Modeling Visual Salience with Excitation-Inhibition Dynamics: A Bottom-Up Approach to Gaze Prediction

Ayaan Iqbal (20995467)
SYDE 552 - University of Waterloo
Mekhael Thaha (21014589)
SYDE 552 - University of Waterloo

Abstract

This final project explores the efficacy of bottom-up spiking neural networks and whether they can generate biologically realistic saccade targets in real-world natural visual scenes. It is hypothesized that such a model would provide saliency-based predictions of gaze that would coincide with human fixation behaviour.To process the data input, a 20x20 LIF neuron grid was simulated using Brian2 to be stimulated by preprocessed saliency maps created with Opencv's Static Saliency Spectral Residual Method. Performance of this pipeline was validated on the testing images of the MIT1003 dataset with a parameter sweep of excitation and inhibition from 1.5-3.5 and 0.3-0.5, respectively. The quantitative metrics used to compare and evaluate were AUC (area under the curve) and NSS (Normalized Scanpath Saliency). Results showed that the optimal saliency selection occurred at excitation values of ~2.0 and inhibition values of ~0.45, creating an Excitation/Inhibition (E/I) ratio of ~4.375:1. While this ratio has research precedent, deeming it accurate, the AUC and NSS values showcased a poor performance by the model (AUC at ~0.49 and NSS at ~ -0.1 - 0.2). This occurred due to the model exhibiting high luminance bias and its inability to prioritize semantic and contrast-based relevance, causing struggles in predicting fixation behaviour. These findings highlight the limitations of purely bottom-up saliency and showcase the need and importance of top-down modulation and temporal mechanisms to include important saliency measurements like habituation. By implementing a top-down model and inhibition of return, results are expected to greatly improve because of the removal of luminance bias and prioritization of semantically relevant data. The exploration of other datasets like the CAT2000, which contains multiple different categories, will ease the training and development of the model, as simpler category images can be used to begin with, and image complexity can improve as model performance improves. Future work should focus on this model and dataset implementation to better approximate human visual attention.

1 Introduction

The style files for NeurIPS and other conference information are available on the World Wide Saccades are rapid and automatic eye movements that reposition the fovea toward new points of interest in the visual field. These movements are essential for visual tasks such as reading, scanning the environment, and recognizing objects. In computational neuroscience, saccadic behaviour is typically modelled as the result of two main processes: top-down control, which is guided by internal goals, memory, or attention, and bottom-up processing, which is driven by the physical salience of visual stimuli.

Although both processes work together in natural vision, most recent models have emphasized top-down influences more. This focus reflects the importance of task relevance and cognitive context in many real-world scenarios. However, bottom-up saccades are a critical component of early visual processing, especially in situations where quick reactions to environmental changes are required.

Previous models, such as the saliency-based framework developed by Itti and Koch, have shown that simple visual features like brightness, contrast, and motion can be used to predict saccade targets [1]. These models are inspired by biological mechanisms found in the dorsal visual stream and the superior colliculus, which are known to support reflexive gaze shifts.

This project builds upon previous research by modelling saccadic gaze prediction driven solely by bottom-up signals. The focus lies in understanding how salience, in the absence of cognitive input, can initiate saccadic movements. The hypothesis proposes that a model relying exclusively on bottom-up salience cues can generate biologically plausible saccadic patterns, consistently targeting high-salience regions across diverse visual scenes. The objective is to assess the effectiveness of this modelling approach and investigate its implications for early visual attention mechanisms and potential applications in artificial vision systems.

2 Methods

A biologically inspired spiking neural network was implemented using the Brian2 simulator to evaluate how varying levels of excitation and inhibition influence the network's ability to predict human eye fixations on visual scenes. The model comprised 400 Leaky Integrate-and-Fire (LIF) neurons arranged in a 20×20 grid. Synaptic dynamics were influenced by both input saliency and lateral interactions among neurons.

Visual stimuli and corresponding ground truth fixation maps were obtained from the MIT1003 dataset, which contains natural images and human fixation data collected through eye-tracking. As a baseline, the Static Saliency Spectral Residual method from OpenCV was used to generate initial saliency maps based on low-level visual features. These maps served as both input to the model and a reference for comparison against the MIT1003 fixation maps.

To explore how the balance of excitation and inhibition affects predictive performance, a systematic parameter sweep was performed. Excitatory input strengths ranged from 1.5 to 3.5, while lateral inhibition levels varied from 0.3 to 0.5. These parameters control the responsiveness of individual neurons and the degree of spatial competition within the network, supporting winner-take-all behaviour frequently observed in biological attention circuits.

The dynamics of each neuron were governed by a set of differential equations that model changes in membrane potential and synaptic input over time. Specifically, the membrane potential v evolves according to Equation 1, which was derived from the LIF Model Equation and Synaptic Conductance Formula:

$$\frac{dv}{dt} = (ge \times (Ee - v) + gi \times (Ei - v) + (v\_rest - v)) \times \frac{1}{tau}$$

Equation 1.

