TPP-Gaze: Modelling Gaze Dynamics in Space and Time with Neural Temporal Point Processes

Alessandro D'Amelio[1], Giuseppe Cartella[2], Vittorio Cuculo[2], Manuele Lucchi[1], Marcella Cornia[2], Rita Cucchiara[2], Giuseppe Boccignone[1]

[1] *University of Milan, Italy* [2] *University of Modena and Reggio Emilia, Italy*

WACV 2025
TUCSON, ARIZONA • FEB 28 - MAR 4

Have a look at our repo!

1. Motivation

Scanpath Prediction is the task of predicting the spatial and temporal patterns of human eye movements, including the sequence and timing of gaze shifts.

- While existing computational models effectively predict spatial aspects of observer's visual scanpaths (***where*** to look), they often overlook the temporal facet of attention dynamics (***when***).
- The few approaches able to predict fixation duration are **fully engineered**, and do not address the problem in a **mathematical principled way**.

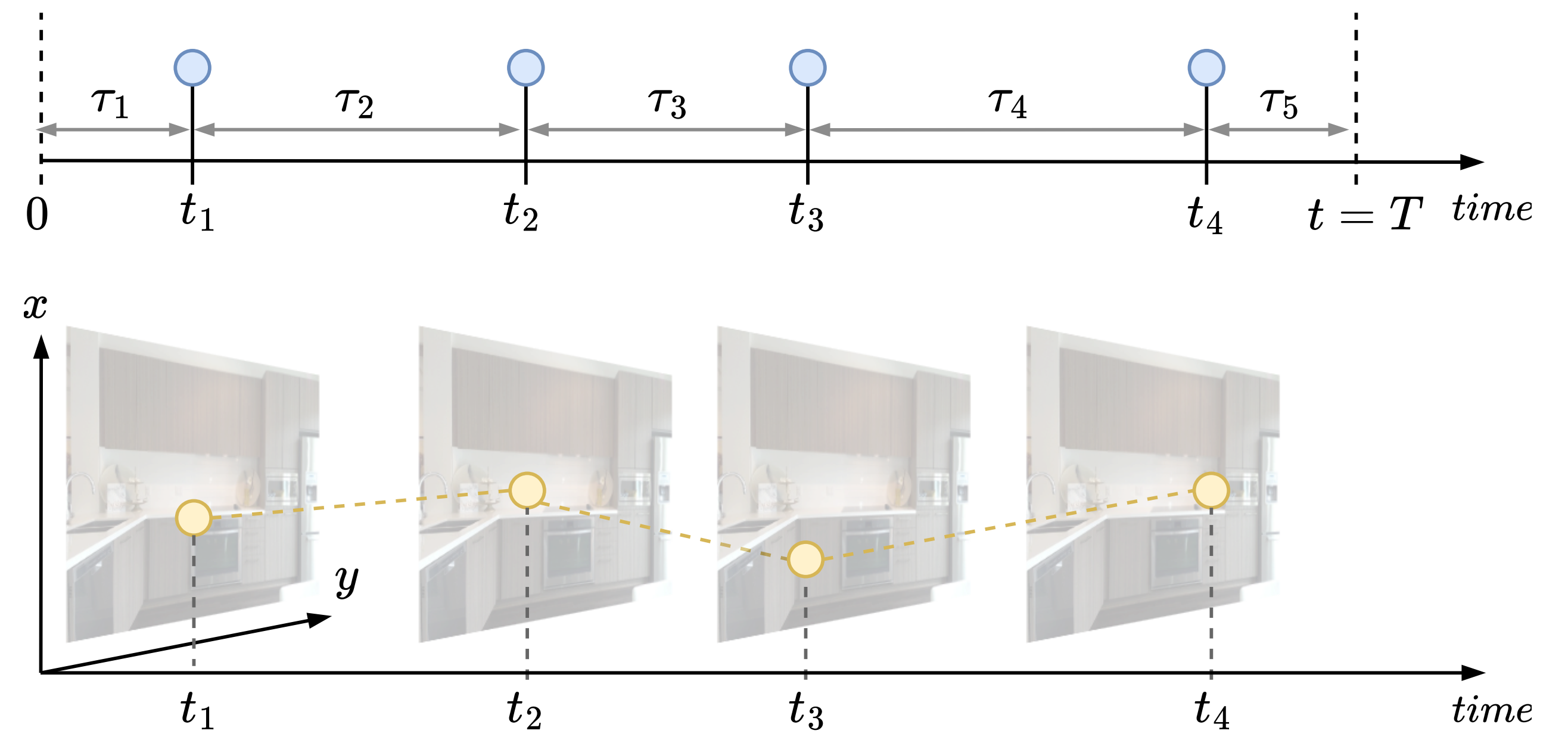
We propose a novel view on the problem of scanpath prediction by considering a scanpath as the realisation of a ***Neural Temporal Point Process*** [1, 2].

