 0.001$). The deficit held across all four spatial metrics on MIT1003: AUC-Judd 0.642 versus 0.827 ($d_z = -1.60$, $p < 0.001$); CC and KL results are consistent in direction and effect size (Appendix J). The same pattern replicated on SALICON and COCO-FreeView (NSS gaps to center-Gaussian baseline: -0.425 and -1.041; full metrics in Appendix J).

4.2 Scanpath Geometry Reveals Large Saccades and Altered Local Return

As shown in Figure 3, AdaptiveNN produced substantially larger saccades than humans on matched first-four fixation sequences (411.2 px versus 115.4 px; ratio $= 3.56$; $D = 0.730$, $p < 0.001$). MultiMatch confirmed this mismatch: AdaptiveNN-human similarity was lower than human-human similarity for both length (0.763 versus 0.939, $d_z = -2.09$, $p < 0.001$) and position (0.782 versus 0.927, $d_z = -2.13$, $p < 0.001$); vector and direction results are in Appendix J. Gamma shape parameters of 4.82 for AdaptiveNN and 1.06 for humans confirmed a shape mismatch beyond mean amplitude (Appendix F). We then tested whether saccade geometry accounts for the local return deficit. Humans showed frequent short-saccade proximity at $k = 1$ ($\mathrm{RR}(1) = 0.351$), whereas AdaptiveNN showed near-absent proximity ($\mathrm{RR}(1) = 0.016$; $d_z = -2.46$, $p < 0.001$). AdaptiveNN also remained below the amplitude-matched human control (0.068; $d_z = -0.79$, $p < 0.001$) and

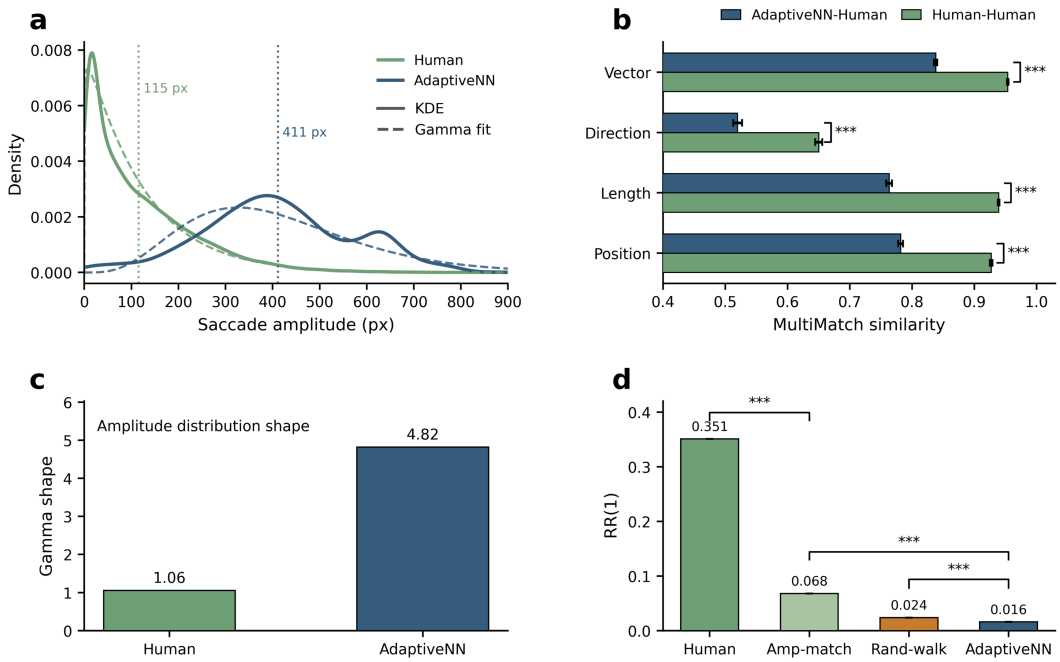

Figure 3: Scanpath geometry of AdaptiveNN first-four fixations on MIT1003. **a**, Saccade amplitude distributions for AdaptiveNN and humans; solid curves show kernel density estimates, dashed curves show Gamma fits, and dotted vertical lines mark mean amplitudes. **b**, MultiMatch similarity across four spatial dimensions for AdaptiveNN-human and human-human scanpaths (*** denotes $p < 0.001$). **c**, Fitted Gamma shape parameters. **d**, Short-lag proximity rate $\mathrm{RR}(1)$ across human fixations, amplitude-matched human paths, model-amplitude random walks, and AdaptiveNN; $\mathrm{RR}(k \geq 2)$ results are in Appendix E. Error bars show 95% bootstrap confidence intervals.

model-amplitude random walk (0.024; $d_z = -1.00$, $p < 0.001$). Because $k = 1$ is a thresholded saccade-amplitude statistic, it does not independently establish return. At $k = 2$ and $k = 3$, model rates remained below humans but slightly exceeded the random-walk control, providing qualified evidence of altered revisitation dynamics; $k = 3$ has only one valid pair per four-step path. Full details are in Appendices F and E.

To test whether task asymmetry explains the observed mismatch, we conducted a matched-image comparison across COCO-FreeView and COCO-Search18. AdaptiveNN remained more peripheral, produced substantially larger saccades, and showed lower $k = 2$ return than humans under both free viewing and visual search; additional COCO-Search18 results across all five audited dimensions are reported in Appendix K.

4.3 Spatial and Social Priors Are Attenuated and Center-Explained

Human first-four fixations were strongly concentrated near the image center, with mean radial eccentricity 0.147 and 79.6% of fixations within the center region ($r < 0.25$). AdaptiveNN fixations were substantially farther from center (mean eccentricity 0.288, 30.5% center; $d_z = 2.27$, $p < 0.001$), falling below the ≈52.6% expected of a uniform policy within the model's feasible action space (Appendix A.3). Cramér's $V = 0.02$ for AdaptiveNN versus uniform, compared with $V = 0.31$ versus humans, confirms weak non-uniformity but no human-like central prior (Figure 4; Appendix D). On the 220 face-positive images, human fixations hit face boxes on 45.1%, compared with 9.4% for AdaptiveNN ($d_z = 1.43$, $p < 0.001$; Figure 4b). AdaptiveNN exceeded uniform sampling in raw face hit rate but fell below the center-Gaussian control. The area-normalized smoothed log-FPI was statistically equivalent to the center-Gaussian baseline ($d_z = -0.023$, 95% CI $[-0.156, 0.110]$; TOST $p = 0.005$; Figure 4c). The model's limited face hits are therefore consistent with image-center statistics rather than face-specific prioritization.

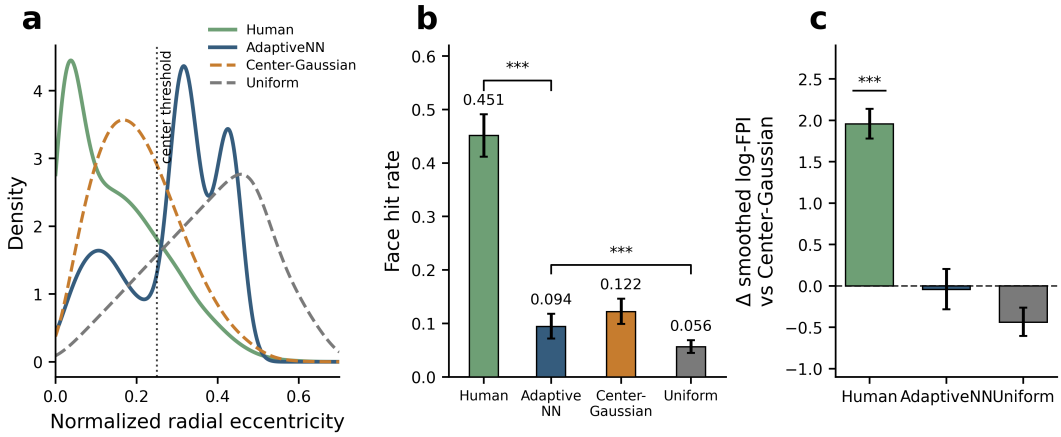

Figure 4: Spatial and social priors on MIT1003. **a**, Radial eccentricity distributions for human fixations, AdaptiveNN, center-Gaussian, and uniform sampling; dotted line marks $r = 0.25$. **b**, Raw face hit rate on 220 face-positive images. **c**, Per-image smoothed log-FPI relative to center-Gaussian; zero indicates face preference statistically equivalent to center statistics. Error bars show 95% bootstrap confidence intervals; brackets show pairwise comparisons (*** $p < 0.001$; ns $p > 0.05$).

