An Active Inference Model of Covert and Overt Visual Attention

Tin Mišić, Karlo Koledić, Fabio Bonsignorio, Ivan Petrović, and Ivan Marković

University of Zagreb Faculty of Electrical Engineering and Computing, Croatia;
Correspondence: ivan.markovic@fer.hr

Abstract. The ability to selectively attend to relevant stimuli while filtering out distractions is essential for agents that process complex, high-dimensional sensory input. This paper introduces a model of covert and overt visual attention through the framework of active inference, utilizing dynamic optimization of sensory precisions to minimize free-energy. This work addresses the lack of active inference models that integrate visual attention with continuous sensory representations and deep generative models for robotics. Our proposed model determines visual sensory precisions based on both current environmental beliefs and sensory input, influencing attentional allocation in both covert and overt modalities. To test the effectiveness of the model, we analyze its behavior in the Posner cueing task and a simple target focus task using two-dimensional (2D) visual data. Reaction times are measured to investigate the interplay between exogenous and endogenous attention, as well as valid and invalid cueing. The results show that exogenous and valid cues generally lead to faster reaction times compared to endogenous and invalid cues. Finally, we show that reflexive saccades are faster than intentional ones, though less adaptable, and discuss the implications for robotic applications.

Keywords: Active inference· Visual attention· Posner cueing task.

1 Introduction

Attention as a cognitive process allows agents to selectively focus on specific stimuli while ignoring others. This ability helps humans avoid sensory overload, and as robots acquire more complex sensory channels it could help decrease the computational load required to perform in daily tasks, such as object tracking and visual search, as well as social interactions [20–22]. Attention is often separated into top-down, or goal-driven attention, and bottom-up or stimulus-driven attention, with some theories including hysteresis as a third component [38]. Top-down attention bilaterally activates dorsal posterior parietal and frontal regions of the brain, while bottom-up attention activates the right-lateralized ventral system, with the dorsal frontoparietal system combining the two into a "salience map" during visual search [6, 23]. Furthermore, visual attention is separated into overt and covert attention [1, 18], with overt attention involving saccadic eye movements to the attentional target, and covert attention referring to attention shifts to the target while the eyes remain fixated elsewhere.

Visual attention and its models are most often tested using the Posner cueing task, i.e., the Posner paradigm. The Posner cueing task is an experimental paradigm used to study covert visual attention [33, 34]. Participants are asked to fixate on a central point while a cue directs attention to a location where a target may appear. The cue can either be endogenous – meaning that attention is voluntarily guided based on symbolic cues (e.g., an arrow pointing left or right), or exogenous – meaning that attention is automatically drawn by a sudden, peripheral stimulus (e.g. a bright flash or a flickering box). Endogenous cueing is considered to be top-down because it requires cognitive processing and active interpretation of the cue, while exogenous cueing is considered to be bottom-up because it does not require conscious interpretation. Reaction times are measured to assess how cues influence attentional shifts.

Through the original Posner paradigm [33, 34] and its variations, valuable insights have been gained about attentional processes. Covert attentional shifts to a target area occur prior to any eye movement [31, 33], and valid cues produce faster responses than invalid cues [33, 34]. Exogenous cues were shown to produce faster reaction times than endogenous cues [4,13], showing that bottom-up attention is faster because it requires no conscious processing. The question of weather attentional selection is object-based or location-based has also been thoroughly researched, and the consensus is that both types are not mutually exclusive, but are dependent on the current task [7,37,43]. Research supporting location-based attention has shown that target eccentricity, i.e. the target distance from the central focus point, plays a role in reaction time, with reaction times increasing as target eccentricity increased [2,17,32].

Multiple approaches exist to model attention, many of which are based on Bayesian inference [9, 12, 24, 26–28, 30, 36, 42]. Previous studies have modeled visual attention and active saccades in visual search tasks [9, 24, 26, 27]. However, the integration of visual attention and bottom-up action within the active inference framework—operating directly on raw two-dimensional visual input—remains largely unexplored. This gap is especially significant in robotics, where vision is a core sensory modality and visual data provide the primary basis for decision-making and interaction with the environment. A key limitation of many existing models [24, 26, 27, 36] is their reliance on discrete sensory inputs or internal states. While suitable for abstract tasks, such discretization is problematic for robotics, where sensory and motor variables are inherently continuous. Treating continuous signals as discrete reduces precision and limits applicability in real-world scenarios. By contrast, continuous representations naturally reflect the analog nature of sensorimotor data and enable more accurate perception, motion control, and sensorimotor learning [5, 8, 41]. In addition, several of these models either omit deep generative models altogether [9, 24, 26, 27] or, in cases where 2D data are used [36], do not leverage deep architectures. Conversely, active inference implementations that employ deep generative models, such as [3, 35], do not address visual attention. Overall, prior work has not yet integrated visual attention, continuous sensory representations, and deep generative models into a single active inference framework suitable for robotics.

In this paper we propose a model of visual attention, shown in Fig. 1, viewed through the lens of active inference [29] – a computational approach derived from the free-energy principle (FEP). According to the FEP, systems adapt and act in a way that minimizes their free-energy [11]. Free-energy is a concept borrowed from physics, statistics, and information theory that limits the surprise on a sample of data given a generative model. This principle helps to explain how biological systems resist the natural tendency to disorder, and their action, perception, and learning processes [10]. In the FEP, attention is theoretically achieved by optimizing sensory precisions, their parameters, and mutual precision weighing [9,14,24,26–28,30]. Biased competition and endogenous/exogenous attention have been studied in this context, and the precision optimization produces behaviors similar to human attention [9, 42].

This study introduces a hierarchical active inference model of overt and covert visual attention, addressing precision optimization for visual data as a mechanism for endogenous and exogenous attention as well as action control. The model integrates both top-down and bottom-up processes, enabling covert and overt shifts of attention. Its performance is demonstrated through the Posner cueing task and a simple target-focus task on two-dimensional visual input. The primary goal is to develop a model that accurately captures human attentional mechanisms while establishing a foundation for robotic applications, including active visual search and joint attention in human–robot interaction. To address previously mentioned limitations, we propose a hierarchical active inference model that incorporates visual attention mechanisms within a deep generative framework, using continuous internal states and raw two-dimensional visual data as input. A variational auto-encoder (VAE) serves as the visual generative model, while model training and experiments are conducted in the Gazebo simulator under the Robot Operating System (ROS).