2. Idea: Modelling Gaze Dynamics as Neural TPPs

Neural TPPs model the next arrival time of an event by conditioning on past events. $\mathcal{H}_t = \{t_n \in \mathcal{T} : t_n < t\}$ denotes the history of arrival times of all events up to time t.

- Time distribution might depend on factors other than the history. A **marked TPP** is a random process whose realisations consist of a sequence of discrete events localised in time, $\{r_{F_n}, t_n\}$, with $t_n \in \mathbb{R}^+$ and the mark $r_{F_n} \in \mathbb{R}^2$.

The modelling assumptions of Neural TPPs align with the structure of scanpath data.

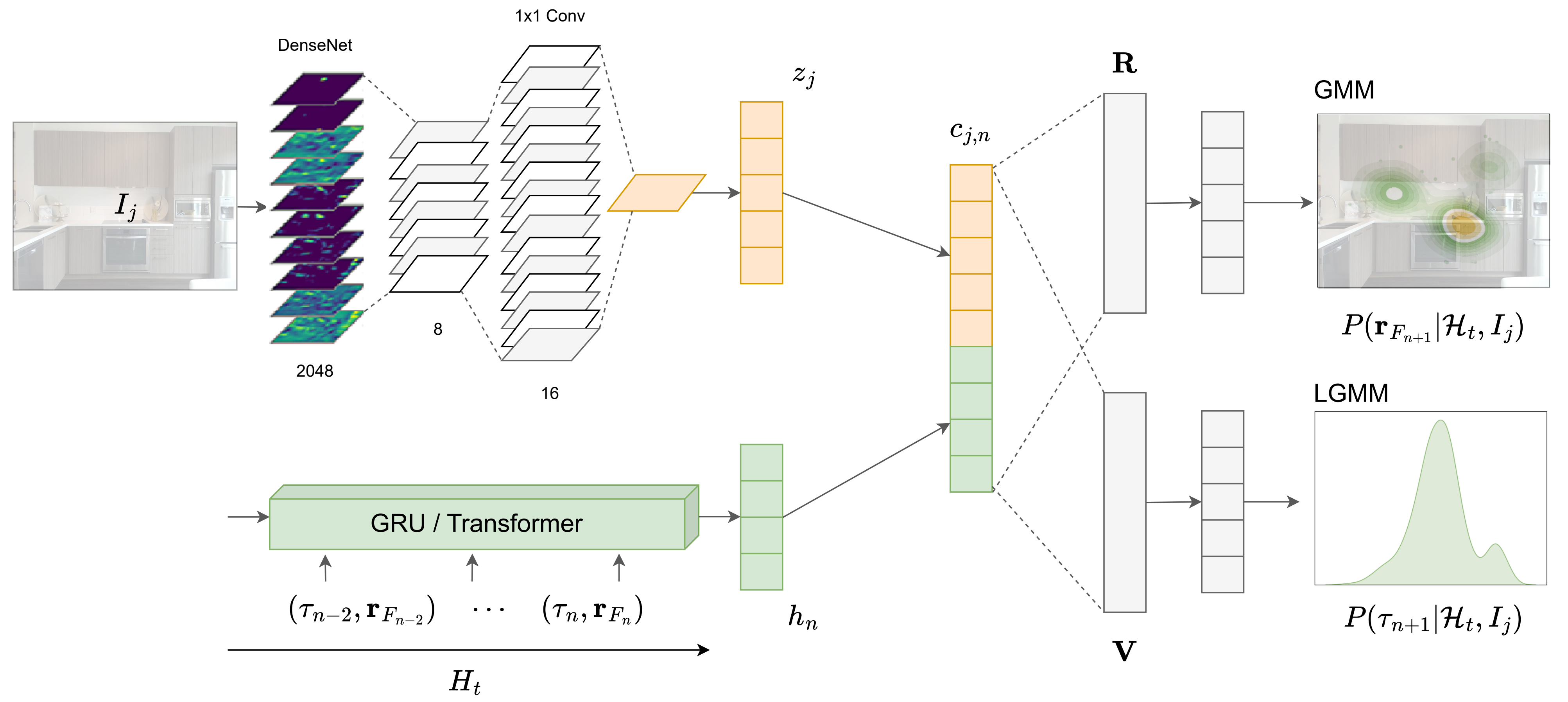


Modelling scanpaths entails defining a mapping from visual stimulus, I to a sequence of time-stamped gaze locations, $S = \{(r_{F_1}, t_1), (r_{F_2}, t_2), \dots, (r_{F_N}, t_N)\}$, where $r_{F_n} \in \mathbb{R}^2$ denotes the two-dimensional spatial coordinates of the n-th fixation on the stimulus I, while $t_n \in \mathbb{R}^+$ represents the corresponding arrival time.

Our Contributions