In this equation, *v* denotes the membrane voltage, *ge* and *gi* represent excitatory and inhibitory conductances, *Ee* and *Ei* are their respective reversal potentials, and *v_rest* is the resting potential. The parameter *tau* represents the membrane time constant.

In addition to voltage dynamics, synaptic conductances decay over time, following exponential decay governed by Equations 2 and 3:

$$\frac{dge}{dt} = -\frac{ge}{tau\_e}$$

Equation 2

$$\frac{dgi}{dt} = -\frac{gi}{tau\_i}$$

Equation 3

As seen in these equations, excitatory and inhibitory conductances decay exponentially over time, controlled by their respective time constants *tau_e* and *tau_i*. When a neuron spikes, its excitatory conductance increases, and an inhibitory signal is sent to all other neurons in the network. This lateral inhibition encourages spatial competition and suppresses widespread activation, allowing only the most salient regions to dominate network activity.

Each input image from the MIT1003 dataset was resized and processed into a saliency map using the Static Saliency Spectral Residual method. To reduce any noise in the image, the resulting saliency map was smoothed with a Gaussian blur. It was then normalized, thresholded to improve contrast, and resized to match the input dimensions of the neuron grid. The processed map was scaled by the excitation parameter and applied as input current. The network was simulated for 300 milliseconds per image, and spike activity was recorded. Spike counts were normalized to form predicted saliency maps representing estimated visual attention.

Model performance was evaluated using two metrics: AUC and NSS [2][3]. Both metrics compared the model's predicted saliency maps to the ground truth fixation maps from the MIT1003 dataset. AUC assessed the model’s ability to distinguish between fixated and non-fixated regions by treating the predicted saliency map as a binary classifier, plotting the true positive rate against the false positive rate across thresholds, and calculating the area under the resulting ROC curve. AUC values closer to 1.0 indicate stronger performance, while 0.5 reflects chance-level prediction. NSS measured alignment between the predicted saliency values and actual fixation locations. After normalizing the predicted map to have zero mean and unit variance, values at fixation points were sampled and averaged to compute the NSS score. Higher NSS scores reflect greater agreement between the model and human gaze behaviour.

AUC and NSS were computed for all combinations of excitation and inhibition parameters and averaged across five selected images from the dataset. A heatmap of AUC values was generated to visualize parameter sensitivity and identify the settings that yielded the most biologically plausible and accurate fixation predictions.

3 Results

The testing and results of the model are split into 2 groups: beta testing and dataset testing. Beta testing refers to the makeshift testing done by group members when creating the model, and uses MS Paint features to create drawings of high contrast and brightness. The dataset testing refers to the conclusions drawn using quantitative markers like AUC and NSS regarding the model's performance with the images provided in the MIT1003 dataset.

Table 1: Figure Observations

Figure	Observation	Quantitative Markers
Figure 1	Showcases the initial spiking response to synthetic brightness stimuli. The synthetic input is bright red circles on a white background, showcasing a high contrast for the model to detect. This test revealed that further corner suppression was needed in the saliency map.	N/A, qualitative markers were used during this portion of testing
Figure 2	Improves filtering for peak clarity in the synthetic inputs and reduces the edge noise of the image, as seen in the saliency and filtered saliency map. This result helped shape early structural revisions to the spiking grid.	N/A, qualitative markers were used during this portion of testing
Figure 3	Begins the dataset testing results, and showcases how the model's saliency focused on the bright sky, instead of the expected sign/building area. The fixation misalignment highlighted the model's luminance bias.	AUC: 0.4924 | NSS: -0.12
Figure 4	Shows how the model output ignored center-biased human fixations, and instead, the bright facade drew the saliency, despite the dataset results indicating the background had minimal viewer attention.	AUC: 0.5254 | NSS: 0.25
Figure 5	Shows how the model ignored the high-contrast central objects and instead peaked in salience in the open sky.	AUC: 0.4911 | NSS: -0.13
Figure 6	The bright sunlight drew false salience despite a lack of viewer attention.	AUC: 0.4925 | NSS: -0.12
Figure 7	Model output targeted the bright sky in the top middle section of the image, rather than the human-like shapes in the bottom middle of the image.	AUC: 0.4949 | NSS: -0.09

In beta testing, neuron spike behaviour revealed a strong positional bias to corners, namely the bottom right corner (19,19), regardless of input structure. This can be seen in Table 1, specifically the saliency maps of Figures 1 and 2, where there is a strong saliency in the bottom right of the screen. To reduce this in the final neuron map, heavy corner suppression and filtering were used. It was also observed that despite the balanced stimuli provided, repetitive winner neurons emerged, suggesting that the winner-takes-all logic required refinement for saccadic diversity.

