Generating Gaze Trajectories via Langevin Monte Carlo: A zero-shot goal-directed gaze simulation approach

Lorenzo Leoncini
PHuSe Lab - Università degli Studi di Milano

February 24, 2025

lorenzo.leoncini@studenti.unimi.it

Abstract. This paper explores the generation of gaze trajectories using Langevin Monte Carlo (LMC) methods applied to saliency maps. Specifically, we employ the Unadjusted Langevin Algorithm (ULA), the Metropolis Adjusted Langevin Algorithm (MALA), and a variant with a Cauchy proposal distribution (MALA-Cauchy) to simulate eye movement patterns. The approach leverages the COCO-Search-18 dataset [1, 2], which provides human eye-tracking data, and the CLIP-Seg [3] model to extract saliency maps [4] for target objects. These saliency maps are transformed into potential functions, guiding the LMC-based sampling process. The generated trajectories are processed using the IVT gaze classification algorithm to identify fixations and assign their duration. We evaluate the realism of the simulated gaze patterns by comparing them with real scanpaths from the dataset, using quantitative similarity metrics. Our results aim to improve the accuracy of gaze simulation and contribute to attention modeling research.

Source code available at:
Using-Langevin-Monte-Carlo-to-Generate-Gaze-Trajectories

1 Introduction

Gaze behavior is deeply intertwined with goals and cognitive processes, as demonstrated by Yarbus' seminal work on eye movements and visual attention [5]. When individuals observe a scene, their gaze is influenced not only by

low-level image properties but also by high-level cognitive factors such as task demands and prior knowledge. Despite this, a significant portion of research in computational visual attention has historically focused on low-level saliency as the primary driver of gaze allocation [4]. The advent of deep learning and language models has opened new possibilities for incorporating goal-directed attention into computational frameworks.

In this work, we explore the use of Langevin Monte Carlo (LMC) methods for simulating human-like gaze trajectories. Specifically, we investigate the Unadjusted Langevin Algorithm (ULA), the Metropolis Adjusted Langevin Algorithm (MALA), and a variant with a Cauchy proposal distribution (MALA-Cauchy) to generate scanpaths that resemble human eye movements. Our approach utilizes the COCO-Search-18 dataset [1, 2], which provides human eye-tracking data, and the CLIP-Seg model [3] to generate saliency maps for target objects. These saliency maps are transformed into potential functions that guide the LMC-based sampling process, simulating gaze trajectories based on the underlying visual features.

To assess the realism of the generated scanpaths, we employ the IVT gaze classification algorithm to segment the trajectories into fixations and assign fixation durations. The generated gaze trajectories are then evaluated using ScanMatch [6] with temporal information and MultiMatch [7] metrics, comparing them with real scanpaths from the dataset. Our results contribute to advancing the modeling of goal-directed visual attention and provide insights into generating realistic gaze trajectories for applications in cognitive modeling, computer vision, and human-computer interaction.

2 Related Work

The study of visual attention has evolved significantly since early computational models, such as the influential work by Itti et al. [4], which focused on low-level saliency derived from image properties. Over time, deep learning has enabled more sophisticated approaches, leading to two primary research directions: saliency-based models, which highlight visually prominent regions [8], and scanpath models, which predict full sequences of gaze fixations and saccades [9]. Although scanpath prediction has gained momentum in recent years, many works have mainly addressed the freeview condition, where no explicit task is assigned to the observer.

However, human gaze behavior is often driven by goals and tasks rather than pure image saliency. This has led to increasing interest in goal-directed attention modeling, where gaze allocation is influenced by an explicit objective,

such as object search in a scene. In this context, our work integrates several aspects drawn from the methods proposed in [10], which introduces ScanDDM, a novel scanpath model enabling generalised zero-shot goal-directed attention prediction.