- We propose `TPP-Gaze`, a novel scanpath model based on Neural TPPs, that jointly learns the temporal dynamics of fixations position and duration.
- We extend recent Neural TPP models to deal with visual data (i.e., images) and connect scanpath modelling and prediction to point process theory.

3. Proposed Method: TPP-Gaze


I_j
DenseNet
2048
8
1x1 Conv
16
z_j
$c_{j,n}$
R
GMM
$P(\mathbf{r}_{F_{n+1}}|\mathcal{H}_t, I_j)$
LGMM
$P(\tau_{n+1}|\mathcal{H}_t, I_j)$
V
GRU / Transformer
$(\tau_{n-2}, \mathbf{r}_{F_{n-2}})$
⋯
$(\tau_n, \mathbf{r}_{F_n})$
h_n
H_t

Representing Scene Semantics: The history of past events also depends on the input visual stimulus **I**. We extract a perceptual representation z_j of I through a DenseNet.

History of Past Events: The pair (r_{F_n}, τ_n) represents the event at time t_n with fixation position r_{F_n} and duration $\tau_n = t_n - t_{n-1}$. The Transformer/GRU state embedding h_n represents the influence of the history up to the $n-$th fixation.

$$S^i_{n+1} \sim p_\theta(\mathbf{r}_{F_{n+1}}, t_{n+1} \mid \mathbf{h}_n, \mathbf{z}_j).$$