4.4 Semantic Guidance Is Selectively Redistributed

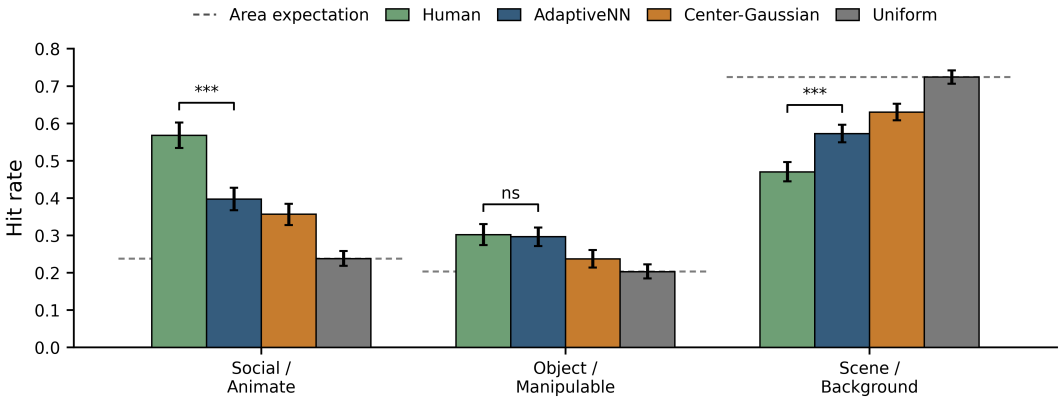

Figure 5: Semantic guidance on MIT1003. **a**, Hit rates across semantic regions; dashed segments mark area-based expectations; brackets show significant human-model hit-rate comparisons only (*** $p < 0.001$). **b**, Smoothed SPR; the dotted line marks SPR $= 1.0$; brackets show independent per-image SPR comparisons (** $p < 0.01$, *** $p < 0.001$). Error bars show 95% bootstrap confidence intervals.

For social or animate regions, humans fixated above the area-based expectation substantially more than AdaptiveNN (hit-rate difference 0.171, $d_z = 0.63$, $p < 0.001$). AdaptiveNN only weakly exceeded the center-Gaussian control for social hit rate ($d_z = 0.18$). For scene and background regions, both conditions showed avoidance relative to image area, but avoidance was weaker for AdaptiveNN (mean difference -0.102, $d_z = -0.44$, $p < 0.001$). Object and manipulable regions showed a qualitatively different pattern: human and model hit rates were nearly identical (0.302 versus 0.297, $d_z = 0.02$, $p = 0.596$), though per-image SPR comparisons revealed a significant preference-strength difference ($p < 0.01$), and both conditions exceeded uniform sampling and the center-Gaussian control. Area-normalized SPR analyses, including the smoothed SPR robustness variant reported in Appendix G, confirmed the selective pattern: attenuated social preference, preserved object hit rate

with attenuated normalized preference, and weakened background avoidance. Semantic guidance was therefore selectively redistributed rather than globally absent.

4.5 Mismatch Is Stable Across Image Complexity

None of the four planned complexity tests on COCO-FreeView survived Bonferroni correction ($\alpha = 0.0125$), and all effect sizes were negligible ($\eta^2 < 0.01$). The NSS gap to the center-Gaussian baseline was effectively flat across low, middle, and high entropy tertiles, and the model-to-human saccade amplitude ratio remained uniformly elevated across all complexity strata. The mechanistic mismatches documented in Sections 4.1–4.4 therefore reflect stable policy-level properties rather than sensitivity to image-statistical regimes. Full stratification results and Spearman correlation summaries are in Appendix C.

5 Discussion

5.1 Cognitive Misalignment as a Systematic Pattern

The five dimensions correspond to known properties of human free viewing: the central prior reflects oculomotor tendencies and scene statistics [39]; face prioritization reflects social attention [5]; local return reflects inhibition and memory for visited locations [27]; saccade amplitude reflects oculomotor constraints; and semantic guidance reflects top-down modulation [20]. AdaptiveNN's divergence across all five dimensions, despite preserved object-level sampling, indicates a systematic mismatch rather than unrelated metric failures. Stability across datasets and complexity strata (Section 4.5) further points to policy-level rather than image-specific effects.

5.2 Mechanistic Interpretation and Implications for Human Free Viewing

In AdaptiveNN, classification reward constrains object-relevant sampling but not center-prior strength, revisitation dynamics, or social priorities: object-level sampling was preserved while central bias, local return, face prioritization, and social guidance were attenuated (Sections 4.3–4.4). The dissociation between coarse animate sampling and center-explained face hits suggests that task utility and cognitive alignment decouple at the level of representational granularity [5].

The local return result further clarifies this dissociation. AdaptiveNN's $k = 1$ rate falls below a model-amplitude random walk, but measures short-saccade proximity rather than an independent return mechanism. At $k \geq 2$, model rates remain below humans but slightly exceed the random walk, suggesting lag-specific reduction rather than global revisitation avoidance. Together with preserved object sampling, this indicates that other human viewing mechanisms require social, oculomotor, or spatial constraints.

5.3 Implications for Glimpse Policy Design

The results suggest three design directions: center-prior regularization to recover human-like spatial concentration [39]; calibration of $k \geq 2$ return dynamics rather than stronger suppression [27]; and semantic supervision for face and animate regions [5]. More broadly, the audit identifies which mechanisms are missing or attenuated, allowing future policies to be evaluated not only by where they look but by which viewing mechanisms they express. Limitations of the present study are discussed in Appendix L.

Acknowledgments and Disclosure of Funding

This work was supported by the National Science Fund for Distinguished Young Scholars of China (Grant No. 62325601), the State Key Program of National Natural Science Foundation of China (Grant No. 62236001), the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 62506037), and the Beijing Municipal Natural Science Foundation (Grant No. L247011).

The authors declare no competing financial interests relevant to this work.

LLM Usage Declaration

Large language models were used solely for language polishing and were not used for any other aspect of this work.

References

[1] Marc Assens Reina, Xavier Giro-i Nieto, Kevin McGuinness, et al. Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2331–2338, 2017.

[2] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple object recognition with visual attention. arXiv preprint arXiv:1412.7755, 2014.

[3] Zoya Bylinskii, Tilke Judd, Aude Oliva, et al. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3): 740–757, 2018.

[4] Alfonso Caramazza. On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: The case for single-patient studies. *Brain and Cognition*, 5(1):41–66, 1986.

[5] Moran Cerf, Jonathan Harel, Wolfgang Einh"auser, et al. Predicting human gaze using low-level saliency combined with face detection. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

[6] Xianyu Chen, Ming Jiang, and Qi Zhao. Beyond average: Individualized visual scanpath prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25420–25431, 2024.

[7] Yu Chen, Zhibo Yang, Seoyoung Ahn, et al. COCO-Search18 fixation dataset for predicting goal-directed attention control. *Scientific Reports*, 11(1):8776, 2021.

[8] Yu Chen, Zhibo Yang, Souradeep Chakraborty, et al. Characterizing target-absent human attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5031–5040, 2022.

[9] Richard Dewhurst, Marcus Nystr"om, Halszka Jarodzka, et al. It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach. *Behavior Research Methods*, 44(4):1079–1100, 2012.

