Appendix A. Model details

On the first frame, the tracker initializes a series of tracks, one for each detection, representing the objects to be tracked. Each track represents the belief about an object's position variables (bounding box center, height, aspect ratio, and velocities) and appearance. New observations are assigned deterministically to tracks/slots and update beliefs about position and appearance.

We denote observations with $\mathbf{o}$ (position observations: $\mathbf{o}^{\text{pos}}$, appearance observations: $\mathbf{o}^{\text{app}}$) and latent variables with $\mathbf{z}$ (position latent variables: $\mathbf{z}^{\text{pos}}$, appearance latent variables: $\mathbf{z}^{\text{app}}$).

A.1. Position model

Each track's position state is an eight-dimensional variable (bounding box center, height, aspect ratio, and velocities), which is updated with new observations by a Kalman filter.

On time step t, first the predicted posterior $\mathcal{N}(\hat{\mathbf{z}}_{i,t}^{\text{pos}}, \hat{\mathbf{P}}_{t}^{\text{pos}})$ is computed with

$$\hat{\mathbf{z}}_{i,t}^{\text{pos}} = \mathbf{F}\mathbf{z}_{i,t-1}^{\text{pos}} \tag{6}$$

$$\hat{\mathbf{P}}_{t}^{\text{pos}} = \mathbf{F}\mathbf{P}_{t-1}^{\text{pos}}\mathbf{F}^{\mathsf{T}} + \mathbf{Q} \tag{7}$$

where $\mathbf{z}_{t-1}^{\text{pos}}$, and $\mathbf{P}_{t-1}^{\text{pos}}$ are the state vector and covariance of the posterior belief from the previous frame, F is the state-transition matrix:

$$\mathbf{F} = \begin{bmatrix} 1 & 0 & 0 & 0 & \Delta t & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & \Delta t & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & \Delta t & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & \Delta t \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and the covariance of the process noise $\mathbf{Q} = \left[2.1, 2.1, 0.01, 2.1, 0.26, 0.26, 1 \times 10^{-5}, 0.26\right] \mathbf{I}$.

The new posterior for time step t is then updated as a combination of the predicted posterior and the assigned new observation $\mathbf{o}_{j,t}$, weighted by the Kalman gain:

$$\mathbf{z}_{i,t}^{\text{pos}} = (\mathbf{I} - \mathbf{K}_t\mathbf{H})\hat{\mathbf{z}}_{i,t}^{\text{pos}} + \mathbf{K}_t\mathbf{o}_{j,t} \tag{8}$$

$$\mathbf{P}_{i,t}^{\text{pos}} = (\mathbf{I} - \mathbf{K}_t\mathbf{H})\,\hat{\mathbf{P}}_{i,t}^{\text{pos}} \tag{9}$$

where $\mathbf{H}$ projects latent beliefs into the observation space:

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

The Kalman gain weighs the contribution of the predicted posterior belief and the current observation to the new state belief, depending on the uncertainty in the predicted posterior $\hat{\mathbf{P}}^{\text{pos}}_{i,t}$ and the uncertainty over the expected observation $\mathbf{S}^{\text{pos}}_{i,t} = \mathbf{H}\hat{\mathbf{P}}^{\text{pos}}_{i,t}\mathbf{H}^{\mathsf{T}} + \mathbf{R}^{\text{pos}}_{j,t}$, where $\mathbf{R}^{\text{pos}}_{j,t}$ is the covariance matrix of the observation noise. Note that observation noise is zero for the base model, constant ($\mathbf{R}^{\text{pos}}_{j,t} = \mathbf{R}^{\text{pos}}$) for the constant model, and a function of the distance between fixation and observation position in the fixation model.