In the MIT1003 testing, the model output consistently correlated with brightness rather than semantically relevant and contrast-rich areas the dataset results normally focused on. This can be seen in Table 1, namely Figures 5, 6 and 7, where the model's output was the bright background, instead of semantically relevant objects. For this portion of testing, the AUC remained at a low value of ~0.5, with NSS values close to 0, confirming a poor prediction of human fixations.

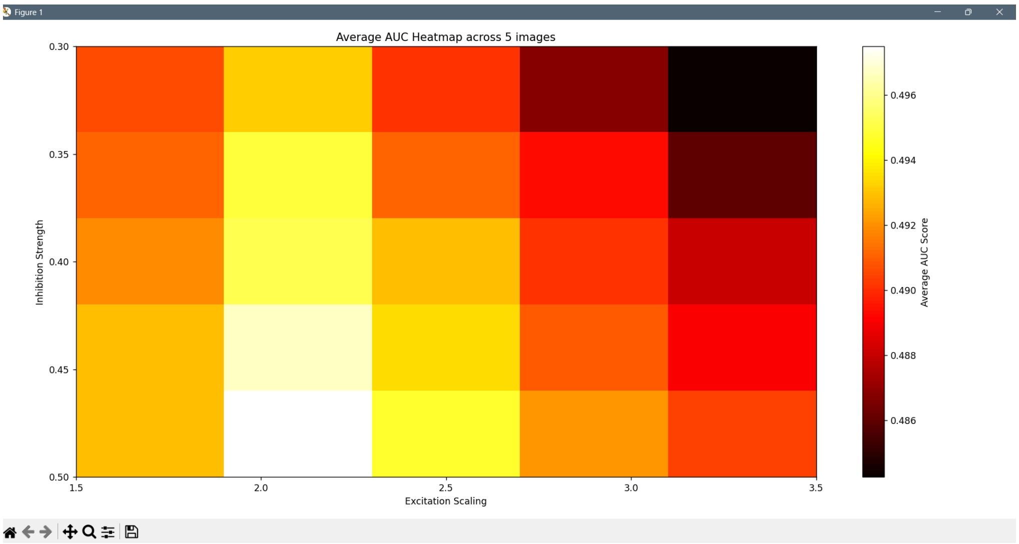

Figure 8: Parameter Sweep AUC Heatmap

Once this testing was completed, Figure 8 showcases the results of the parameter sweep across the tests and the average AUC scores that came with each E/I combination. It is found that an excitation of ~2.0 and an inhibition of ~0.45 (4.375:1 E/I ratio) is most optimal in terms of saliency accuracy with an AUC value of ~0.497. However, it is noted that comparing the optimal E/I ratio to the other combinations showcases a minute difference in accuracy, as seen by the small range of AUC values.

4 Discussion

Our parameter sweep revealed that the model performed best with moderate excitation and lower inhibition values, producing biologically grounded saccade responses. At peak performance, the excitation to inhibition ratio was approximately 4.375 to 1, closely aligning with Zhang's reported optimal ratio of 4 to 1 for Relu models [4]. This suggests that the spiking neuron model used here operates near a biologically realistic regime. Prior research, including Marino et al., has shown that winner-take-all systems achieve high accuracy when combining local excitation with broader, distal inhibition [5]. These dynamics amplify each neuron's local activity while enforcing global competition, allowing the network to isolate salient regions by enhancing contrast and suppressing background noise.

This balance also reflects the shunting dynamics described by Grossberg, who showed that neurons governed by membrane potential rules can mitigate noise and saturation when arranged in on-center, off-surround networks. These networks rely on a stable balance of excitation and inhibition to process distributed activity patterns [6]. The consistent excitation and inhibition ratios observed across test images support this principle, suggesting that the model maintains stable behaviour across diverse visual inputs.

A key limitation of the model is its reliance on simplified input. Although it incorporates lateral inhibition to support spatial competition, it does not fine-tune center-surround filters or adjust lateral connection strengths as seen in more advanced saliency models. For instance, Itti and Koch used biologically inspired techniques such as the difference of Gaussians to model retinal contrast sensitivity [1]. The additional mechanism enabled better local feature detection and likely explained the superior predictive performance observed in their work. This limitation became particularly evident when the model was tested on natural scene datasets.

Despite its biological plausibility, the model underperformed on natural scenes such as those in MIT1003. While it performed well in synthetic tests, its results on real images exposed clear weaknesses. AUC scores hovered around 0.5 and NSS values ranged from -0.1 to 0.2, indicating performance close to chance or even negatively correlated with human fixations [7]. These metrics suggest that the model often failed to identify regions that attract human attention.

Further analysis showed that the model frequently directed saccades toward bright, high contrast areas rather than semantically meaningful objects. For example, Figure 7 shows the model fixating on bright sunlight near the top center of the image, a region with no clear visual importance. A reliable saliency system should prioritize relative contrast, not absolute brightness. This model lacks local normalization, preventing it from comparing visual differences between adjacent regions. As a result, uniformly bright areas such as skies or blank walls can incorrectly dominate as salient.