3 The Proposed Approach

At every moment during visual exploration, an observer must make two key decisions: where to direct their gaze next and how long to maintain fixation at a given location. This ongoing process results in a sequence of fixations and saccades, influenced not only by the visual characteristics of the scene but also by cognitive factors such as goals and intent. In goal-directed attention, eye movements are purposefully modulated by a specific task, dynamically steering fixations toward the most relevant regions of interest.
In this work, we model this process as a stochastic exploration of a potential landscape, where gaze movements emerge from a combination of deterministic attraction to salient regions and random variability inherent in human behavior. To achieve this, we leverage Langevin Monte Carlo (LMC) sampling, a probabilistic approach that allows us to simulate scanpaths as sequences of gaze transitions governed by an underlying potential field. The observation goal is represented as a textual description, which is processed by CLIP-Seg [3] to generate a saliency map conditioned by the goal. This saliency map is then transformed into a potential function, shaping the trajectory of a Markovian random walk.
To extract fixations and their duration from the generated trajectories, we apply the IVT algorithm, which classifies gaze events based on velocity thresholds, distinguishing fixations from saccades.
Through this framework, we integrate visual information, probabilistic modeling, and stochastic dynamics, providing a principled approach to zero-shot goal-directed gaze simulation. The following sections describe how this model is constructed and evaluated.

3.1 The Model

We propose a model based on ULA, MALA, and MALA-Cauchy to simulate gaze trajectories using a potential map derived from a saliency map, which serves as a prior for visual attention. Each pixel p represents a fixation point with an associated saliency value V_p. From this, we construct a potential function where lower values guide gaze shifts toward relevant regions. Tra-

jectories are generated using stochastic differential equations (SDEs) driven by Langevin dynamics. Although ULA and MALA follow the potential gradient, MALA-Cauchy enhances exploration. The approach aims to produce realistic scanpaths, but its effectiveness remains to be tested through future experiments.

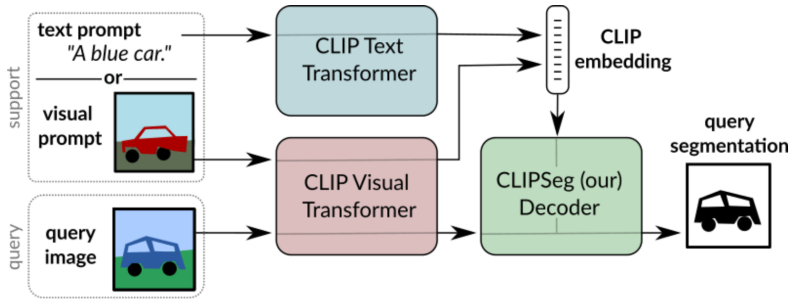

Figure 1: CLIPSeg: image segmentation with CLIP

3.1.1 Setting the Prior Value

It is assumed that the prior value assigned to each patch p of a stimulus depends on the probability that p belongs to an object of interest O_G under the current goal G of the observer: $V_p = P(p \in O_G)$ [3.1.1]. If patches are equated with the pixels that make up the image, the prior V_p can be operationally set by computing the segmentation map for the object of interest. Recent developments in computer vision provide approaches that perform zero-shot segmentation of any object based on arbitrary prompts written in natural language. For example, the ClipSeg model [3] can be used to generate such a segmentation map, which defines V_p and sets the prior probability for each pixel p to be gazed at.

3.1.2 Generating the Potential Map Using a Gaussian Mixture Model (GMM)

To obtain a smooth and continuous potential map from the saliency map [3.1.1], we model the underlying distribution using a Gaussian Mixture Model (GMM). The GMM is a probabilistic model that represents the data as a mixture of multiple Gaussian components, capturing both local and global structures in the saliency distribution.
We fit the GMM to the saliency values using the Expectation-Maximization