[10] Michael C. Dorris, Raymond M. Klein, Stefan Everling, et al. Contribution of the primate superior colliculus to inhibition of return. *Journal of Cognitive Neuroscience*, 14(8):1256–1263, 2002.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[12] John C. Dunn and Kim Kirsner. What can we infer from double dissociations? *Cortex*, 39(1): 1–8, 2003.

[13] Wolfgang Einh"auser, Merrielle Spain, and Pietro Perona. Objects predict fixations better than early saliency. *Journal of Vision*, 8(14):18–18, 2008.

[14] Alex Graves. Adaptive computation time for recurrent neural networks. arXiv preprint arXiv:1603.08983, 2016.

[15] Yizeng Han, Gao Huang, Shiji Song, et al. Dynamic neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7436–7456, 2021.

[16] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems*, volume 19, 2006.

[17] James V. Haxby, Elizabeth A. Hoffman, and M. Ida Gobbini. The distributed human neural system for face perception. *Trends in Cognitive Sciences*, 4(6):223–233, 2000.

[18] Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. *Trends in Cognitive Sciences*, 9(4):188–194, 2005.

[19] John M. Henderson. Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, 7(11):498–504, 2003.

[20] John M. Henderson. Regarding scenes. *Current Directions in Psychological Science*, 16(4): 219–222, 2007.

[21] John M. Henderson and Taylor R. Hayes. Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour*, 1(10):743–747, 2017.

[22] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001.

[23] Laurent Itti, Christof Koch, and Ernst 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.

[24] Max Jaderberg, Karen Simonyan, and Andrew Zisserman. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[25] Ming Jiang, Shengsheng Huang, Juanyong Duan, et al. SALICON: Saliency in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1072–1080, 2015.

[26] Tilke Judd, Krista Ehinger, Frédo Durand, et al. Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2106–2113. IEEE, 2009.

[27] Raymond M. Klein. Inhibition of return. *Trends in Cognitive Sciences*, 4(4):138–147, 2000.

[28] Matthias K"ummerer, Lucas Theis, and Matthias Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. arXiv preprint arXiv:1411.1045, 2014.

[29] Matthias K"ummerer, Thomas S. A. Wallis, Leon A. Gatys, et al. Understanding low- and high-level contributions to fixation prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4789–4798, 2017.

[30] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. DeepGaze III: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision*, 22(5):7, 2022.

[31] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[32] Junting Pan, Elisa Sayrol, Xavier Giro-i Nieto, et al. Shallow and deep convolutional networks for saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 598–606, 2016.

[33] Adam Pardyl, Michał Wronka, Maciej Wołczyk, Kamil Adamczewski, Tomasz Trzciński, and Bartosz Zieliński. AdaGlimpse: Active visual exploration with arbitrary glimpse position and scale. In *European Conference on Computer Vision*, pages 112–129, 2024.

[34] Derrick Parkhurst, Klinton Law, and Ernst Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107–123, 2002.

[35] Ronald A. Rensink. The dynamic representation of scenes. *Visual Cognition*, 7(1–3):17–42, 2000.

[36] Blake A. Richards, Timothy P. Lillicrap, Philippe Beaudoin, et al. A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761–1770, 2019.

[37] Martin Schrimpf, Jonas Kubilius, Ha Hong, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, page 407007, 2018.

[38] Tim Shallice. *From Neuropsychology to Mental Structure*. Cambridge University Press, 1988.

[39] Benjamin W. Tatler. 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, 2007.

[40] Benjamin W. Tatler, Roland J. Baddeley, and Iain D. Gilchrist. Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, 45(5):643–659, 2005.

[41] Benjamin W. Tatler, Mary M. Hayhoe, Michael F. Land, et al. Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, 11(5):5–5, 2011.

[42] Yulin Wang, Yang Yue, Yufeng Yue, et al. Emulating human-like adaptive vision for efficient and flexible machine visual perception. *Nature Machine Intelligence*, pages 1–19, 2025.

[43] Calden Wloka, Iuliia Kotseruba, and John K. Tsotsos. Active fixation control to predict saccade sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3184–3193, 2018.

[44] Daniel L. K. Yamins and James J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016.

[45] Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

[46] Zhibo Yang, Lihan Huang, Yu Chen, et al. 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.

[47] Alfred L. Yarbus. Eye movements during fixation on stationary objects. In *Eye Movements and Vision*, pages 103–127. Springer US, Boston, MA, 1967.

A Model, Dataset, and Sampling Details

A.1 Auxiliary RAM and DRAM Controls

We additionally evaluated RAM and DRAM as auxiliary recurrent-attention controls. These models were not used as primary comparison targets for two reasons. First, under the same behavioral audit used in the main analysis, both RAM and DRAM showed stronger deviations from human free viewing than AdaptiveNN. Second, the original AdaptiveNN study reports that RAM and DRAM perform substantially worse than AdaptiveNN on sequential visual tasks, making them unsuitable as performance-matched alternatives to the ImageNet-scale AdaptiveNN policy [42].

For this auxiliary analysis, RAM and DRAM each produced four policy-selected fixation locations. Their predicted normalized fixation coordinates were mapped back to the original MIT1003 image resolution using the same coordinate convention as the other model outputs. We evaluated three diagnostic dimensions that do not require semantic or face-specific recognition capacity: spatial priors, scanpath geometry, and lagged fixation proximity. The same MIT1003 human fixation sequences and first-four fixation matching procedure were used as in the main temporal analysis.

Table 2: Auxiliary RAM and DRAM controls on MIT1003. Eccentricity and saccade amplitudes are normalized; MM-L and MM-P denote MultiMatch length and position.

Condition	Ecc.	Center	Sacc.	PR(1)	MM-L	MM-P
AdaptiveNN	0.288	0.306	0.461	0.016	0.763	0.782
RAM	0.574	0.002	0.669	0.000	0.637	0.587
DRAM	0.613	0.001	0.825	0.000	0.497	0.554
Human-human ref.	–	–	–	–	0.938	0.927

As shown in Table 2, RAM and DRAM did not provide human-like alternatives to AdaptiveNN. Both recurrent-attention controls exhibited almost no center occupancy, with center proportions of 0.002 for RAM and 0.001 for DRAM, compared with 0.306 for AdaptiveNN. Their scanpaths were also more divergent from human scanpaths: mean normalized saccade amplitude increased from 0.461 for AdaptiveNN to 0.669 for RAM and 0.825 for DRAM, while MultiMatch length and position similarities fell substantially below AdaptiveNN-human similarity. Finally, both RAM and DRAM showed near-zero adjacent-fixation proximity.

These results support using RAM and DRAM as historical recurrent-attention controls rather than primary baselines. Their inclusion confirms that the proposed audit is not tied to a single implementation of sequential attention, but their substantially poorer behavioral alignment makes them unsuitable as performance-matched comparison models for the main analysis. We therefore retain AdaptiveNN as the primary audit target: it is the strongest classification-trained glimpse policy considered here, the model for which explicit human-like behavioral claims have been made, and the most stringent case for testing whether high task performance entails human-like viewing mechanisms.

A.2 Contemporary Policy and Gaze-Supervised Controls

We applied the same audit to AdaGlimpse [33], a classification-trained policy that selects glimpses of arbitrary position and scale through Soft Actor-Critic, and to DeepGaze III [30], a scanpath model trained directly on human gaze. AdaGlimpse provides an architecture-level control; DeepGaze III tests whether the audit can recover human-like structure when gaze supervision is available.

Table 3 summarizes the full AdaGlimpse comparison. AdaGlimpse remains below the center-Gaussian baseline on spatial alignment across all three datasets, and its face preference is statistically equivalent to center bias. In contrast, its adjacent-fixation proximity and saccade geometry are substantially closer to humans than AdaptiveNN's. The audit therefore separates recurring spatial and face-prioritization deficits from architecture-dependent scanpath structure. AdaGlimpse also has lower ImageNet top-1 accuracy than AdaptiveNN (76.3% versus 82.2%).