The paper is organized as follows. In Sec. 2 we give an overview of the theoretical background and elaborate the proposed approach that is based on free-energy minimization with 2D precision optimization and overt saccades through active inference. Section 3 shows the results of the Posner cueing tasks and active attention trials. Section 4 provides the discussion of the results while Sec. 5 concludes the paper and provides directions for future work.

2 Proposed Method

2.1 Free-energy Minimization

Free-energy is defined as the negative evidence lower bound (ELBO), or as the sum of the Kullback-Leibler (KL) divergence and the surprise [9–11]:

$$F(\boldsymbol{z}, \boldsymbol{s}) = -\mathcal{L}(q) = D_{KL}[q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{s})] - \ln p(\boldsymbol{s}), \tag{1}$$

where $\boldsymbol{z}$ and $\boldsymbol{s}$ represent latent system states and sensory observations, respectively, while the KL-divergence is computed between the posterior $p(\boldsymbol{z}|\boldsymbol{s})$ and the approximate variational density $q(\boldsymbol{z})$. Given that, the surprise is defined as the negative log-probability of an outcome $-\ln p(\boldsymbol{s})$. If the variational density $q(\boldsymbol{z})$ is assumed to factor into Gaussian probability density functions (pdfs) [9,11,35]:

$$q(\boldsymbol{z}) = \prod_i q(\boldsymbol{z}_i) = \prod_i \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Pi}_i^{-1}), \tag{2}$$

the free-energy then becomes dependent only on the most probable hypotheses, beliefs $\boldsymbol{\mu_i}$, and precision matrices $\boldsymbol{\Pi}_i$ of the latent system states $\boldsymbol{z}$ [9,35]:

$$\begin{aligned} F(\boldsymbol{\mu}, \boldsymbol{\Pi}, \boldsymbol{s}) &= -\ln p(\boldsymbol{s}, \boldsymbol{\mu}, \boldsymbol{\Pi}) + C \\ &= -\ln p(\boldsymbol{s}|\boldsymbol{\mu}, \boldsymbol{\Pi}) - \ln p(\boldsymbol{\mu}, \boldsymbol{\Pi}) + C \end{aligned} \tag{3}$$

Furthermore, sensory observations $\boldsymbol{s}$ and beliefs $\boldsymbol{\mu}$ are defined in the context of hierarchical dynamic models [9–11,35]:

$$\begin{aligned} \tilde{\boldsymbol{s}} &= \tilde{\boldsymbol{g}}(\tilde{\boldsymbol{\mu}}) + \boldsymbol{w}_s \\ D\tilde{\boldsymbol{\mu}} &= \tilde{\boldsymbol{f}}(\tilde{\boldsymbol{\mu}}) + \boldsymbol{w}_\mu. \end{aligned} \tag{4}$$

Here, $\tilde{\boldsymbol{\mu}}$ indicates generalized coordinates of beliefs with multiple temporal orders, $\tilde{\boldsymbol{\mu}} = \{\boldsymbol{\mu}, \boldsymbol{\mu}', \boldsymbol{\mu}'', \cdots\}$, which allow for a richer approximation of the environment dynamics, D stands for the differential shift operator $D\tilde{\boldsymbol{\mu}} = \{\boldsymbol{\mu}', \boldsymbol{\mu}'', \cdots\}$ in the generalized equation of system dynamics $\tilde{\boldsymbol{f}}(\tilde{\boldsymbol{\mu}})$, while $\tilde{\boldsymbol{g}}(\tilde{\boldsymbol{\mu}})$ is the sensor model that maps current beliefs to sensory observations. The amplitudes of random fluctuations $\boldsymbol{w}_s$ and $\boldsymbol{w}_\mu$ are state dependent and are defined as Gaussian pdfs with covariances $\boldsymbol{\Sigma_s}$ and $\boldsymbol{\Sigma_\mu}$, respectively [9,35]:

$$\begin{aligned} \boldsymbol{w_s} &\sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_s(\boldsymbol{z}, \boldsymbol{s}, \boldsymbol{\gamma})) \\ \boldsymbol{w_\mu} &\sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_\mu(\boldsymbol{z}, \boldsymbol{s}, \boldsymbol{\gamma})). \end{aligned} \tag{5}$$

The covariances $\boldsymbol{\Sigma}_i$ are the inverses of precisions, $\boldsymbol{\Sigma}_i := \boldsymbol{\Sigma}_i(\boldsymbol{z}, \boldsymbol{s}, \boldsymbol{\gamma}) = \boldsymbol{\Pi}_i(\boldsymbol{z}, \boldsymbol{s}, \boldsymbol{\gamma})^{-1}$, with precision parameters $\boldsymbol{\gamma}$ that control the amplitudes [9,42]. The precisions are dynamic and depend on the current states and sensory input. It is through optimization of precisions and their parameters that attention is achieved [9,14, 24,26–28,30].