A.2. Appearance model

For each bounding box observation, the corresponding image crop is extracted and embedded into a latent space to yield a 128-dimensional appearance observation (extracted via a pre-trained re-identification model with a ResNet50 backbone). A track's belief about the object's appearance is modeled as an empirical distribution over past observations of the object. In particular, the Gaussian belief distribution $\mathcal{N}(\mathbf{z}^{\text{app}}_{i,t}, \mathbf{P}^{\text{app}}_{i,t})$ is parameterized via the precision-weighted mean $\mathbf{z}^{\text{app}}_{t}$ and covariance $\mathbf{P}^{\text{app}}_{i,t}$ over the past K ($K = \min\{t, 10\}$) appearance embeddings. The precision weight of a sample in memory corresponds to the inverse variance of the observation noise associated with the observation.

The predicted appearance observation for a particular track i at time-point t is then modeled as $\mathcal{N}(\hat{\mathbf{o}}^{\text{app}}_{i,t}, \mathbf{S}^{\text{app}}_{i,t})$ with $\hat{\mathbf{o}}^{\text{app}}_{i,t} = \mathbf{z}^{\text{app}}_{i,t-1}$ and $\mathbf{S}^{\text{app}}_{i,t} = \mathbf{P}^{\text{app}}_{t-1} + \mathbf{R}^{\text{app}}_{j,t}$ (the projection $\mathbf{H}$ from latent into observation space is simply the identity matrix $\mathbf{I}$).

A.3. Assignment

New observations are associated with those tracks that minimize the distances between the tracker's belief about object positions and appearances and the new observations.

In particular, we compute the distance $d^{\text{pos}}(i, j)$ between the tracker's belief about the i-th track's position and the position of the j-th observation $\mathbf{o}^{\text{pos}}_{j,t}$ as the negative log probability of the observation under the model's probabilistic prediction of the object's position, $\mathcal{N}(\hat{\mathbf{o}}^{\text{pos}}_{i,t}, \mathbf{S}^{\text{pos}}_{i,t})$, with $\hat{\mathbf{o}}^{\text{pos}}_{i,t} = \mathbf{H}\hat{\mathbf{z}}^{\text{pos}}_{i,t}$ as the predicted position in observation space.

$$d^{\text{pos}}(i,j) = \frac{1}{2}\Big((\mathbf{o}^{\text{pos}}_{j,t} - \hat{\mathbf{o}}^{\text{pos}}_{i,t})^{\mathsf{T}}\Big(\mathbf{S}^{\text{pos}}_{i,t}\Big)^{-1}(\mathbf{o}^{\text{pos}}_{i,t} - \hat{\mathbf{o}}^{\text{pos}}_{i,t}) + \log|\mathbf{S}^{\text{pos}}_{i,t}| + k\log(2\pi)\Big) \tag{10}$$

Similarly, the distance between distance $d^{\text{app}}(i, j)$ between the tracker's belief about the i-th track's appearance and the observed appearance is computed as the negative log probability of the observed appearance embedding $\mathbf{o}^{\text{app}}_{j,t}$ under the model's belief about the track's appearance $\hat{\mathbf{o}}^{\text{app}}_{i,t}$ and the associated uncertainty $\mathbf{S}^{\text{app}}_{i,t}$

$$d^{\text{app}}(i,j) = \frac{1}{2}\Big((\mathbf{o}^{\text{app}}_{j,t} - \hat{\mathbf{o}}^{\text{app}}_{i,t})^{\mathsf{T}}\Big(\mathbf{S}^{\text{app}}_{i,t}\Big)^{-1}(\mathbf{o}^{\text{app}}_{i,t} - \hat{\mathbf{o}}^{\text{app}}_{i,t}) + \log|\mathbf{S}^{\text{app}}_{i,t}| + k\log(2\pi)\Big) \tag{11}$$

on each time step, the entries of the assignment cost matrix $\mathbf{C}t$ is computed as a weighted sum of the position cost D^{pos} and the appearance costs D^{app} between the i-th track and the j-th observation.

$$c_{i,j} = \lambda d^{\text{pos}}(i,j) + (1-\lambda)d^{\text{app}}(i,j) \tag{12}$$

Appendix B. Object motion trajectories

Initial object positions were sampled randomly such that no object was occluded and objects had a minimal inter-object distance of half an object width. Initial angular motion directions were sampled from a uniform distribution. Object speed was always constant. Using rejection sampling, motion trajectories were sampled such that at the end of the motion period, object centroids were separated by at least half an object distance.