Table 4 reports the DeepGaze III control results. DeepGaze III normally conditions each prediction on a real human fixation prefix. Using such a prefix here would inject human behavior into the model output, so we initialized its history with four distinct non-maximum-suppressed peaks from DeepGaze IIE and then selected each fixation by the argmax of the predicted density. This independent warm

Table 3: AdaGlimpse and AdaptiveNN across the three free-viewing datasets.

Dataset	Dimension	Metric	AdaptiveNN	AdaGlimpse	Human	Center-Gaussian
MIT1003	Spatial alignment	NSS	0.510	0.709	3.202	1.518
MIT1003	Spatial alignment	AUC-Judd	0.642	0.680	0.912	0.827
MIT1003	Spatial alignment	CC	0.136	0.114	0.613	0.416
MIT1003	Spatial alignment	KL	9.888	10.455	1.161	1.329
MIT1003	Spatial prior	Center occupancy	0.305	0.602	0.796	0.679
MIT1003	Face prioritization	Face hit rate	0.094	0.142	0.451	0.122
MIT1003	Face prioritization	FPI vs. center	equivalent	equivalent	–	–
MIT1003	Short-saccade structure	PR(1)	0.016	0.385	0.351	–
MIT1003	Saccade geometry	Amplitude ratio	3.56	1.28	1.00	–
MIT1003	Saccade geometry	Gamma shape	4.820	0.877	1.058	–
MIT1003	Semantic guidance	Object hit rate	0.297	0.233	0.302	–
SALICON	Spatial alignment	NSS	0.520	0.480	2.009	0.945
SALICON	Spatial alignment	AUC-Judd	0.682	0.673	0.884	0.766
SALICON	Spatial alignment	CC	0.247	0.245	0.551	0.505
SALICON	Spatial alignment	KL	5.886	7.961	0.928	0.851
COCO-FreeView	Spatial alignment	NSS	0.256	0.484	2.477	1.297
COCO-FreeView	Spatial alignment	AUC-Judd	0.571	0.589	0.879	0.809
COCO-FreeView	Spatial alignment	CC	0.087	0.151	0.434	0.471
COCO-FreeView	Spatial alignment	KL	12.054	11.780	2.317	1.097

Table 4: DeepGaze III as a gaze-supervised control on MIT1003.

Dimension	Metric	Human	DeepGaze III	AdaptiveNN
Spatial alignment	NSS	3.202	2.868	0.510
Spatial alignment	AUC-Judd	0.912	0.902	0.642
Central bias	Center occupancy	0.796	0.730	0.305
Face prioritization	Raw hit rate	0.452	0.628	0.094
Face prioritization	FPI vs. center	distinct ($d_z = 1.42$)	distinct ($d_z = 1.53$)	equivalent
Short-saccade structure	PR(1)	0.351	0.229	0.016
Lagged proximity	RR(2)	0.126	0.836	–
Saccade amplitude	Ratio to human	1.00	0.72	3.56
MultiMatch-Length	Similarity	0.939	0.938	0.763

start avoids a human-authored prefix but creates a disclosed artifact: when an image contains few dominant peaks, deterministic argmax selection can revisit the same peak every other step, inflating RR(2) to 0.836. We therefore do not interpret this value as a substantive model property.

On four of five audited dimensions, DeepGaze III is markedly closer to humans than AdaptiveNN. Its MultiMatch-Length score is statistically indistinguishable from the human–human reference ($p = 0.355$), and its face preference remains distinct from center statistics.

A.3 Coordinate conventions

Images were resized to 288×288 before model inference and normalized with ImageNet mean and standard deviation statistics. The policy output at step t is the normalized top-left coordinate of a focus patch of size $P = 112$ pixels in an input image of size $S = 288$ pixels.

Model actions were produced in the 288×288 input coordinate system and mapped back to the original image resolution. If the policy output at step t is the normalized top-left coordinate (a_t^y, a_t^x) of a focus patch of size $P = 112$ in an input image of size $S = 288$, the corresponding fixation point in input coordinates was computed as

$$x_t^{\text{in}} = a_t^x(S - P) + P/2, \qquad y_t^{\text{in}} = a_t^y(S - P) + P/2.$$

For an original image of width W and height H, the reported fixation coordinate was

$$x_t = \frac{x_t^{\text{in}}}{S} W, \qquad y_t = \frac{y_t^{\text{in}}}{S} H.$$

All reported fixation coordinates were clipped to image bounds. MIT1003 and SALICON were evaluated at their original image resolutions. COCO-FreeView was evaluated on the released 1680×1050 canvas. The coordinate mapping imposes a structural bound on model fixations: with $P = 112$ and $S = 288$, the accessible normalized range is $[P/(2S),\ (S - P/2)/S] \approx [0.194, 0.806]$

along each axis; the accessible side length is $(S - P)/S \approx 0.611$, covering approximately 37.3% of the full image area. The uniform and center-Gaussian baselines cover the full image and therefore constitute a mild structural asymmetry. However, the center region ($r < 0.25$) lies entirely within the model's accessible space, and the model's center occupancy (30.5%) falls below the expected rate of a uniform policy constrained to the same action space ($\approx$52.6%), indicating that the spatial prior deficit is not an artifact of action space bounds.

A.4 Dataset inclusion and exclusion

Table 5: Dataset inclusion and exclusion summary.

Dataset	Available	Evaluated	Inclusion rule
MIT1003	1003	999	Matched stimulus, fixation map, raw gaze, and model output
SALICON-val	5000	1000	First 1000 validation images with valid fixation maps
COCO-FreeView	2243 usable	1000	Seed 42 sample, at least 5 subjects per image

Table 5 summarizes the evaluated images and inclusion rules. For MIT1003, four images were excluded due to missing stimulus, fixation map, or model output files. For SALICON, the first 1000 validation images with valid fixation maps were used to maintain a sample size comparable to MIT1003 while limiting computational cost. For COCO-FreeView, 1000 images were drawn by random sampling (seed 42) from the 2243 usable images meeting the minimum subject threshold of five observers per image. Uniform and center-Gaussian controls used at least 1000 sampled points per image. Lagged-proximity controls used 50 amplitude-matched human samples and 100 random-walk samples per image.

B Human Fixation Extraction Details

B.1 I-DT parameters

Table 6: I-DT fixation extraction parameters.

Parameter	Value
Dispersion threshold	35 px
Minimum duration	100 ms
Assumed sampling rate	240 Hz
Minimum samples per fixation	3
Minimum valid path length	4 fixations
Duplicate handling	remove consecutive duplicates

For MIT1003, human fixation sequences were extracted from raw gaze samples using I-DT. Table 6 lists the parameters. MIT1003 raw files used in this analysis did not provide timestamps, so timestamps were assigned synthetically using the assumed 240 Hz sampling rate.

Dispersion was defined as the sum of the horizontal and vertical coordinate ranges within a temporal window. A window was classified as a fixation when its dispersion was at most 35 pixels and its duration was at least 100 ms. We required at least three gaze samples per fixation cluster. The fixation location was the mean (x, y) coordinate of the samples in the cluster. For each subject and image, we retained finite gaze samples within image bounds, removed consecutive duplicate samples, and assigned timestamps at intervals of $1000/240$ ms. When duplicate subject-image records were encountered, we retained the longer valid fixation sequence.

B.2 Extraction sanity checks

Table 7 summarizes the MIT1003 fixation extraction sanity checks. The extracted sequences produced plausible fixation counts and durations while preserving enough subjects for within-image human-human baselines.

Table 7: MIT1003 fixation extraction sanity summary.

Quantity	Value
MAT files found	15076
Valid subject-image fixation sequences	14861
Images with valid fixation sequences	999
Mean subjects per image	14.82
Minimum subjects per image	13
Maximum subjects per image	15
Mean fixations per subject	11.97
Median fixations per subject	12
Minimum fixations per subject	4
Maximum fixations per subject	19
Mean fixation duration	197.89 ms
Median fixation duration	187.82 ms

C Image Complexity Stratification

C.1 Entropy computation

Image complexity was measured as grayscale entropy computed from a 256-bin intensity histogram:

$$\mathcal{H}(I) = -\sum_{b=1}^{256} p_b \log_2 p_b.$$

Images in COCO-FreeView were divided into low, middle, and high complexity groups using tertiles of the entropy distribution. The four planned complexity outcomes were AdaptiveNN NSS, ΔNSS relative to the center-Gaussian baseline, overall MultiMatch similarity, and the model-to-human saccade amplitude ratio. Bonferroni-corrected threshold was $\alpha = 0.0125$.