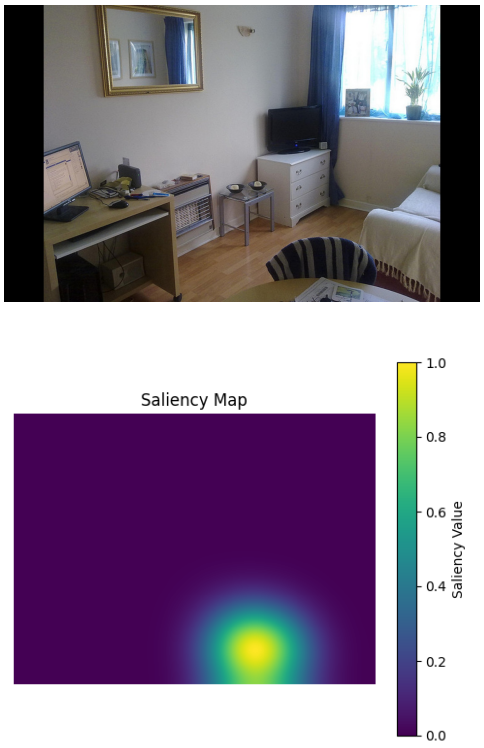

Figure 2: Generating Saliency Map with a "chair" task

(EM) algorithm, which iteratively estimates the parameters of the mixture model. The estimated probability density function (PDF) $p(x)$ provides a smooth approximation of the saliency distribution.
The potential map is then derived as follows.

$$U(x) = -\log p(x) \tag{1}$$

where $U(x)$ represents the energy landscape that will guide the sampling process. This transformation ensures that areas of high saliency correspond to low potential values, aligning with the principles of goal-directed attention.
In the end, we reshape the potential map back to the original image dimensions.
Mathematically, given a GMM with K Gaussian components parameterized by means μ_k, covariances Σ_k, and mixing coefficients π_k, the probability density function is:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \tag{2}$$

where $\mathcal{N}(x|\mu_k, \Sigma_k)$ represents a multivariate Gaussian density function. The parameters are estimated using the EM algorithm, which iteratively maximizes the likelihood function.
This approach ensures that the potential map is both continuous and well-suited for guiding the Langevin-based sampling methods, maintaining smooth gradients that facilitate efficient sampling trajectories.

3.1.3 Generating Gaze Trajectories with ULA

The Unadjusted Langevin Algorithm (ULA) is derived from the Langevin Stochastic Differential Equation (SDE):

$$d\mathbf{x}(t) = -\nabla U(\mathbf{x}(t))\, dt + \sqrt{2\Delta t}\, dW(t) \tag{3}$$

where $\mathbf{x}(t)$ represents the position, $U(\mathbf{x}(t))$ is the potential, and $W(t)$ is a Wiener process (representing noise). The discrete form of the Langevin equation is as follows:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \Delta t\, \nabla U(\mathbf{x}_t) + \sqrt{2\Delta t} \cdot \epsilon_t \tag{4}$$

where $\epsilon_t \sim \mathcal{N}(0, 1)$ is Gaussian noise and Δt is the step size.
In our method, the gaze trajectory is initialized at a given point and iteratively updated using the gradient of the potential map derived from the saliency map combined with Gaussian noise. This approach captures the dynamics of gaze patterns by integrating the influence of saliency (through the potential map) and stochasticity (through noise).

3.1.4 Generating Gaze Trajectories with MALA-Norm

The Metropolis-Adjusted Langevin Algorithm (MALA) extends ULA by incorporating a Metropolis step, ensuring that the proposed samples are accepted based on a probability criterion. It is derived from the same Langevin Stochastic Differential Equation (SDE):

$$d\mathbf{x}(t) = -\nabla U(\mathbf{x}(t))\, dt + \sqrt{2\Delta t}\, dW(t) \tag{5}$$

which leads to the discretized Langevin update:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \Delta t \, \nabla U(\mathbf{x}_t) + \sqrt{2\Delta t} \cdot \epsilon_t \tag{6}$$

where $\epsilon_t \sim \mathcal{N}(0, 1)$ represents Gaussian noise. Unlike ULA, MALA applies a Metropolis acceptance step after each Langevin update.
The proposal distribution Q determines the acceptance probability based on the log ratio:

$$\alpha = \min \left\{ 1, \frac{p(x_{t+1})Q(x_t \mid x_{t+1})}{p(x_t)Q(x_{t+1} \mid x_t)} \right\} \tag{7}$$

,where:

$$Q(x^* \mid x) \propto \exp\left(-\frac{1}{4\Delta t} \|x^* - x + \Delta t \nabla U(x)\|^2 \right) \tag{8}$$

where $\mathbf{x}$ is the current position, and $\mathbf{x}^*$ is the new position proposed.
The trajectory is generated iteratively in two main steps. First, a new position $\mathbf{x}^*$ is proposed using the Langevin update. Then, the proposal is either accepted or rejected based on the Metropolis criterion, ensuring that the trajectory follows the underlying potential landscape while maintaining stochasticity.
By balancing stochastic exploration and the potential gradient, MALA provides a principled approach to simulating gaze trajectories based on the saliency-derived potential map.