To address these issues, future versions of the model should incorporate dynamic, top-down components that better reflect the temporal and semantic nature of human attention. Gaze behaviour is shaped not only by visual stimuli but also by task goals, prior fixations, memory, and contextual relevance. Mechanisms like inhibition of return (IOR) help reduce repeated fixations and encourage exploration of the image that resembles real human eye movements [8].

Top-down modulation can further enhance the model by prioritizing biologically relevant categories such as faces or skin tones, through feature weighting and excitatory input adjustments. This could be implemented via category-specific enhancements to excitatory neurons. Habituation mechanisms may also simulate cortical feedback and adaptive inhibition. These ideas are supported by findings in the superior colliculus and frontal eye fields, which integrate bottom-up and top-down inputs into a unified attentional map [9]. Prior research shows that top-down influence increases the salience of task-relevant regions, even when they are not the most visually striking [10]. Without this type of adjustment, the model defaults to the most intense stimuli, as seen in both high and low resolution testing.

Another significant limitation stems from the dataset used in the evaluation. MIT1003 includes 1003 natural images with eye tracking data from 15 participants under free viewing conditions. Its diversity in lighting, composition, and scene content makes it a valuable dataset for stimulus-driven attention. However, MIT1003 is known to exhibit a center bias, influenced by both human oculomotor behaviour and photographic framing. Models that do not account for this bias, such as the one tested here, often underperform on metrics like AUC. In this case, the lack of a center weighting mechanism is likely what caused the model to highlight peripheral brightness while ignoring central salient regions [11].

Other datasets were also considered, MIT300 was excluded due to its limited access, lack of fixation and low-level blind evaluation format. Since it is derived from MIT1003, it was not expected to offer significantly new insights. CAT2000, by contrast, offers broader coverage. It includes 4000 images across 20 diverse categories, including natural scenes, line drawings, satellite views, and abstract textures. It is especially useful for evaluating generalization. While CAT2000 was designed for models with both bottom-up and top-down mechanisms, it could still highlight specific domains where this model performs well, particularly those reliant on low-level features. Given the model’s lack of semantic processing and center bias, a range of outcomes would be expected.

Despite these challenges, CAT2000 could serve as a valuable benchmark for future testing. Categories such as low-resolution textures or grayscale patterns align with the model’s strengths in low-level contrast detection. However, categories like social or cartoon scenes would likely reveal their weaknesses, since they require semantic interpretation. This expectation is supported by studies showing that models lacking top-down processing and center bias often struggle with complex, naturalistic images [12]. Testing this model across such domains could help prioritize which improvements would be most impactful.

In summary, while the model demonstrates a biologically plausible excitation and inhibition balance and performs well in synthetic tasks, it lacks generalizability to real-world scenes. This limitation stems from its over-reliance on brightness, absence of top-down control, and failure to account for center bias. Although MIT1003 effectively revealed these issues, a dataset like CAT2000 may offer a broader assessment of where the model succeeds and fails. Moving forward, integrating temporal dynamics, semantic sensitivity, and attentional modulation will be essential to developing a more accurate and human-aligned saliency prediction system.

5 Citations

[1] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, Jan. 1998, doi: 10.1109/34.730558.

[2] A. P. Bradley, “The use of the area under the ROC curve in the evaluation of machine learning algorithms,” Pattern Recognition, vol. 30, no. 7, pp. 1145–1159, Jul. 1997, doi: 10.1016/s0031-3203(96)00142-2.

[3] F. Urban, B. Follet, C. Chamaret, O. L. Meur, and T. Baccino, “Medium spatial frequencies, a strong predictor of salience,” Cognitive Computation, vol. 3, no. 1, pp. 37–47, Nov. 2010, doi: 10.1007/s12559-010-9086-8.

[4] X. Zhang, X. Long, S.-J. Zhang, and Z. S. Chen, “Excitatory-inhibitory recurrent dynamics produce robust visual grids and stable attractors,” Cell Reports, vol. 41, no. 11, p. 111777, Dec. 2022, doi: 10.1016/j.celrep.2022.111777.

[5] R. A. Marino, T. P. Trappenberg, M. Dorris, and D. P. Munoz, “Spatial interactions in the superior colliculus predict saccade behavior in a neural field model,” Journal of Cognitive Neuroscience, vol. 24, no. 2, pp. 315–336, Sep. 2011, doi: 10.1162/jocn_a_00139.

[6] G. E. Cox, T. J. Palmeri, G. D. Logan, P. L. Smith, and J. D. Schall, “Salience by competitive and recurrent interactions: Bridging neural spiking and computation in visual attention.,” Psychological Review, vol. 129, no. 5, pp. 1144–1182, Apr. 2022, doi: 10.1037/rev0000366.

[7] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, “What do different evaluation metrics tell us about saliency models?,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 3, pp. 740–757, Mar. 2018, doi: 10.1109/tpami.2018.2815601.