C.2 Complexity group summary

Table 8: COCO-FreeView complexity stratification summary. MultiMatch overall is the average of vector, direction, length, and position similarity.

Group	n	Entropy	NSS	ΔNSS	MultiMatch overall	Saccade ratio
Low	333	6.518	0.277	-1.044	0.746	2.752
Middle	334	7.043	0.272	-1.040	0.746	2.833
High	333	7.436	0.219	-1.040	0.744	2.716

As shown in Table 8, the NSS gap to the center-Gaussian baseline was nearly flat across entropy tertiles, MultiMatch overall varied minimally, and the model-to-human saccade ratio remained elevated in all groups.

D Spatial Priors and Face Prioritization

D.1 Spatial prior analysis

Normalized radial eccentricity was defined as

$$r(x, y) = \sqrt{\left(\frac{x}{W} - 0.5\right)^2 + \left(\frac{y}{H} - 0.5\right)^2}.$$

The center region was defined by $r < 0.25$. Non-center fixations were assigned to left, right, top, or bottom according to the dominant signed displacement from the image center.

Tables 9 and 10 summarize the radial eccentricity and regional distribution results, respectively. Global chi-square comparisons over the five regions showed that the model-human difference was

Table 9: Radial eccentricity summary on MIT1003.

Condition	Mean	SD	Median	n
Model	0.288	0.125	0.311	3996
Human	0.147	0.118	0.122	59204
Uniform	0.382	0.142	0.398	999000
Center-Gaussian	0.207	0.106	0.195	999000

Table 10: Region proportions on MIT1003. Values are proportions of all evaluated fixations or sampled control points.

Condition	Center	Left	Right	Top	Bottom
Model	0.305	0.172	0.197	0.112	0.214
Human	0.796	0.060	0.045	0.064	0.035
Uniform	0.197	0.201	0.200	0.202	0.200
Center-Gaussian	0.679	0.080	0.080	0.081	0.080

substantial, Cramer's $V = 0.313$, whereas the model-uniform difference was small, Cramer's $V = 0.020$.

D.2 Face detection and prioritization metrics

Faces were detected with MTCNN using the parameters in Table 11. Fifty quality-control overlays were saved for manual inspection. The smoothed face preference index was computed as an auxiliary ratio-based statistic to prevent degenerate values when the number of fixations is small:

$$\mathrm{FPI}_{\mathrm{smooth}} = \frac{(h+\alpha)/(N+\beta)}{a_{\mathrm{face}}}, \quad \alpha = 0.05, \quad \beta = 0.10,$$

where h is the number of face hits and N is the number of evaluated fixations.

Table 11: MTCNN face detection parameters.

Parameter	Value
Confidence threshold	0.90
Minimum face size	20 px
Bounding box expansion	10 percent
Minimum face area ratio	0.001
Maximum face area ratio	0.50
QC overlays	50 images

As summarized in Table 12, MTCNN produced 800 raw detections across 999 images, of which 453 remained after filtering by confidence threshold and area ratio constraints, yielding 220 face-positive images used in the prioritization analysis.

D.3 Face prioritization results

Table 13 summarizes face hit rates and area-normalized face preference indices on face-positive MIT1003 images. Human face hit rate was much higher than model face hit rate, $d_z = 1.43$, $p < 0.001$. The model was modestly above uniform sampling in raw face hit rate, but its weakly smoothed log-FPI ($d_z = -0.023$, 95% CI $[-0.156, 0.110]$; TOST $p = 0.005$) was statistically equivalent to the center-Gaussian control, indicating that the model's limited face hits are consistent with image-center statistics rather than face-specific prioritization.

Equivalence analysis. To confirm that the null FPI result reflects genuine equivalence rather than insufficient power, we conducted a two one-sided tests (TOST) procedure with bounds $\pm\Delta = 0.2$,

Table 12: Face detection summary on MIT1003.

Quantity	Value
Images scanned	999
Images with valid faces	220
Raw MTCNN detections	800
Valid faces after filtering	453
Filtered by low confidence	216
Filtered by area ratio	131
Detection errors	0

Table 13: Face prioritization summary on face-positive MIT1003 images.

Condition	FHR	FHR minus area	FPI raw	FPI smooth
Human	0.451	0.395	20.083	19.611
Model	0.094	0.038	1.623	3.000
Uniform	0.056	-0.000	0.986	0.991
Center-Gaussian	0.122	0.065	2.140	2.146

the conventional small-effect threshold for d_z. With $n = 220$ face-positive images, the minimum detectable effect at 80% power is $d_z \approx 0.19$, confirming adequate sensitivity.

The TOST tests H_0: $|\delta| \geq \Delta$ against H_1: $|\delta| < \Delta$. For the smoothed log-FPI ($d_z = -0.023$, $\mathrm{SE} = 1/\sqrt{220} = 0.0674$):

$$t_{\text{lower}} = \frac{d_z + \Delta}{\mathrm{SE}} = \frac{0.177}{0.0674} = 2.625, \quad p_{\text{lower}} = 0.005$$

$$t_{\text{upper}} = \frac{d_z - \Delta}{\mathrm{SE}} = \frac{-0.223}{0.0674} = -3.307, \quad p_{\text{upper}} = 0.001$$

$p_{\text{TOST}} = \max(0.005,\ 0.001) = 0.005 < 0.05$, establishing equivalence within the small-effect bound. The 95% CI $[-0.156,\ 0.110]$ lies entirely within ± 0.2, consistent with this conclusion.

E Lagged Fixation Proximity Controls and Amplitude Matching

E.1 Proximity-rate definition

All scanpaths were converted to normalized coordinates in $[0, 1] \times [0, 1]$. For $k \in \{1, 2, 3\}$, proximity rate was defined as

$$\mathrm{RR}(k) = \frac{1}{T-k} \sum_{t=k+1}^{T} \mathbf{1}\left[\|f_t - f_{t-k}\|_2 \leq 0.05\right].$$

At $k = 1$, PR(1) measures adjacent-fixation proximity and is mathematically equivalent to thresholding saccade amplitude. Only $k \geq 2$ can contain a return to a previously left location; the pairwise rates reported here do not impose an additional leave-and-return criterion.

E.2 Control construction

The amplitude-matched human control used normalized mean saccade amplitude as the matching variable. For each model image, we selected candidate human paths from a global pool using the 300 nearest paths in mean amplitude space and sampled 50 matched paths. The random-walk control started from the model's initial fixation, used the model's own normalized saccade amplitudes, sampled directions uniformly from $[0, 2\pi)$, and used boundary reflection to keep paths inside the unit square. We generated 100 random walks per image.

Table 14: Amplitude matching quality. Amplitudes are normalized to the unit image coordinate system.

Quantity	Mean	SD	Median	n
Model mean amplitude per image	0.461	0.109	0.463	999
Same-image human mean amplitude	0.127	0.039	0.125	999
Absolute matching difference	0.107	0.087	0.096	999

Table 15: Lagged-proximity summary on MIT1003. PR(1) is an adjacent-fixation measure.

Condition	PR(1)	RR(2)	RR(3)
Model	0.016	0.019	0.029
Human	0.351	0.126	0.063
Amplitude-matched human	0.068	0.025	0.027
Model-amplitude random walk	0.024	0.015	0.011

Table 16: Key paired comparisons for adjacent-fixation proximity PR(1).