3.1.5 Generating Gaze Trajectories with MALA-Cauchy

The Metropolis-Adjusted Langevin Algorithm (MALA) can be modified by using a Cauchy proposal distribution instead of a Gaussian one. This adaptation, known as MALA-Cauchy, keeps the fundamental structure of MALA but incorporates a noise model with heavier tails, which promotes better exploration and helps alleviate challenges posed by narrow or flat gradients in the potential.
Starting with the Langevin Stochastic Differential Equation (SDE):

$$d\mathbf{x}(t) = -\nabla U(\mathbf{x}(t)) \, dt + \sqrt{2\Delta t} \, dW(t) \tag{9}$$

we derive the discretized Langevin update:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \Delta t \, \nabla U(\mathbf{x}_t) + \Delta t \cdot \epsilon_t \tag{10}$$

where ϵ_t represents noise sampled from a standard Cauchy distribution with a location parameter of 0 and a scale parameter of 1. The parameter γ controls the dispersion of the noise, influencing the step size and ensuring that the proposal distribution can account for long-range correlations in the potential. Like MALA-Norm, MALA-Cauchy applies a Metropolis acceptance step after each Langevin update.
The proposal distribution Q determines the acceptance probability based on the log ratio:

$$\alpha = \min \left\{ 1, \frac{p(x_{t+1})Q(x_t \mid x_{t+1})}{p(x_t)Q(x_{t+1} \mid x_t)} \right\} \tag{11}$$

,where

$$Q(x^* \mid x) = \frac{1}{\pi\gamma \left(1 + \frac{(x^* - x + \Delta t \nabla U(x))^2}{\gamma^2}\right)} \tag{12}$$

where $\mathbf{x}$ is the current position, and $\mathbf{x}^*$ is the proposed new position.
The gaze trajectory is generated iteratively in two key steps: first, a new position $\mathbf{x}^*$ is proposed using the modified Langevin update with Cauchy noise. Then, the proposal is accepted or rejected based on the Metropolis criterion, ensuring that the trajectory follows the potential map while maintaining the inherent stochastic nature of the process.
By leveraging the Cauchy distribution's heavy tails, MALA-Cauchy enables a wider exploration of the potential map, reducing the likelihood of getting trapped in local minima and enhancing the robustness of gaze trajectory simulations.

4 Experiments

4.1 Experimental Setup

Dataset. To test the zero-shot visual search capabilities of the proposed approach, we use the COCO-Search18 dataset. Originally designed for training deep learning models on goal-directed behavior, this dataset contains eye-tracking data from 10 participants searching for 18 target-object categories within 6,202 natural-scene images. In total, it includes approximately 300,000

recorded search fixations.
A "gold standard" was created for split 2 of the COCO-Search18 [1] validation set regarding the comparison between human scanpaths. For each photo, average values of comparisons using MM and SM with T were calculated. Once these average values were obtained, a final mean of the values was computed for the entire split. Photos with one or more observations having fewer than 3 fixations were excluded from the analysis.

Metrics. Several metrics have been developed to quantitatively evaluate the similarity between real and simulated eye movements. In this study, we employ the MultiMatch (MM) [7] and ScanMatch (SM) [6] with T metrics.
MultiMatch analyzes scanpaths by considering five key aspects: shape, direction, length, position, and duration. Scanpaths are temporally aligned and compared using the Dijkstra algorithm, and similarity scores are computed through vector arithmetic on matched saccade pairs. The final MM score is derived as the average across these measures.
ScanMatch, on the other hand, represents scanpaths as sequences of letters by segmenting them into spatial and temporal bins. The encoded scanpaths are aligned and compared, where higher scores indicate greater similarity in spatial, temporal, and sequential patterns.