2.2 Perceptual and Active Inference

Perception, action, and learning can all be optimized through the minimization of free-energy. In this paper we only consider perception and action, and leave the learning processes of attention for future work. Action and beliefs are optimized through gradient descent [10,11,29,35]:

$$\begin{aligned} &\dot{\tilde{\boldsymbol{\mu}}} - D\tilde{\boldsymbol{\mu}} = -\partial_{\tilde{\mu}} F(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Pi}}, \tilde{\boldsymbol{s}}) \\ &\dot{\boldsymbol{a}} = -\partial_{\boldsymbol{a}} F(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Pi}}, \tilde{\boldsymbol{s}}). \end{aligned} \tag{6}$$

The likelihood and prior in (3) also become generalized and can be partitioned within and across temporal orders d, respectively [35]:

$$\begin{aligned} p(\tilde{\boldsymbol{s}}|\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Pi}}_{\boldsymbol{s}}) &= \prod_{d} p(\boldsymbol{s}^{[d]}|\boldsymbol{\mu}^{[d]}, {\boldsymbol{\Pi}_{\boldsymbol{s}}}^{[d]}) \\ p(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Pi}}_{\boldsymbol{\mu}}) &= \prod_{d} p(\boldsymbol{\mu}^{[d+1]}|\boldsymbol{\mu}^{[d]}, {\boldsymbol{\Pi}_{\boldsymbol{\mu}}}^{[d]}). \end{aligned} \tag{7}$$

These partitions are also assumed to take the following Gaussian pdf form:

$$\begin{aligned} p(\boldsymbol{s}^{[d]}|\boldsymbol{\mu}^{[d]}, {\boldsymbol{\Pi}_{\boldsymbol{s}}}^{[d]}) &= \frac{|{\boldsymbol{\Pi}_{\boldsymbol{s}}}^{[d]}|^{\frac{1}{2}}}{\sqrt{(2\boldsymbol{\pi})^{L}}} \exp\left(-\frac{1}{2}{\boldsymbol{e}_{s}^{[d]}}^{T} {\boldsymbol{\Pi}_{\boldsymbol{s}}}^{[d]} \boldsymbol{e}_{s}^{[d]}\right) \\ p(\boldsymbol{\mu}^{[d+1]}|\boldsymbol{\mu}^{[d]}, {\boldsymbol{\Pi}_{\boldsymbol{\mu}}}^{[d]}) &= \frac{|{\boldsymbol{\Pi}_{\boldsymbol{\mu}}}^{[d]}|^{\frac{1}{2}}}{\sqrt{(2\boldsymbol{\pi})^{M}}} \exp\left(-\frac{1}{2}{\boldsymbol{e}_{\mu}^{[d]}}^{T} {\boldsymbol{\Pi}_{\boldsymbol{\mu}}}^{[d]} \boldsymbol{e}_{\mu}^{[d]}\right), \end{aligned} \tag{8}$$

where L and M are the respective dimensions of sensory observations $\boldsymbol{s}$ and internal beliefs $\boldsymbol{\mu}$. Therein, $\boldsymbol{e}_{s}^{[d]}$ and $\boldsymbol{e}_{\mu}^{[d]}$ represents sensory and system dynamics prediction errors:

$$\begin{aligned} \boldsymbol{e}_{s}^{[d]} &= \boldsymbol{s}^{[d]} - \boldsymbol{g}^{[d]}(\boldsymbol{\mu}^{[d]}) = \boldsymbol{s}^{[d]} - \boldsymbol{p}^{[d]} \\ \boldsymbol{e}_{\mu}^{[d]} &= \boldsymbol{\mu}^{[d+1]} - \boldsymbol{f}^{[d]}(\boldsymbol{\mu}^{[d]}), \end{aligned} \tag{9}$$

where $\boldsymbol{p}^{[d]} = \boldsymbol{g}^{[d]}(\boldsymbol{\mu}^{[d]})$ are sensory predictions generated by the generative sensor model. Note that in our case the system dynamics model is defined through flexible intentions $\boldsymbol{h}^{(k)}$ [35], where for each intention $k \in (0, K-1)$:

$$\boldsymbol{f}^{(k)}(\boldsymbol{\mu}) = l \cdot \boldsymbol{E}_{i}^{(k)} + \boldsymbol{w}_{\mu}^{(k)} = l \cdot (\boldsymbol{h}^{(k)} - \boldsymbol{\mu}) + \boldsymbol{w}_{\mu}^{(k)}, \tag{10}$$

with l being the empirically-determined gain of intention errors $\boldsymbol{E}_{i}^{(k)}$. This gain could potentially be optimized through learning with the FEP. Flexible intentions $\boldsymbol{h}^{(k)}$ represent top-down attractors which are generated from current beliefs and drive beliefs towards dynamic goals.The implementation of the generative sensor models $\boldsymbol{g}^{[d]}$ is presented in subsection 3.1.

Belief update With state- and sensory-dependent precisions, the belief update takes the following form:

$$\begin{aligned} \dot{\tilde{\boldsymbol{\mu}}} =& D\tilde{\boldsymbol{\mu}} + \frac{\partial\tilde{\boldsymbol{g}}}{\partial\tilde{\boldsymbol{\mu}}}^{T} \tilde{\boldsymbol{\Pi}}_{s}\tilde{\boldsymbol{e}}_{s} + \frac{\partial\tilde{\boldsymbol{f}}}{\partial\tilde{\boldsymbol{\mu}}}^{T} \tilde{\boldsymbol{\Pi}}_{\mu}\tilde{\boldsymbol{e}}_{\mu} - D^{T}\tilde{\boldsymbol{\Pi}}_{\mu}\tilde{\boldsymbol{e}}_{\mu} \\ &+ \frac{1}{2}\mathrm{Tr}\left[\tilde{\boldsymbol{\Pi}}_{\boldsymbol{s}}^{-\mathbf{1}}\frac{\partial\tilde{\boldsymbol{\Pi}}_{\boldsymbol{s}}}{\partial\tilde{\boldsymbol{\mu}}}\right] - \frac{1}{2}\tilde{\boldsymbol{e}}_{s}^{T}\frac{\partial\tilde{\boldsymbol{\Pi}}_{\boldsymbol{s}}}{\partial\tilde{\boldsymbol{\mu}}}\tilde{\boldsymbol{e}}_{s} \\ &+ \frac{1}{2}\mathrm{Tr}\left[\tilde{\boldsymbol{\Pi}}_{\boldsymbol{\mu}}^{-\mathbf{1}}\frac{\partial\tilde{\boldsymbol{\Pi}}_{\boldsymbol{\mu}}}{\partial\tilde{\boldsymbol{\mu}}}\right] - \frac{1}{2}\tilde{\boldsymbol{e}}_{\mu}^{T}\frac{\partial\tilde{\boldsymbol{\Pi}}_{\boldsymbol{\mu}}}{\partial\tilde{\boldsymbol{\mu}}}\tilde{\boldsymbol{e}}_{\mu}, \end{aligned} \tag{11}$$

with Tr being the trace of a matrix. The terms that comprise the belief update equation are:

- $\frac{\partial \tilde{\boldsymbol{g}}}{\partial \tilde{\boldsymbol{\mu}}}^T \tilde{\boldsymbol{\Pi}}_s \tilde{\boldsymbol{e}}_s$: likelihood error computed at the sensory level, representing the free-energy gradient of the likelihood relative to the belief $\tilde{\boldsymbol{\mu}}^{[d]}$ in (9)
- $\frac{\partial \tilde{\boldsymbol{f}}}{\partial \tilde{\boldsymbol{\mu}}}^T \tilde{\boldsymbol{\Pi}}_\mu \tilde{\boldsymbol{e}}_\mu$: backward error from the next temporal order, representing the free-energy gradient relative to the belief $\tilde{\boldsymbol{\mu}}^{[d+1]}$ in (9)
- $-D^T \tilde{\boldsymbol{\Pi}}_\mu \tilde{\boldsymbol{e}}_\mu$: forward error coming from the previous temporal order, representing the free-energy gradient relative to the belief $\tilde{\boldsymbol{\mu}}^{[d]}$ in (9)
- $\frac{1}{2}\mathrm{Tr}\left[\tilde{\boldsymbol{\Pi}}_{\boldsymbol{s}}^{\mathbf{-1}} \frac{\partial \tilde{\boldsymbol{\Pi}}_{\boldsymbol{s}}}{\partial \tilde{\boldsymbol{\mu}}}\right] - \frac{1}{2}\tilde{\boldsymbol{e}}_s^T \frac{\partial \tilde{\boldsymbol{\Pi}}_{\boldsymbol{s}}}{\partial \tilde{\boldsymbol{\mu}}} \tilde{\boldsymbol{e}}_s$: free-energy gradients from the sensory precisions, serves as bottom-up attention
- $\frac{1}{2}\mathrm{Tr}\left[\tilde{\boldsymbol{\Pi}}_{\boldsymbol{\mu}}^{\mathbf{-1}} \frac{\partial \tilde{\boldsymbol{\Pi}}_{\mu}}{\partial \tilde{\boldsymbol{\mu}}}\right] - \frac{1}{2}\tilde{\boldsymbol{e}}_\mu^T \frac{\partial \tilde{\boldsymbol{\Pi}}_{\mu}}{\partial \tilde{\boldsymbol{\mu}}} \tilde{\boldsymbol{e}}_\mu$: free-energy gradients from the system dynamics precisions, serves as top-down attention.

Action update Action is also updated through the minimization of free-energy [10, 11, 29, 35]:

$$a = \arg\min_a F(\boldsymbol{\mu}, \boldsymbol{\Pi}, \boldsymbol{s}), \tag{12}$$

with the action update taking the following form:

$$\dot{a} = -\partial_a F(\boldsymbol{\mu}, \boldsymbol{\Pi}, \boldsymbol{s}) = -\frac{\partial \tilde{\boldsymbol{s}}}{\partial a}^T \tilde{\boldsymbol{\Pi}}_s \tilde{\boldsymbol{e}}_s + \frac{1}{2}\mathrm{Tr}\left[\tilde{\boldsymbol{\Pi}}_{\boldsymbol{s}}^{\mathbf{-1}} \frac{\partial \tilde{\boldsymbol{\Pi}}_{\boldsymbol{s}}}{\partial \tilde{\boldsymbol{s}}}\right] \frac{\partial \tilde{\boldsymbol{s}}}{\partial a} - \frac{1}{2}\tilde{\boldsymbol{e}}_s^T \frac{\partial \tilde{\boldsymbol{\Pi}}_{\boldsymbol{s}}}{\partial \tilde{\boldsymbol{s}}} \tilde{\boldsymbol{e}}_s \frac{\partial \tilde{\boldsymbol{s}}}{\partial a}, \tag{13}$$

with bottom-up attention components in relation to sensory input, analogous to those in relation to belief in (11). These control signals act as reflexive saccades [16, 45]. The gradient $\frac{\partial \tilde{\boldsymbol{s}}}{\partial a}$ is an inverse mapping from sensory data to actions.

3 Results

3.1 Implementation of the proposed model

A graphical overview of the model[1] is shown in Fig. 1. The current belief state $\boldsymbol{\mu}$ is passed to exteroceptive, proprioceptive, and interoceptive generative models. Their predictions $\boldsymbol{p}$ are compared to actual inputs $\boldsymbol{s}$, with prediction errors $\boldsymbol{e_s}$ driving both action and belief updates. Proprioceptive (camera pitch/yaw) and interoceptive (symbolic cue) generative models are identity mappings, while the exteroceptive visual model is the decoder of a disentangled VAE. The VAE encodes target presence and position, simplifying conversion from intrinsic image coordinates to extrinsic camera orientation.

The belief state consists of:

- **Symbolic cue belief** – position of an endogenous cue on the image, mirroring sensory input
- **Camera orientation belief** – proprioceptive pitch and yaw of the camera

[1] Code, video examples, and details on VAE training and simulations are available at: https://unizgfer-lamor.github.io/ainf-visual-attention/

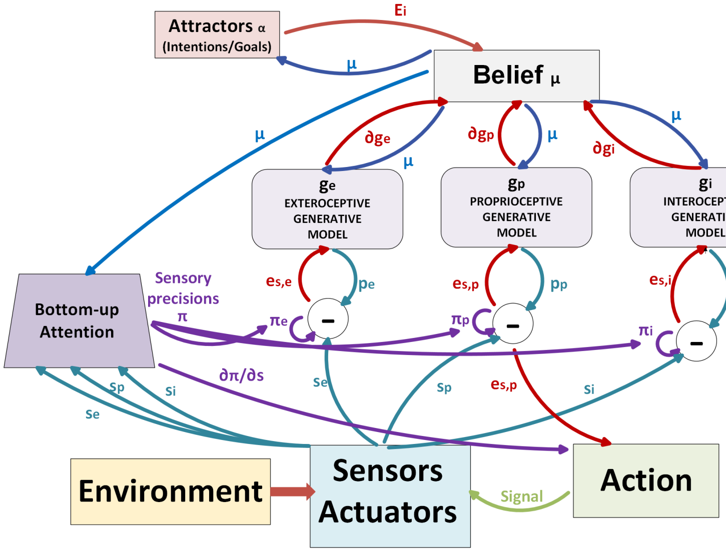

Fig. 1: Core structure of the proposed model. Beliefs about sensory causes are updated through attractor goals and error signals to minimize free-energy. Bottom-up attention is regulated by dynamic sensory precisions.