Comparison	Mean difference	d_z	95 percent bootstrap CI
Model vs human	-0.335	-2.46	[-0.343, -0.326]
Model vs amplitude-matched human	-0.052	-0.79	[-0.056, -0.048]
Model vs model-amplitude random walk	-0.008	-1.00	[-0.008, -0.007]
Human vs amplitude-matched human	0.283	2.33	[0.276, 0.291]

E.3 Amplitude matching quality

Table 14 reports the quality of the amplitude-matching procedure. The matched human paths remained lower in amplitude than the model targets, making this a conservative control for adjacent-fixation proximity.

E.4 Full lagged-proximity results

Table 15 reports proximity across all three lags. Humans exceeded AdaptiveNN at PR(1) (0.351 versus 0.016), a short-saccade difference that is not independent of amplitude. At $k = 2$ and $k = 3$, AdaptiveNN remained below humans (0.019 versus 0.126 and 0.029 versus 0.063) but slightly exceeded the model-amplitude random walk (0.015 and 0.011). A four-fixation path supplies two pairs at $k = 2$ and one at $k = 3$, limiting the precision of longer-lag comparisons. Table 16 reports the PR(1) comparisons as a geometric control.

F Saccade Amplitude Distribution Fitting

F.1 Amplitude extraction

Amplitude fitting used first-four scanpaths and pixel saccade lengths. Saccades shorter than 5 pixels were excluded. Table 17 summarizes the resulting amplitude distributions, which included 2986 model saccades and 42840 human saccades after filtering.

F.2 Distribution fitting results

As shown in Table 18, Gamma provided the best AIC for both model and human amplitudes, but the fitted shape parameters were markedly different, indicating a distributional shape mismatch rather than only a difference in mean amplitude.

Table 17: Amplitude summary after filtering saccades shorter than 5 pixels.

Condition	Mean px	SD px	Median px	Min px	Max px	n
Model	412.674	158.426	403.505	6.145	884.551	2986
Human	119.492	112.895	85.308	5.000	938.809	42840

Table 18: Distribution fitting summary. Lower AIC is better. ΔAIC is relative to the best model within each condition.

Condition	Distribution	AIC	ΔAIC	Key parameter
Model	Gamma	39313.984	0.000	shape = 4.820
Model	Lognormal	40247.463	933.479	sigma = 0.551
Model	Exponential	41941.318	2627.334	scale = 412.674
Model	Truncated power law	45249.517	5935.533	alpha = 1.010
Human	Gamma	495427.809	0.000	shape = 1.058
Human	Exponential	495510.905	83.096	scale = 119.492
Human	Lognormal	498082.954	2655.145	sigma = 1.166
Human	Truncated power law	505236.781	9808.972	alpha = 1.010

G Semantic Segmentation and Semantic Guidance

G.1 Mask2Former configuration

Semantic segmentation was performed with Mask2Former using the ADE20K checkpoint `facebook/mask2former-swin-large-ade-semantic`. Table 19 summarizes the segmentation run. Auxiliary log-transformed and smoothed semantic preference ratios were used for robustness checks. The smoothed variant first applied weak additive smoothing constants ($\alpha = 0.05$, $\beta = 0.10$) to the hit rate before area normalization, and was capped at 100:

$$\mathrm{SPR}_{\mathrm{smooth},g} = \min\left(\frac{(h_g + \alpha) \,/\, (N + \beta)}{a_g},\ 100\right),$$

where h_g is the number of fixation hits in group g, N is the total number of evaluated fixations, and a_g is the group area ratio. The log-transformed variant was then defined as:

$$\log \mathrm{SPR}_g = \log(\mathrm{SPR}_{\mathrm{smooth},g} + \varepsilon), \quad \varepsilon = 10^{-6}.$$

These auxiliary variants were not used as primary hit-rate outcomes.

Table 19: Mask2Former segmentation summary.

Quantity	Value
Segmentation model	`facebook/mask2former-swin-large-ade-semantic`
Images segmented	999
Segmentation errors	0
QC overlays	50 images
Minimum group area ratio	0.005

G.2 ADE20K category mapping

ADE20K labels were mapped into three semantic groups before analyzing fixation hits: social or animate, object or manipulable, and scene or background. The mapping was generated before computing the semantic statistics using the keyword rules implemented in the analysis code. Table 20 summarizes the deterministic keyword rules used to map ADE20K labels into semantic groups.

G.3 Semantic area and inclusion

As summarized in Table 21, a group was included for a given image only when its area ratio was at least 0.005.

Table 20: Keyword rules used to map ADE20K labels into semantic groups. Rules were applied in the following order: social or animate, object or manipulable, scene or background, and otherwise other.

Group	Keyword families
Social or animate	person, people, man, woman, boy, girl, child, baby, face, head, torso, leg, arm, animal, dog, cat, bird, horse, cow, sheep, elephant, bear, zebra, giraffe, fish
Object or manipulable	chair, armchair, seat, sofa, couch, bench, table, desk, bed, cabinet, shelf, bookcase, door, window, screen, monitor, television, tv, computer, laptop, keyboard, phone, telephone, book, bottle, cup, glass, plate, bowl, fork, knife, spoon, bag, backpack, umbrella, car, bus, truck, bicycle, motorcycle, boat, train, airplane, vase, clock, lamp, pillow, curtain, rug, mat, basket, box, signboard, stove, microwave, oven, refrigerator, sink, toilet, bathtub, shower, counter, countertop
Scene or background	sky, road, street, sidewalk, pavement, grass, tree, plant, flower, field, earth, mountain, hill, rock, sand, sea, water, river, lake, pool, snow, floor, ceiling, wall, building, house, skyscraper, bridge, path, runway, stairway, stairs, escalator, fence, railing, ground, terrain, land, background, forest, column, pole, curb, platform, pier, waterfall
Other	no keyword match

Table 21: Semantic area summary over MIT1003 images.

Group	Mean area	SD	Median area	Included rows
Social or animate	0.108	0.188	0.001	452
Object or manipulable	0.114	0.198	0.014	558
Scene or background	0.714	0.299	0.834	985
Other	0.065	0.157	0.000	NA

G.4 Full semantic guidance results

Table 22: Semantic guidance summary. HR denotes hit rate. Smoothed SPR denotes a weakly smoothed area-normalized semantic preference ratio used as a robustness statistic.

Group	Condition	HR	Smoothed SPR
Social or animate	Human	0.568	4.823
Social or animate	Model	0.397	2.860
Social or animate	Uniform	0.238	1.005
Social or animate	Center-Gaussian	0.357	1.618
Object or manipulable	Human	0.302	3.119
Object or manipulable	Model	0.297	2.367
Object or manipulable	Uniform	0.203	0.997
Object or manipulable	Center-Gaussian	0.237	1.231
Scene or background	Human	0.470	0.543
Scene or background	Model	0.573	0.737
Scene or background	Uniform	0.724	0.999
Scene or background	Center-Gaussian	0.630	0.779

As summarized in Tables 22 and 23, human and model fixations differed for social or animate regions and scene or background regions, but not for object or manipulable regions. The object hit-rate comparison yielded $d_z = 0.02$, $p = 0.596$.

H Statistical Procedures

All analyses were conducted at the image level unless explicitly stated otherwise. Paired comparisons used paired t-tests with Cohen's d_z:

$$d_z = \frac{\overline{x - y}}{s_{x-y}}.$$

Table 23: Semantic guidance paired comparisons on MIT1003. Human-model hit-rate differences with effect sizes.