Implementation Details. Our approach requires setting several hyperparameters related to the different stages of the pipeline: potential map generation, trajectory simulation, and fixation detection.
For the potential map, derived from the saliency map, we apply Gaussian smoothing with a chosen standard deviation σ and normalize the values to ensure consistency. Additionally, we determine the number of components for the Gaussian Mixture Model (GMM), which is used to transform the saliency map into a potential map. If necessary, we interpolate or resize the saliency map before conversion.
During trajectory simulation with ULA, MALA, or MALA-Cauchy, we define the step size ϵ [4][6] and the number of iterations, which influence the trajectory length and exploration behavior. The proposal distribution varies depending on the algorithm: Gaussian for MALA and ULA, and Cauchy for MALA-Cauchy.
For the MALA-Cauchy algorithm, the dispersion parameter γ for the Cauchy distribution is an important part of the proposal distribution. It influences the spread or "heaviness" of the tails of the distribution, affecting the exploration behavior of the gaze trajectory. Finally, for fixation detection using the IVT algorithm, we set a velocity threshold to distinguish fixations from saccades

and a minimum fixation duration to filter out transient points.
These parameters are tuned to generate realistic gaze trajectories and maximize similarity with human fixations, as evaluated by ScanMatch and MultiMatch.

4.2 Results

Zero-Shot Visual Search. We evaluate the proposed approach on the COCO-Search18 dataset, using the "gold standard" we previously created for split 2 of the validation set.
Results in terms of average MM and SM scores are reported in the table below. For comparison, we provide results from our approaches, as well as those obtained when comparing human-recorded scanpaths with one another (using the "gold standard").

Model	MultiMatch						ScanMatch
	Shape	Length	Direction	Position	Duration	**Avg MM**	**w/Dur**
Humans	0.926	0.916	0.762	0.885	0.698	0.837	0.457
ULA	0.906	0.9	0.623	0.708	0.602	0.750	0.113
MALA	0.905	0.896	0.633	0.711	0.6	0.749	0.12
MALA-C	0.903	0.895	0.642	0.712	0.604	0.751	0.118

Table 1: Comparison of MultiMatch and ScanMatch metrics across Humans, MALA-C, MALA, and ULA. Observations with fewer than 3 fixations were excluded, resulting in approximately 120 images with 10 observations each. Done on the val set of COCO-Search-18.

4.3 Problems

The proposed algorithms — ULA, MALA, and MALA-Cauchy — are primarily based on the gradient of the potential map to guide gaze trajectories. If the gradient is flat across large areas, these algorithms will struggle to generate meaningful trajectories, especially ULA.

Unadjusted Langevin Algorithm (ULA). ULA is computationally efficient but highly dependent on the gradient. In flat regions, it fails to guide the gaze effectively, making it unsuitable for complex environments, small objects, or multi-goal object detection. Step size and the number of steps are critical: too large a step size causes overshooting, while too small results in

slow trajectories. Excessive smoothing of the potential map can further weaken gradient details, worsening these limitations.

Metropolis-Adjusted Langevin Algorithm (MALA). MALA improves upon ULA by incorporating a Metropolis-Hastings correction, reducing bias in the random walk. However, if large portions of the potential map have a flat gradient, MALA will still struggle, as the Gaussian proposal distribution may fail to explore these areas effectively. The number of steps and step size should be carefully balanced to avoid overly long or slow trajectories. Smoothing must be done with caution to preserve important gradient information.

MALA-Cauchy. MALA-Cauchy, utilizing a Cauchy distribution, introduces more randomness, making it better suited for complex or cluttered scenes with weak gradients. However, excessive randomness may result in erratic trajectories in more structured scenes. The dispersion parameter γ is critical: too large a value leads to excessive exploration, while too small a value limits it. As with the other algorithms, careful smoothing of the potential map is required to avoid discontinuities that could hinder performance.

Parameter Considerations. The number of steps and step size are crucial for all algorithms: too few steps restrict exploration, while too many steps lead to overly long trajectories. The step size should be small enough to ensure smooth transitions without causing abrupt movements. Additionally, smoothing of the potential map should not be too aggressive, as it can remove critical gradient information and introduce discontinuities.