- **Visual belief** – disentangled encoding of target presence and position
- **Covert attention belief** – amplitude and center of an RBF governing visual precisions

Beliefs are updated through bottom-up prediction errors and top-down attractors $\boldsymbol{\alpha}$, following the flexible intentions theory in [35]. In our model, attractors for proprioceptive, visual, and covert attention beliefs are generated from current visual and cueing inputs, allowing dynamic goals. These intentions drive overt actions (camera orientation) and covert shifts (RBF updates).

Visual precision $\boldsymbol{\Pi_s}$ is defined as a diagonal matrix:

$$\boldsymbol{\Pi_s} = \begin{bmatrix} \pi_1(\boldsymbol{\mu}, \boldsymbol{s}) & 0 & \cdots & 0 \\ 0 & \pi_2(\boldsymbol{\mu}, \boldsymbol{s}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pi_L(\boldsymbol{\mu}, \boldsymbol{s}) \end{bmatrix}_{L \times L}, \tag{14}$$

where $L = 32 \times 32\,(\times 3)$ is the dimensionality of the visual data. Each precision term is determined by an RBF centered on the covert attention belief

$[\mu_{amp}, \mu_u, \mu_v]$ and the centroid of the largest red object $[\boldsymbol{r_u}(\boldsymbol{s}), \boldsymbol{r_v}(\boldsymbol{s})]$:

$$\begin{aligned}
&\pi_i(\boldsymbol{\mu}, \boldsymbol{s}) = \pi(x, y, \boldsymbol{\mu}, \boldsymbol{s}) = \\
&\frac{\mu_{amp}}{2}\left(\ln\left(-\tfrac{(x-\mu_u)^2+(y-\mu_v)^2}{b^2}+1\right)+c\right) \\
&+\frac{1}{2}\left(\ln\left(-\tfrac{(x-\boldsymbol{r_u}(\boldsymbol{s}))^2+(y-\boldsymbol{r_v}(\boldsymbol{s}))^2}{b^2}+1\right)+c\right),
\end{aligned} \tag{15}$$

with parameters $b = 2.6$ and $c = 1$, chosen to normalize RBF values between 0 and 1. This RBF form ensures that covert attention is drawn toward areas of high prediction error, unlike Gaussian RBFs which repel from error. As shown in Fig. 2, the resulting precision decreases with distance from the focus center, mimicking human foveation [2, 32].

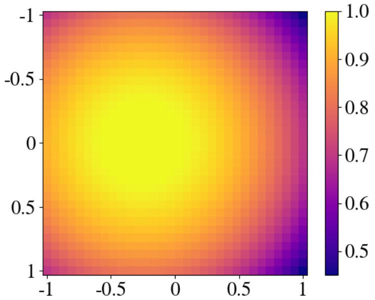

Fig. 2: An example of an RBF precision matrix with the center in (-0.25, 0.0).

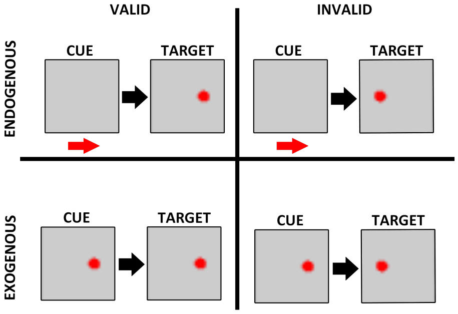

Fig. 3: Experimental setups for the four variations of the Posner Cueing Task, with visual data used in experiments. The red arrows represent symbolic endogenous cues not visible through sensory input.

3.2 Simulating the Posner Cueing Task

We evaluated the model's implementation of exogenous and endogenous covert attention using the Posner cueing task. The model received three types of sensory input: (1) camera orientation (proprioceptive), (2) a symbolic cue signal (interoceptive), and (3) visual input of an empty scene in which a red sphere may appear as a target. Unlike traditional versions, the endogenous cue was delivered as a symbolic interoceptive signal rather than a visual arrow, but still required voluntary attentional shifts. Motor actions (overt attention) were disabled.

Four task variations were tested by combining two cue types (endogenous, exogenous) with two validity conditions (valid, invalid):

- **Endogenous cueing**: the symbolic cue is internally processed to form a top-down intention, shifting both covert focus and the target-position belief.
- **Exogenous cueing**: a brief appearance of the target object triggers a bottom-up shift of covert attention via free-energy gradients and updates the belief over the object's position through the VAE likelihood error.

In valid trials, the target appeared at the cued location; in invalid trials, it appeared on the opposite side of the visual field. The task variations are illustrated in Fig. 3.

Each condition was tested with $N = 200$ trials. A single trial proceeded as follows:

- A random target position (with varying eccentricity) was generated, and the model initialized (10 steps),
- A cue (endogenous or exogenous) was presented (50 steps),
- After a specified cue-target onset asynchrony (CTOA), the target appeared,
- The trial ended either upon detection, defined as the latent variable for target presence becoming positive (marking the reaction time, RT), or after 300 steps if undetected.