Group	Mean difference	d_z	p
Social or animate	0.171	0.63	< 0.001
Object or manipulable	0.005	0.02	0.596
Scene or background	-0.102	-0.44	< 0.001
Social vs center-Gaussian (model)		0.18	

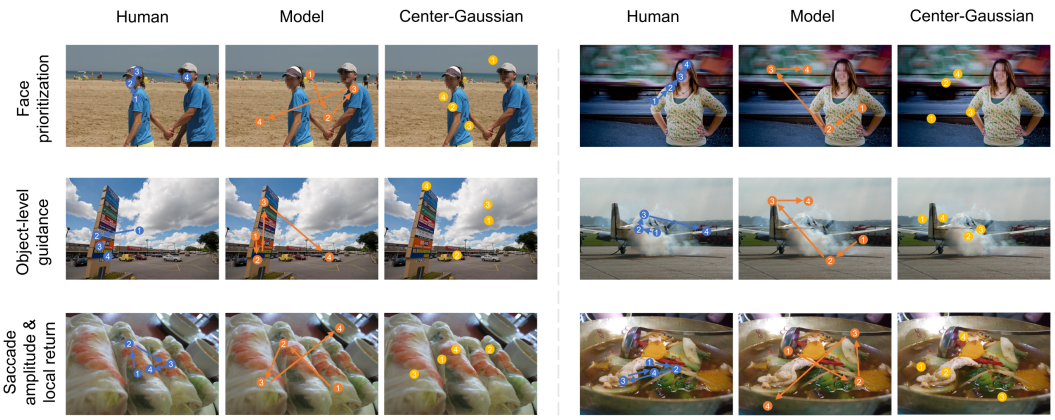

Figure 6: Representative scanpath examples from MIT1003. Each panel shows first-four fixations for humans (blue), AdaptiveNN (orange), and center-Gaussian samples (yellow); numbers indicate fixation order. **Top row (face prioritization):** Human fixations cluster on face regions; AdaptiveNN fixations fall outside and spread peripherally; center-Gaussian samples are image-center biased regardless of face location. **Middle row (object-level guidance):** Both humans and AdaptiveNN fixate on salient objects (signage, aircraft), consistent with preserved object-level sampling (Section 4.4). **Bottom row (saccade amplitude and lagged proximity):** Human paths are compact; AdaptiveNN paths span the full image. Face regions are mosaicked for privacy.

Independent comparisons used Cohen's d, and ANOVA analyses reported η^2. For categorical pooled hit analyses, we used chi-square tests and odds ratios. For radial and amplitude distributions, we used two-sample Kolmogorov-Smirnov tests. For distribution fitting, we reported AIC and fitted distribution parameters. All bootstrap confidence intervals used 5000 resamples with 95 percent percentile intervals. The nominal significance threshold was $\alpha = 0.05$. For the four planned image-complexity tests, we applied Bonferroni correction, giving $\alpha = 0.0125$. Effect sizes are reported together with p-values throughout; results are interpreted primarily through the consistency of effects across datasets, metrics, and diagnostic controls.

I Representative Scanpath Examples

Figure 6 illustrates three dissociations quantified in the main text. In the face prioritization examples, human fixations consistently target facial regions while AdaptiveNN fixations fall outside them, confirming no face-specific prioritization beyond center statistics (Section 4.3). In the object-level guidance examples, both humans and AdaptiveNN fixate on task-relevant objects, demonstrating that semantic guidance is selectively redistributed rather than globally absent (Section 4.4). In the saccade-amplitude examples, human paths are spatially compact, whereas AdaptiveNN paths traverse the full image extent, consistent with the quantitative geometry results (Section 4.2). These examples confirm that the reported dissociations reflect stable policy-level properties.

J Full Spatial and Temporal Alignment Results

J.1 MIT1003 spatial alignment

Table 24: MIT1003 spatial alignment metrics. Higher is better for NSS, AUC-Judd, and CC. Lower is better for KL.

Condition	NSS	AUC-Judd	CC	KL
AdaptiveNN first 2	0.477	0.602	0.125	11.884
AdaptiveNN first 3	0.488	0.624	0.132	10.775
AdaptiveNN first 4	0.510	0.642	0.136	9.888
Center-Gaussian	1.518	0.827	0.416	1.329
Human ceiling	3.202	0.912	0.613	1.161

Table 24 reports the full first-k spatial alignment results on MIT1003. The first-four AdaptiveNN maps performed worse than the center-Gaussian baseline across all four spatial metrics: NSS and AUC-Judd and CC were lower, while KL divergence was higher ($d_z = -1.39$, $d_z = -1.60$, $d_z = -1.32$, and $d_z = 2.87$ respectively, all $p < 0.001$).

J.2 SALICON spatial alignment

Table 25: SALICON validation spatial alignment metrics.

Condition	NSS	AUC-Judd	CC	KL
AdaptiveNN first 2	0.493	0.646	0.227	8.877
AdaptiveNN first 3	0.502	0.669	0.237	7.159
AdaptiveNN first 4	0.520	0.682	0.247	5.886
Center-Gaussian	0.945	0.766	0.505	0.851
Human ceiling	2.009	0.884	0.551	0.928

As shown in Table 25, the first-four AdaptiveNN maps performed worse than the center-Gaussian baseline across all spatial metrics: NSS, AUC-Judd, and CC were lower, while KL divergence was higher ($d_z = -0.77$, $d_z = -0.84$, $d_z = -1.06$, and $d_z = 2.36$ respectively, all $p < 0.001$).

J.3 COCO-FreeView spatial alignment

Table 26: COCO-FreeView spatial alignment metrics.

Condition	NSS	AUC-Judd	CC	KL
AdaptiveNN first 2	0.231	0.550	0.080	13.070
AdaptiveNN first 3	0.236	0.560	0.081	12.566
AdaptiveNN first 4	0.256	0.571	0.087	12.054
Center-Gaussian	1.297	0.808	0.471	1.097
Human ceiling	2.477	0.879	0.434	2.317

For first-four AdaptiveNN fixations, the mean ΔNSS relative to the center-Gaussian baseline was -1.041 (Table 26). The paired comparison between AdaptiveNN and the center-Gaussian baseline yielded $d_z = -2.31$, $p < 0.001$.

J.4 Temporal alignment with MultiMatch

Table 27 reports MultiMatch results on MIT1003. The duration dimension was excluded because AdaptiveNN does not produce fixation durations. Table 28 reports the corresponding COCO-FreeView temporal alignment results.

Table 27: MIT1003 first-four MultiMatch comparison.

Comparison	Vector	Direction	Length	Position
AdaptiveNN-human	0.838	0.520	0.763	0.782
Human-human	0.954	0.650	0.939	0.927
Paired d_z	-2.54	-0.86	-2.09	-2.13

Table 28: COCO-FreeView first-four MultiMatch comparison.

Comparison	Vector	Direction	Length	Position
AdaptiveNN-human	0.840	0.544	0.795	0.802
Human-human	0.931	0.622	0.925	0.901
Paired d_z	-2.36	-0.54	-1.72	-1.53

Table 29: First-four saccade amplitude summary.

Dataset	Model mean px	Human mean px	Ratio	KS D
MIT1003	411.2	115.4	3.56	0.730
COCO-FreeView	644.6	245.6	2.62	0.700

J.5 Saccade amplitude summary

As shown in Table 29, the model-to-human amplitude ratio exceeded 2.5 on both datasets, and the large KS statistics ($D > 0.70$) confirm that the distributional mismatch extends beyond mean amplitude differences to the full shape of the amplitude distribution.

K Matched-Image Task-Condition Control

K.1 Motivation and design

A potential concern is that AdaptiveNN's spatial and oculomotor deviations from free-viewing humans reflect task asymmetry, as the model is trained for classification while humans perform free viewing, rather than genuine cognitive misalignment. To address this, we added a task-driven human anchor condition using COCO-Search18 [7], a target-present visual-search dataset in which human observers searched COCO images for targets from 18 object categories. We constructed a matched-image three-way comparison among free-viewing humans (COCO-FreeView, FV), task-driven humans (COCO-Search18, TD), and AdaptiveNN on the intersection of images for which all three signals were available.

The diagnostic logic follows a gap-ratio framework. For each metric m:

$$\mathrm{Gap}_{\mathrm{task}} = |m_{\mathrm{TD}} - m_{\mathrm{FV}}|, \qquad \mathrm{Gap}_{\mathrm{model}} = |m_{\mathrm{AdaptiveNN}} - m_{\mathrm{FV}}|, \qquad \mathrm{Ratio} = \frac{\mathrm{Gap}_{\mathrm{model}}}{\mathrm{Gap}_{\mathrm{task}}}.$$

A ratio above 1.0 indicates a larger model-to-free-viewing gap than task-to-free-viewing gap.