5 Final Considerations

A key limitation of the proposed algorithms arises when fewer than three fixations are generated or when the saliency area is misidentified. In such cases, performance metrics are significantly reduced, highlighting the dependence on accurate saliency detection. Due to the variability in saliency areas across different images, generalizing these algorithms for diverse image types remains challenging.

This study evaluated ULA, MALA, and MALA-Cauchy for simulating gaze trajectories. While ULA is efficient, its reliance on gradients limits its effectiveness for complex or multi-goal tasks. MALA mitigates some of these issues but

still struggles with flat-gradient areas. MALA-Cauchy improves exploration with greater randomness but may result in erratic trajectories in structured scenes.
A critical challenge across all algorithms is the balance between step size and number of steps, as well as proper smoothing of the potential map, which directly impacts performance. Accurate saliency detection remains crucial for reliable gaze trajectory simulation.

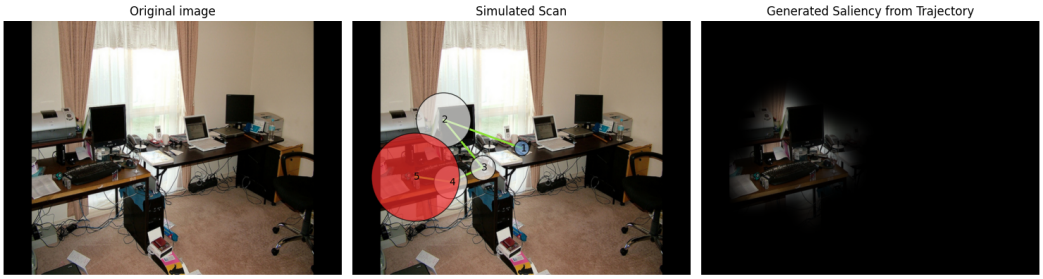

Figure 3: Results of a *"keyboard" goal-directed task.*

For example, in a "keyboard" goal-directed task, MALA could produce a gaze trajectory that simulates human-like focus on key areas, but if the saliency map is not precise, the gaze trajectory may not match the desired focus, leading to less accurate task completion.
Future improvements should focus on adaptive strategies for different images and tasks to enhance gaze trajectory reliability in various environments.

References

[1] Y. Chen, Z. Yang, S. Ahn, D. Samaras, M. Hoai, and G. Zelinsky, "Coco-search18 fixation dataset for predicting goal-directed attention control," *Scientific Reports*, vol. 11, no. 1, pp. 1–11, 2021.

[2] Z. Yang, L. Huang, Y. Chen, Z. Wei, S. Ahn, G. Zelinsky, D. Samaras, and M. Hoai, "Predicting goal-directed human attention using inverse reinforcement learning," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[3] T. Lüddecke and A. Ecker, "Image segmentation using text and image prompts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7086–7096, June 2022.

[4] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, pp. 1254 – 1259, 12 1998.

[5] A. L. Yarbus, *Eye movements and vision*. Springer, 2013.

[6] F. Cristino, S. Mathôt, J. Theeuwes, and I. Gilchrist, "Scanmatch: A novel method for comparing fixation sequences," *Behavior research methods*, vol. 42, pp. 692–700, 08 2010.

[7] R. Dewhurst, M. Nyström, H. Jarodzka, T. Foulsham, R. Johansson, and K. Holmqvist, "It depends on how you look at it: Scanpath comparison in multiple dimensions with multimatch, a vector-based approach," *Behavior research methods*, vol. 44, 05 2012.

[8] A. Borji, "Saliency prediction in the deep learning era: Successes and limitations," *IEEE transactions on pattern analysis and machine intelligence*, vol. PP, 08 2019.

[9] M. Kümmerer and M. Bethge, "State-of-the-art in human scanpath prediction," 02 2021.

[10] A. D'Amelio, M. Lucchi, and G. Boccignone, "ScanDDM: Generalised Zero-Shot Neuro-Dynamical Modelling of Goal-Directed Attention," in *Proceedings of the European Conference on Computer Vision*, 2024.