The results are shown in Fig. 4. The left panel plots RTs as a function of target eccentricity, while the right panel shows internal dynamics of covert focus and target-belief updates, with clear facilitation in valid-cue conditions. The model reproduces several well-established effects in human data and location-based models:

- **Cue validity:** valid cues yield faster RTs than invalid ones [33,34], consistent with the spotlight theory of attention [7, 37, 43]. Invalid cues increase RTs due to larger spotlight shifts at target onset.
- **Cue type:** exogenous cues produce faster RTs than endogenous cues [4,13], as bottom-up signals propagate directly via error gradients, while endogenous cues require symbolic interpretation and intentional updating.
- **Target eccentricity:** RTs increase with target distance from fixation [2, 17, 32], reflecting both location-based encoding and the shape of the RBF precision function.

As shown in Fig. 4(right), covert attention centers update more rapidly than beliefs over target location in both cueing conditions. This aligns with empirical findings that covert shifts precede conscious target perception [31,33] and overt attention [16,45].

To examine the effect of cue-target onset asynchrony, all four conditions were repeated across different CTOA values. Fig. 5(left) shows that valid cues consistently yield faster RTs than invalid ones across CTOAs. The endogenous cueing results, shown in Fig. 5(right), replicate classic Posner findings [9,33,34], including the asymmetric pattern where invalid cues impose a greater cost than the benefit provided by valid cues.

3.3 Action Signals from Bottom-up Attention

Since action can be determined from free-energy optimization, overt attention in the form of eye saccades or camera orientation changes can be also implemented. Here we examined focus reach times for two action-update contributions:

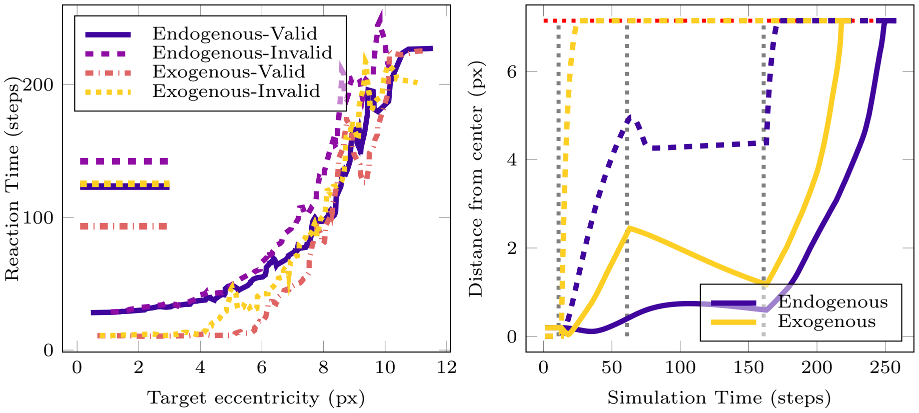

Fig. 4: Reaction time and target eccentricity relationship for endogenous and exogenous cueing. **Left:** reaction times and their averages as a function of target distance from focus point (CTOA = 100 for each trial) **Right:** covert attention center (dashed lines) and sphere position beliefs (solid lines) during valid trials, for both endogenous and exogenous cues. The horizontal line is the true target distance from center, and the vertical lines indicate trial events: the cue appears at step 10, disappears at step 60, target appears at step 160.

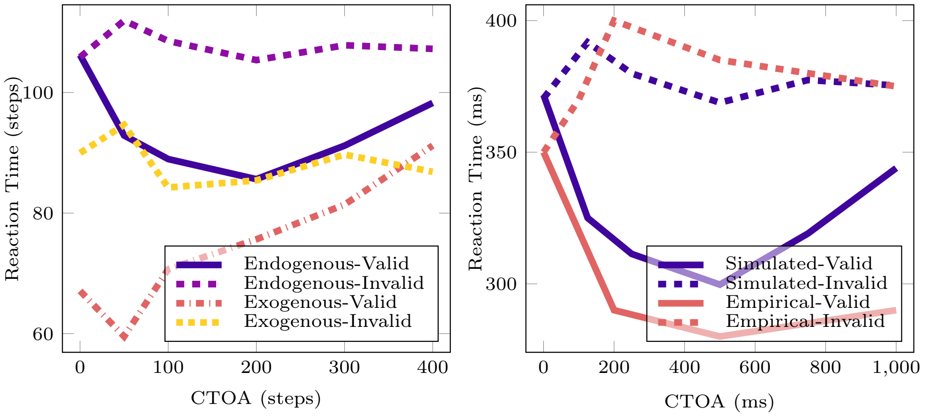

Fig. 5: Comparison of reaction times from simulated and empirical results. **Left:** average trial reaction time as a function of CTOA. Results are shown for endogenous-valid, endogenous-invalid, exogenous-valid, exogenous-invalid task variations. **Right:** comparison of simulated and empirical human data [34] for endogenous cueing. Simulated reaction times are shown up to an arbitrary constant reflecting the scale gap between human times and simulation steps.

– Top-down proprioceptive action signals: $-\frac{\partial \tilde{\boldsymbol{s}}}{\partial a}^T \tilde{\boldsymbol{\Pi}}_s \tilde{\boldsymbol{e}}_s$ – these are determined from the prediction error of the proprioceptive channel, between the proprioceptive input and current proprioceptive beliefs

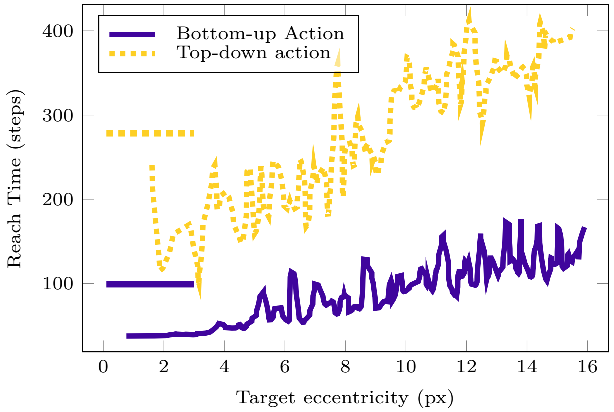

Fig. 6: Reach times and their averages for different initial target distances.