Object motion dynamics could be either linear or following a complex generative motion model. In the linear case, object positions were deterministically simulated forward using the initial start position and motion vector. In the complex motion model, the angular motion direction was perturbed on each frame. At the start of the motion, the motion perturbation angle was sampled from a von Mises distribution centered on $\mu = 0$ degrees with a precision of $\kappa = 100.0$ degrees. This motion perturbation angle was then applied for the next T_1 frames. At frame $T_1 + 1$, a new motion perturbation angle was sampled from the same von Mises distribution and applied for the next T_2 frames. Intervals $T_1, T_2, \ldots$ were sampled from a Poisson distribution with $\lambda = 10$. Hence, on average every 10 frames, the motion direction of the object changed. This procedure yielded complex but smooth motion trajectories.

Appendix C. Human gaze tracking

Nine participants (7 female, mean $\pm$ age 26.9 ± 8.3) with normal or corrected-to-normal vision were recruited from the participant pool of the Institute of Neuroscience and Psychology, University of Glasgow. All participants gave informed consent and the study was approved by the ethics committee of the College of Medical, Veterinary & Life Sciences of the University of Glasgow. Participants viewed stimuli in the lab on a monitor (1920×1080 resolution, 60Hz refresh rate). Monocular gaze (left eye) was recorded at a sampling rate of 1000Hz using an infrared camera (Eyelink 1000, SR Research). The camera was positioned under the display monitor facing the participants. Participants used a chin-rest, which allowed us to control the distance between the eyes and the monitor (distance: 57cm) and minimized head motion. The eye tracker was calibrated before each block.

Appendix D. Human gaze behavior reveals beliefs about tracked objects

Fixation behavior is a core feature of human visual inference enabling targeted sampling of the environment. To gauge to what extent fixations are directly subserving the task rather than being a behavior coincidental to the task, we plotted the distance of all objects to the fixation center as a function of time in the trial and object type (Figure 5). We observe that fixations were closer to targets compared to distractor objects in line with previous findings (Hyönä et al., 2019). Moreover, fixation behavior revealed the underlying beliefs of participants about target and distractor objects. In particular, distractor objects which were (incorrectly) selected by participants in the responses period, were closer to the fixation during the motion period compared to objects which were not selected. This demonstrates the relevance of human fixation behavior for multiple object tracking.

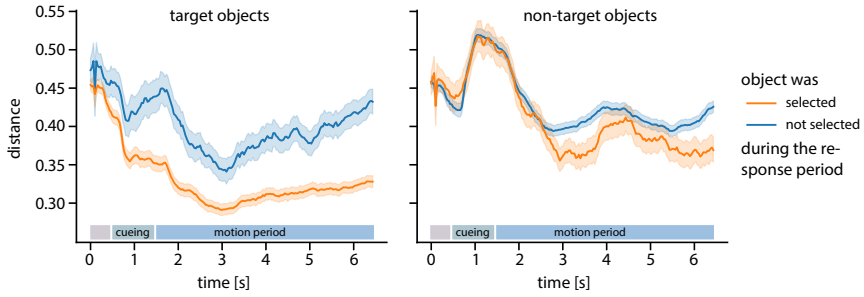

Figure 5: **Fixation behavior is related to the tracking task**. Average distance of target (left) and non-target objects (right) to the fixation and as a function of whether the object was believed to be a target object as indicated by the behavioral response. Distance as a fraction of the vertical (horizontal) extent of the motion area (i.e., relative to a motion area of 1×1).