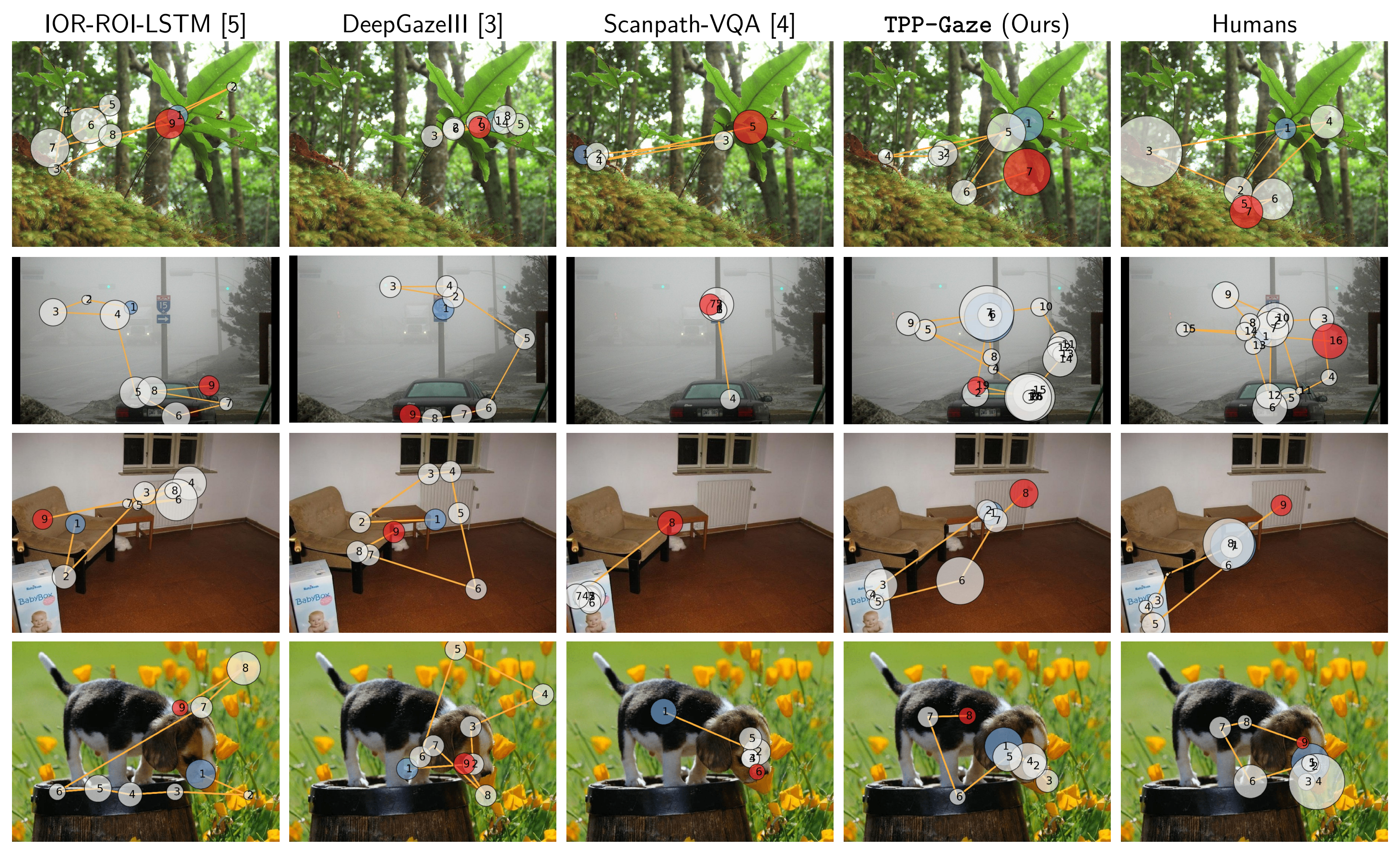
We model the conditional dependence of the distribution $p_\theta(\tau_{n+1} \mid \mathbf{h}_n, \mathbf{z}_j)$ on both past events and stimulus by concatenating the history embedding and semantic vectors into a context vector $\mathbf{c}_{j,n} = [\mathbf{h}_n \mid\mid \mathbf{z}_j]$.

Fixation Duration Generation: The context vector $\mathbf{c}_{j,n}$ is employed to learn the parameters of a Log-Gaussian Mixture Model.
Fixation Position Generation: The conditional probability of the next mark (fixation position), $p_\theta(\mathbf{r}_{F_{n+1}} \mid \mathbf{h}_n, \mathbf{z}_j)$ is defined as a 2D Gaussian Mixture Model.