K.2 Data

Unlike the main COCO-FreeView analysis, this control used all usable COCO-FreeView images before intersection, because the goal was a matched-image comparison across datasets rather than a fixed-size random-sample evaluation. COCO-FreeView and COCO-Search18 contain independent observers. The main analysis used COCO-Search18 validation split2, yielding 344 matched images and 1,456 valid task-driven fixation sequences. We evaluated all five audited dimensions.

K.3 Results

Table 30 reports the five-dimension matched-image comparison. Object hit rate is mechanically confounded because all 18 COCO-Search18 targets are object categories, and is excluded from

Table 30: COCO-Search18 task-condition control on matched images. FV denotes free viewing and TD target-directed search. Ratio is $|\text{Model} - \text{FV}|/|\text{TD} - \text{FV}|$. PR(1) denotes adjacent-fixation proximity.

Dimension	Metric	n	FV	TD	AdaptiveNN	Ratio
Spatial prior	Mean eccentricity	344	0.163	0.209	0.279	2.52
Spatial prior	Center proportion	344	0.739	0.580	0.327	2.59
Short-saccade structure	PR(1)	344	0.143	0.154	0.006	12.09
Lagged proximity	RR(2)	344	0.055	0.036	0.015	2.10
Saccade amplitude	Mean amplitude (px)	343	251.1	273.0	640.6	17.8
Saccade amplitude	Gamma shape	343	1.73	1.63	6.80	54.0
Semantic guidance	Object hit rate	342	0.536	0.672	0.551	0.11
Semantic guidance	Scene hit rate	333	0.228	0.193	0.284	1.59
Semantic guidance	Social hit rate	28	0.068	0.064	0.070	0.50
Face prioritization	Face hit rate	17	0.028	0.022	0.000	5.07

task-fairness conclusions. Face and social estimates are reported for completeness but have small matched samples ($n = 17$ and $n = 28$). Table 31 reports the corresponding split1 robustness check.

Spatial priors. Visual search shifted human gaze peripherally, as expected: mean eccentricity increased from 0.163 to 0.209 and center-region occupancy decreased from 0.739 to 0.580. AdaptiveNN deviated further still, reaching eccentricity 0.279 and center occupancy 0.327. The gap-ratio was 2.52 for eccentricity and 2.59 for center occupancy, stable across both splits (2.35 and 2.36 on split1).

Lagged proximity. PR(1) changed from 0.143 in free viewing to 0.154 in search, while AdaptiveNN reached 0.006. Because PR(1) is a short-saccade measure, it does not independently establish return. At $k = 2$, search reduced human proximity from 0.055 to 0.036, while AdaptiveNN reached 0.015.

Saccade amplitude and semantic guidance. Task effects on mean amplitude and Gamma shape were modest relative to AdaptiveNN, producing ratios of 17.8 and 54.0. Scene hit rate showed a smaller ratio of 1.59.

Overall. The matched-image comparison shows material task dependence, especially for spatial priors and $k \geq 2$ proximity. It does not explain AdaptiveNN's much larger saccade-amplitude mismatch, and the spatial-prior deviation remains larger than the corresponding task shift.

Table 31: Robustness check on split1 (284 matched images).

Dimension	Metric	FV	TD	AdaptiveNN	Gap_{task}	$\text{Gap}_{\text{model}}$	Ratio
Spatial prior	Mean eccentricity	0.162	0.212	0.280	0.050	0.117	2.35
Spatial prior	Center proportion	0.746	0.568	0.327	0.177	0.419	2.36
Short-saccade structure	PR(1)	0.137	0.152	0.007	0.015	0.130	8.41
Lagged proximity	RR(2)	0.052	0.037	0.012	0.015	0.040	2.63

L Limitations

AdaptiveNN-DeiT-S represents the state-of-the-art classification-trained glimpse policy and the contemporary architecture of this class trained at ImageNet scale, and is moreover the specific variant whose authors report that "the perceptual behaviours of AdaptiveNN are indistinguishable from people in many cases" [42]; it is therefore the most theoretically motivated target for a cognitive alignment audit. The model's action space covers approximately 37.3% of the full image area (Appendix A.3), a mild asymmetry relative to full-image baselines; however, AdaptiveNN's center occupancy (30.5%) falls below the constrained-uniform expectation ($\approx$52.6%), so this structural bound does not explain the spatial prior deficit.

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: The abstract and introduction clearly state three contributions: (1) behavioral evidence for partially separable mechanisms in human free viewing; (2) demonstrating that spatial overlap is not cognitive alignment; (3) a modular diagnostic audit framework. All three are directly supported by the experimental results in Sections 4.

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 5.4 contains a dedicated Limitations subsection discussing: (1) single-model scope; (2) reliance on Mask2Former segmentation for semantic guidance analyses; and (3) restriction to first-four fixations. Task mismatch between classification training and free-viewing evaluation is addressed in Appendix K via a dedicated task-condition control.

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: [N/A]

   Justification: This paper presents an empirical behavioral audit and does not include theoretical results, theorems, or proofs.

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: All experimental details are reported in the appendices: model coordinate conventions (Appendix A), fixation extraction parameters and sanity checks (Appendix B), spatial/face analysis parameters (Appendix D), return-rate control construction (Appendix E), amplitude fitting procedures (Appendix F), segmentation configuration and label mappings (Appendix G), and statistical procedures (Appendix H). Random seeds are specified in Appendix A.

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: [No]

   Justification: All three datasets (MIT1003, SALICON, COCO-FreeView) are publicly available. AdaptiveNN is a published model. The analysis code for this study has not been publicly released. Experimental procedures are described in Appendices A–I.

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) necessary to understand the results?

   Answer: [Yes]

   Justification: All evaluation details are specified in the appendices: model coordinate conventions and random seeds (Appendix A), I-DT parameters (Appendix B), face detection parameters (Appendix D), return-rate control construction (Appendix E), amplitude fitting (Appendix F), segmentation configuration and ADE20K label mapping (Appendix G), and statistical procedures (Appendix H).

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: All paired comparisons report Cohen's d_z, p-values, and 95% bootstrap confidence intervals (5000 resamples). Figures include 95% bootstrap CI error bars. Effect sizes are reported alongside p-values throughout, and Bonferroni correction is applied for the four planned complexity tests (Appendix C).

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: [No]

   Justification: The paper does not report compute resources. The experiments consist of inference with a pretrained model (AdaptiveNN-DeiT-S) and statistical analyses on three datasets of 1000 images each; the total compute is modest and does not meaningfully affect reproducibility.

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: The research uses only publicly available datasets and pretrained models, involves no new human subjects, and raises no foreseeable ethical concerns. Face regions in qualitative figures are mosaicked for privacy.

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: This work is a diagnostic behavioral audit of an existing model using existing datasets. It introduces no new model, dataset, or deployable system, and has no direct path to negative societal applications. The findings are relevant to foundational research on cognitively aligned vision models.

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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases no new models, datasets, or systems with misuse potential. All assets used are existing publicly available resources.

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 datasets (MIT1003, SALICON, COCO-FreeView, COCO-Search18) and models (AdaptiveNN, Mask2Former, MTCNN) are cited with their original papers. All are publicly released for academic research use, and are used strictly within their intended scope.

13. **New assets**

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

Answer: [N/A]

Justification: The paper introduces no new dataset or model. The diagnostic audit framework is described methodologically; associated analysis code will be released upon acceptance.

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: [N/A]

    Justification: The paper uses only pre-existing publicly released eye-tracking datasets (MIT1003, SALICON, COCO-FreeView). No new human subjects experiments were conducted.

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: [N/A]

    Justification: No new human subjects research was conducted. All human gaze data were sourced from publicly released datasets collected under their respective original institutional protocols.

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 rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used as any component of the core methodology, experimental design, or quantitative analysis.