- Bottom-up visual precision action signals: $\frac{1}{2}\mathrm{Tr}\left[\boldsymbol{\tilde{\Pi}_s^{-1}}\frac{\partial\boldsymbol{\tilde{\Pi}_s}}{\partial\boldsymbol{\tilde{s}}}\right]\frac{\partial\boldsymbol{\tilde{s}}}{\partial a}-\frac{1}{2}\boldsymbol{\tilde{e}}_s^T\frac{\partial\boldsymbol{\tilde{\Pi}_s}}{\partial\boldsymbol{\tilde{s}}}\boldsymbol{\tilde{e}}_s\frac{\partial\boldsymbol{\tilde{s}}}{\partial a}$ – these are determined through the bottom-up derivative of the precision matrix. Since the action update is dependent only on the sensory input, only the second half of (15) contributes to the action update.

The trials start with a 10-step initialization interval, after which the target appears at a random position in the agent's field of view. The trial is finished when the agent successfully focuses the target at the center of its field of view.

The reach times as a function of the initial target distance can be seen in Fig. 6. The results show that bottom-up overt orienting is overall faster than top-down intentional orienting, which is explained by the sensitivity of the precision to red objects (or any predetermined visual object of interest, like faces [16]). This is similarly reflected in how reaction time changes with distance. Both forms of orienting exhibit an increasing trend in reaction time as distance increases; however, top-down orienting shows a steeper rise, indicating a greater sensitivity to distance compared to bottom-up orienting.

4 Discussion

Our proposed model was evaluated on exogenous, endogenous, valid, and invalid variations of the Posner paradigm, as well as a simple target reach task. It successfully reproduces key attentional effects observed in human data and location-based models, including the influence of exogenous versus endogenous cues, cue validity, and overt behaviors such as involuntary saccades.

A central contribution of this work is the integration of visual attention mechanisms, continuous sensory representations, and deep generative models within an active inference framework. By operating directly on raw two-dimensional visual input, the model overcomes the limitations of prior approaches that rely on discrete inputs or lack deep generative components, making it more suitable for robotic applications.

The overt attention experiments reveal a trade-off between speed and flexibility: bottom-up orienting enables rapid but singular shifts to individual objects, while top-down orienting is slower but allows flexible allocation across multiple objects. This trade-off has direct implications for robotics. One application is active visual search, where robots adjust their viewpoint to locate a target object, potentially using intermediate objects to guide attention through top-down processes [19, 39, 40, 44, 46]. For instance, when searching for a keyboard, a desk might serve as a top-down cue directing attention upward, while the keyboard itself, once detected, would attract bottom-up attention due to its salient features. Another application is joint attention in human–robot interaction, where human gaze and head movements guide the robot's attention toward relevant objects in the shared environment [15, 25]. In both cases, our model integrates bottom-up attention, where salient objects attract attention directly, and top-down attention, where contextual cues guide attention across the scene—both essential for natural and adaptive active vision in robots.

5 Conclusion

In this paper, we have proposed an active inference model of covert and overt visual attention. The proposed model successfully demonstrates known attentional phenomena and mechanisms in the context of the Posner cueing task and a simple active orienting task. It shows that valid cues produce faster reaction times than invalid cues, and that exogenous cues produce faster reaction times than endogenous cues. The model also successfully demonstrates location-based attention, with reaction times increasing with target eccentricity.

Future work will extend the model with multiple possible targets/intentions to further test object-based and location-based effects, as well as with top-down attentional mechanisms that lead to inhibition of return. Overt saccades will also be examined further, with a focus on varying attraction to different objects in tasks such as active visual search and joint attention. We plan to further develop and test this framework as a model of perception, learning, and action in autonomous robots.

Acknowledgments. This research has been supported by the H2020 project AIFORS under Grant Agreement No 952275.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Blair, C.D., Ristic, J.: Attention combines similarly in covert and overt conditions. Vision (Basel) **3**(2), 16 (Apr 2019)

2. Carrasco, M., Evert, D.L., Chang, I., Katz, S.M.: The eccentricity effect: target eccentricity affects performance on conjunction searches. Percept. Psychophys. **57**(8), 1241–1261 (Nov 1995)
3. Çatal, O., Wauthier, S., Verbelen, T., De Boom, C., Dhoedt, B.: Deep active inference for autonomous robot navigation (Mar 2020)
4. Cheal, M., Lyon, D.R.: Central and peripheral precuing of forced-choice discrimination. Q. J. Exp. Psychol. A **43**(4), 859–880 (Nov 1991)
5. Cioffi, G., Cieslewski, T., Scaramuzza, D.: Continuous-time vs. discrete-time vision-based slam: A comparative study. IEEE Robotics and Automation Letters **7**(2), 2399–2406 (2022). https://doi.org/10.1109/LRA.2022.3143303
6. Corbetta, M., Shulman, G.L.: Control of goal-directed and stimulus-driven attention in the brain. Nat. Rev. Neurosci. **3**(3), 201–215 (Mar 2002)
7. Egly, R., Driver, J., Rafal, R.D.: Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects. J. Exp. Psychol. Gen. **123**(2), 161–177 (1994)
8. Eliasmith, C., Furlong, P.M.: Continuous then discrete: A recommendation for building robotic brains. In: Faust, A., Hsu, D., Neumann, G. (eds.) Proceedings of the 5th Conference on Robot Learning. Proceedings of Machine Learning Research, vol. 164, pp. 1758–1763. PMLR (08–11 Nov 2022), https://proceedings.mlr.press/v164/eliasmith22a.html
9. Feldman, H., Friston, K.J.: Attention, uncertainty, and free-energy. Front. Hum. Neurosci. **4** (2010)
10. Friston, K.: The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. **11**(2), 127–138 (Feb 2010)
11. Friston, K., Kilner, J., Harrison, L.: A free energy principle for the brain. J. Physiol. Paris **100**(1-3), 70–87 (Jul 2006)
12. Itti, L., Baldi, P.: Bayesian surprise attracts human attention. Vision Res. **49**(10), 1295–1306 (Jun 2009)
13. Jonides, J.: Voluntary versus automatic control over the mind's eye's movement. In: Attention and Performance IX, pp. 187–203 (1981)
14. Kanai, R., Komura, Y., Shipp, S., Friston, K.: Cerebral hierarchies: predictive processing, precision and the pulvinar. Philos. Trans. R. Soc. Lond. B Biol. Sci. **370**(1668), 20140169 (May 2015)
15. Kaplan, F., Hafner, V.V.: The challenges of joint attention. Interaction Studies **7**(2), 135–169 (2006). https://doi.org/https://doi.org/10.1075/is.7.2.04kap, https://www.jbe-platform.com/content/journals/10.1075/is.7.2.04kap
16. Kauffmann, L., Peyrin, C., Chauvin, A., Entzmann, L., Breuil, C., Guyader, N.: Face perception influences the programming of eye movements. Sci. Rep. **9**(1), 560 (Jan 2019)
17. Klein, R.M.: Inhibition of return. Trends Cogn. Sci. **4**(4), 138–147 (Apr 2000)
18. Kulke, L.V., Atkinson, J., Braddick, O.: Neural differences between covert and overt attention studied using EEG with simultaneous remote eye tracking. Front. Hum. Neurosci. **10**, 592 (Nov 2016)
19. Kunze, L., Doreswamy, K.K., Hawes, N.: Using qualitative spatial relations for indirect object search. In: 2014 IEEE International Conference on Robotics and Automation (ICRA). IEEE (May 2014)
20. Lanillos, P., Dean-Leon, E., Cheng, G.: Multisensory object discovery via self-detection and artificial attention. In: 2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob). IEEE (Sep 2016)