[8] F. Shic and B. Scassellati, “A behavioral analysis of computational models of visual attention,” International Journal of Computer Vision, vol. 73, no. 2, pp. 159–177, Sep. 2006, doi: 10.1007/s11263-006-9784-6.

[9] R. A. Marino, T. P. Trappenberg, M. Dorris, and D. P. Munoz, “Spatial interactions in the superior colliculus predict saccade behavior in a neural field model,” Journal of Cognitive Neuroscience, vol. 24, no. 2, pp. 315–336, Sep. 2011, doi: 10.1162/jocn_a_00139.

[10] T. Judd, K. Ehinger, F. Durand, and A. Torralba, “Learning to predict where humans look,” 2009 IEEE 12th International Conference on Computer Vision, pp. 2106–2113, Sep. 2009, doi: 10.1109/iccv.2009.5459462.

[11] D. Berga, X. R. F. Vidal, X. Otazu, and X. M. Pardo, “SID4VAM: a benchmark dataset with synthetic images for visual attention modeling,” 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 8788–8797, Oct. 2019, doi: 10.1109/iccv.2019.00888.

6 Appendix

6.1 Individual Model Testing

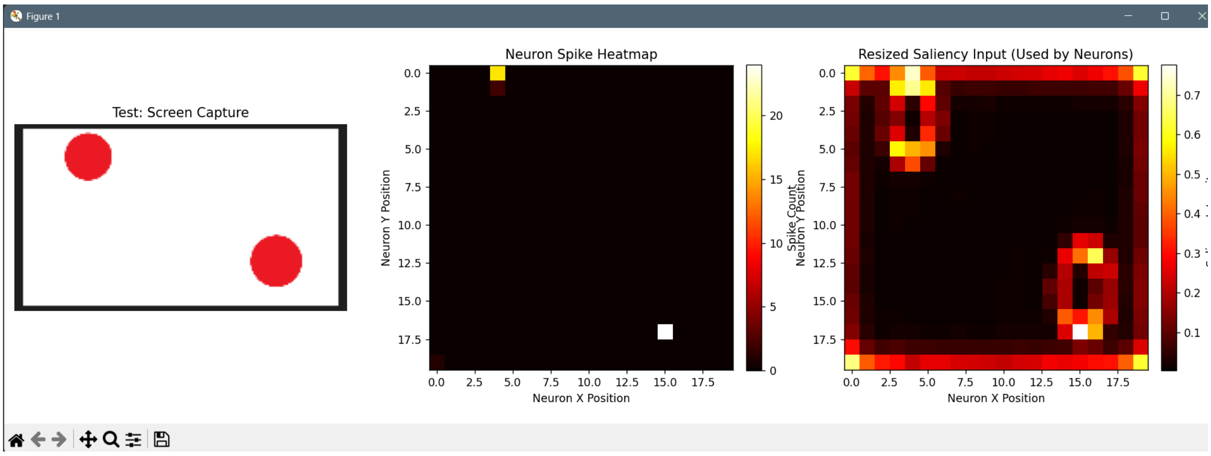

Figure 1: Initial Spiking Response to Synthetic Brightness Stimuli

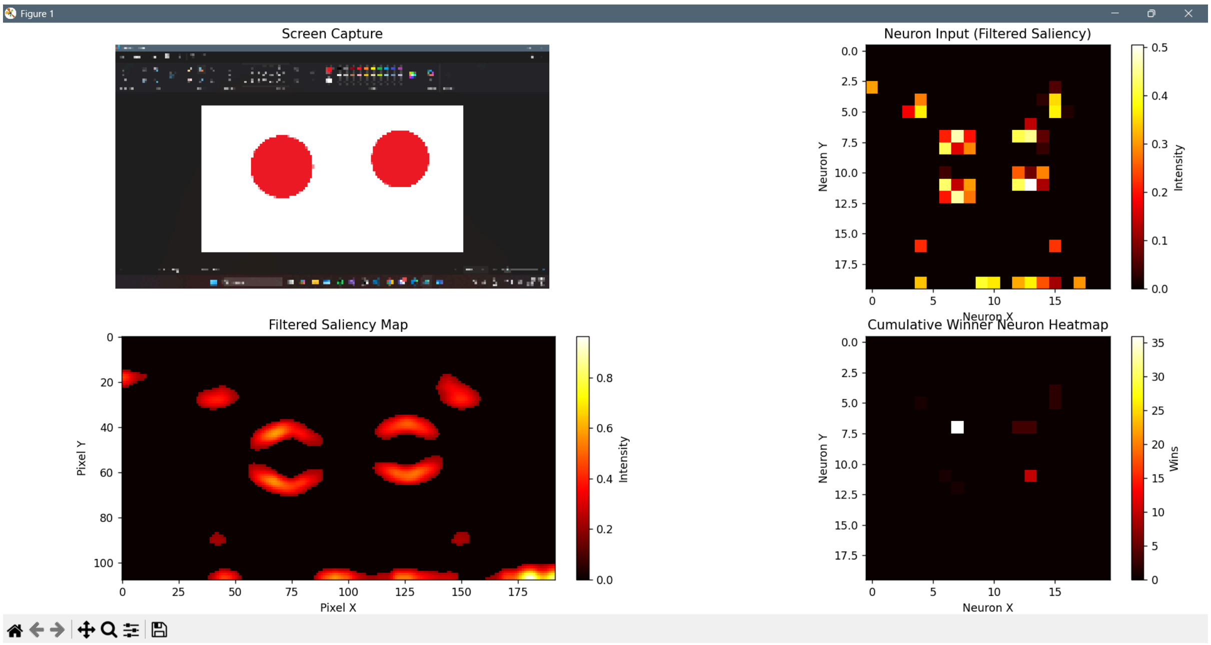

Figure 2: Filtered Saliency and Winner Neuron Selection

6.2 Dataset-Integrated Model Testing

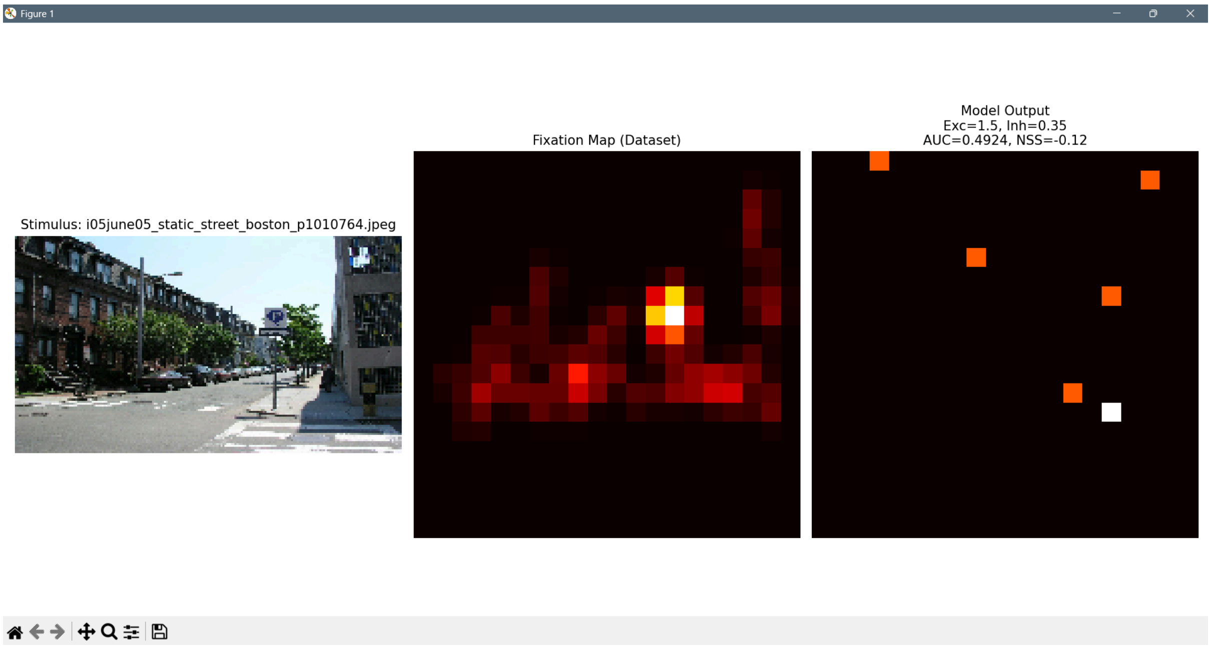

Figure 3: Testing Image 1 - Street Scene Stimulus (MIT1003)