21. Lanillos, P., Ferreira, J.F., Dias, J.: Designing an artificial attention system for social robots. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE (Sep 2015)
22. Lanillos, P., Ferreira, J.F., Dias, J.: Multisensory 3d saliency for artificial attention systems (09 2015)
23. Mengotti, P., Käsbauer, A.S., Fink, G.R., Vossel, S.: Lateralization, functional specialization, and dysfunction of attentional networks. Cortex **132**, 206–222 (Nov 2020)
24. Mirza, M.B., Adams, R.A., Friston, K., Parr, T.: Introducing a bayesian model of selective attention based on active inference. Sci. Rep. **9**(1), 13915 (Sep 2019)
25. Nagai, Y., Hosoda, K., Morita, A., Asada, M.: A constructive model for the development of joint attention. Conn. Sci. **15**(4), 211–229 (Dec 2003)
26. Parr, T., Benrimoh, D.A., Vincent, P., Friston, K.J.: Precision and false perceptual inference. Front. Integr. Neurosci. **12**, 39 (Sep 2018)
27. Parr, T., Friston, K.J.: Uncertainty, epistemics and active inference. J. R. Soc. Interface **14**(136), 20170376 (Nov 2017)
28. Parr, T., Friston, K.J.: Attention or salience? Curr. Opin. Psychol. **29**, 1–5 (Oct 2019)
29. Parr, T., Pezzulo, G., Friston, K.J.: Active inference. The MIT Press (2022)
30. Parvizi-Wayne, D.: How preferences enslave attention: calling into question the endogenous/exogenous dichotomy from an active inference perspective. Phenomenol. Cogn. Sci. (Sep 2024)
31. Peterson, M.S., Kramer, A.F., Irwin, D.E.: Covert shifts of attention precede involuntary eye movements. Percept. Psychophys. **66**(3), 398–405 (Apr 2004)
32. Pinker, S., Downing, C.J.: Attention and Performance XI: Mechanisms of attention and visual search. Erlbaum, Hillsdale, NJ (1985)
33. Posner, M.I.: Orienting of attention. Q. J. Exp. Psychol. **32**(1), 3–25 (Feb 1980)
34. Posner, M., Nissen, M., Ogden, W.: Attended and unattended processing modes: The role of set for spatial location. Modes of Perceiving and Processing Information **137** (01 1978)
35. Priorelli, M., Stoianov, I.P.: Flexible intentions: An active inference theory. Front. Comput. Neurosci. **17**, 1128694 (Mar 2023)
36. Rao, R.P.: An optimal estimation approach to visual perception and learning. Vision Research **39**(11), 1963–1989 (1999). https://doi.org/https://doi.org/10.1016/S0042-6989(98)00279-X, https://www.sciencedirect.com/science/article/pii/S004269899800279X
37. Reppa, I., Schmidt, W.C., Leek, E.C.: Successes and failures in producing attentional object-based cueing effects. Atten. Percept. Psychophys. **74**(1), 43–69 (Jan 2012)
38. Shomstein, S., Zhang, X., Dubbelde, D.: Attention and platypuses. Wiley Interdiscip. Rev. Cogn. Sci. **14**(1), e1600 (Jan 2023)
39. Shubina, K., Tsotsos, J.K.: Visual search for an object in a 3D environment using a mobile robot. Comput. Vis. Image Underst. **114**(5), 535–547 (May 2010)
40. Sjöö, K., Aydemir, A., Jensfelt, P.: Topological spatial relations for active visual search. Rob. Auton. Syst. **60**(9), 1093–1107 (Sep 2012)
41. Škrjanc, I., Klančar, G.: A comparison of continuous and discrete tracking-error model-based predictive control for mobile robots. Rob. Auton. Syst. **87**, 177–187 (Jan 2017)
42. Spratling, M.W.: Predictive coding as a model of biased competition in visual attention. Vision Res. **48**(12), 1391–1408 (Jun 2008)

43. Vecera, S.P., Farah, M.J.: Does visual attention select objects or locations? J. Exp. Psychol. Gen. **123**(2), 146–160 (1994)
44. Vogel, J., Murphy, K.: A non-myopic approach to visual search. In: Fourth Canadian Conference on Computer and Robot Vision (CRV '07). IEEE (May 2007)
45. Walker, R., Walker, D.G., Husain, M., Kennard, C.: Control of voluntary and reflexive saccades. Exp. Brain Res. **130**(4), 540–544 (Feb 2000)
46. Wixson, L.E., Ballard, D.H.: Using intermediate objects to improve the efficiency of visual search. Int. J. Comput. Vis. **12**(2-3), 209–230 (Apr 1994)