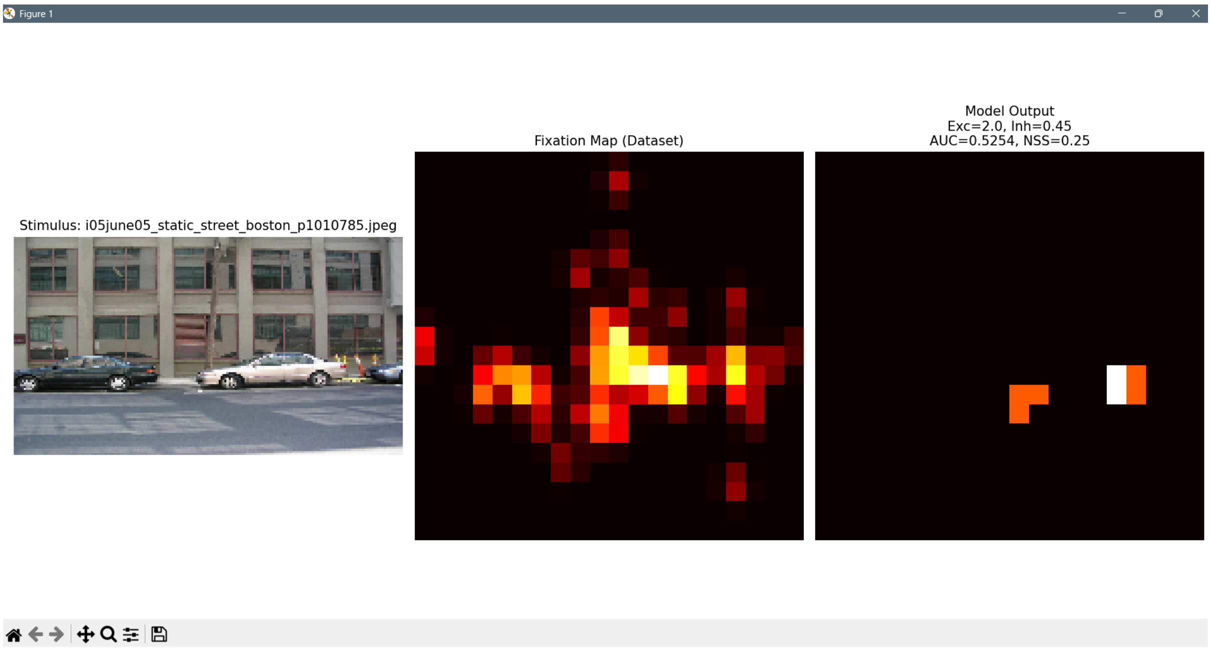

Figure 4: Testing Image 2 - Building Front Stimulus (MIT1003)

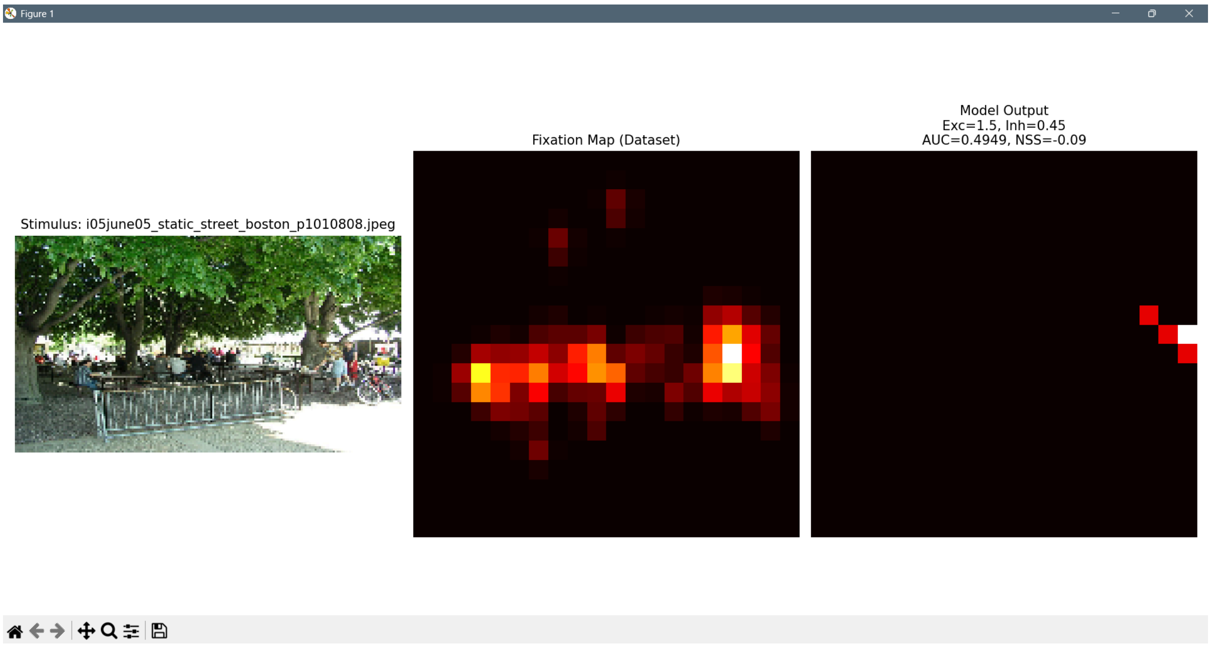

Figure 5: Testing Image 3 - Bike Rack Stimulus (MIT1003)

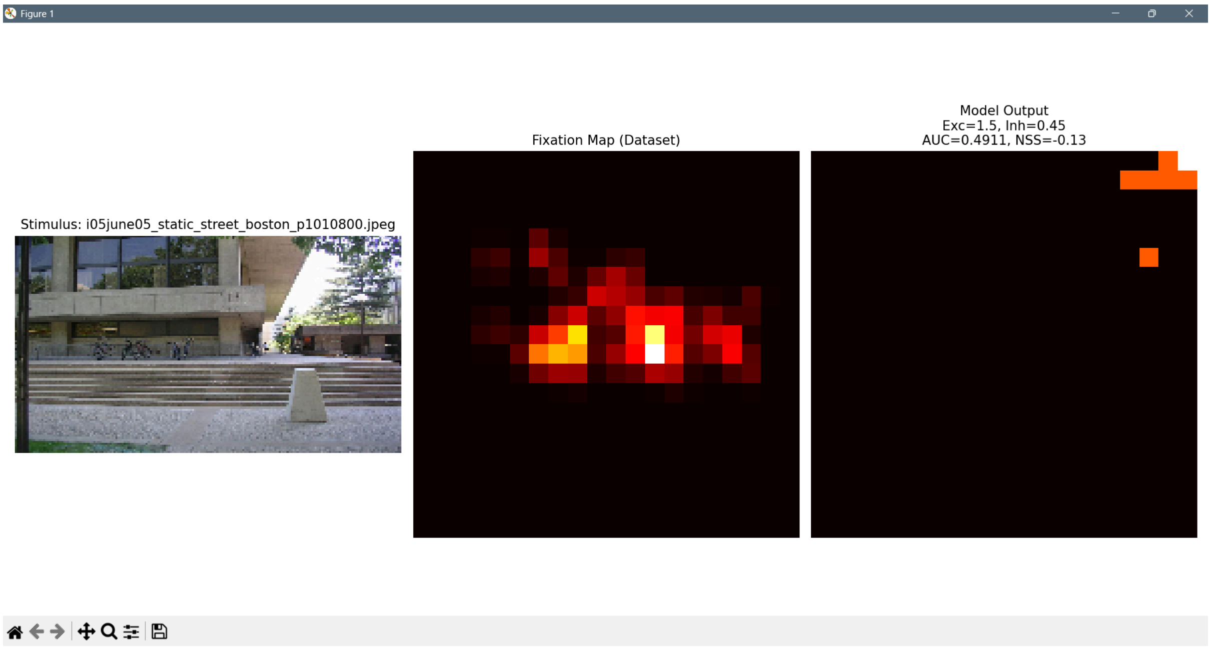

Figure 6: Testing Image 4 - Concrete Steps Stimulus (MIT1003)

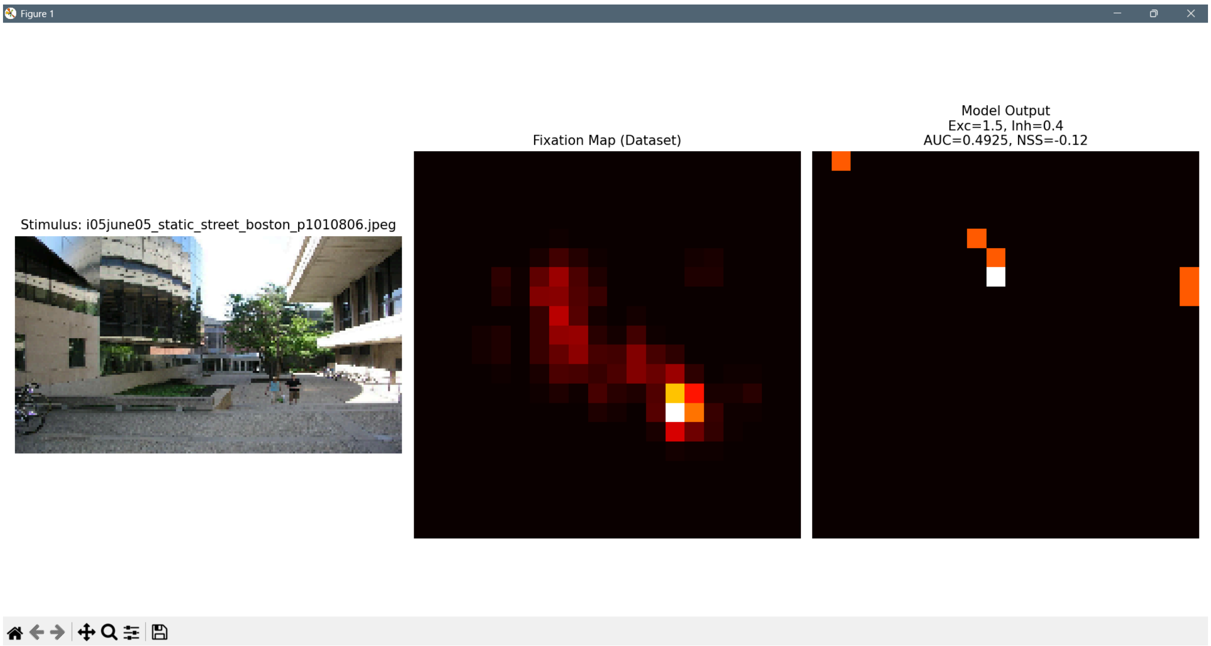

Figure 7: Testing Image 5 - Concrete Atrium Stimulus (MIT1003)