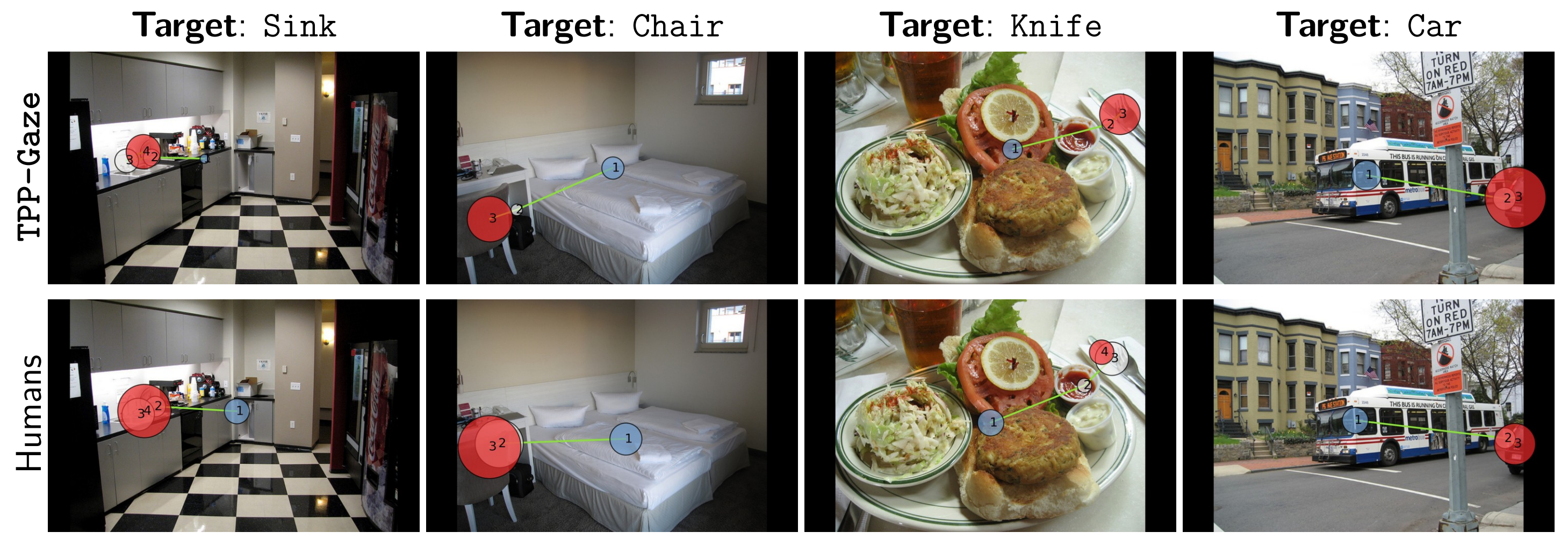
The loss function is a negative log-likelihood defined as: $\mathcal{L}(\boldsymbol{\theta}) = -\sum_j \sum_i \sum_n \left[\log p^*_\theta(\tau^i_n \mid \mathbf{c}_{j,n}) + \log p^*_\theta(\mathbf{r}^i_{F_n} \mid \mathbf{c}_{j,n})\right]$.

4. Qualitative Examples

`TPP-Gaze` predicts scanpaths better aligned with those from human subjects.



5. Extension to Visual Search



Let $\mathbf{F}_{target}$ be the embedding vector representing the search objective obtained through the RoBERTa language model.

The visual backbone for the visual search model is modified to output $M = 256$ feature maps. Let $X = [x_0, \dots, x_M] \in \mathbb{R}^{M»d}$ represent the matrix of flattened image features. The task-specific semantic representation for the j-th image, $\mathbf{z}_{j,target}$, is:

$$\mathbf{z}_{j,target} = \sum_{i=1}^{M} w_i \mathbf{x}_i, \quad w = \mathrm{softplus}(\mathrm{MLP}(\mathbf{F}_{target}))$$

6. Comparison with the State of the Art

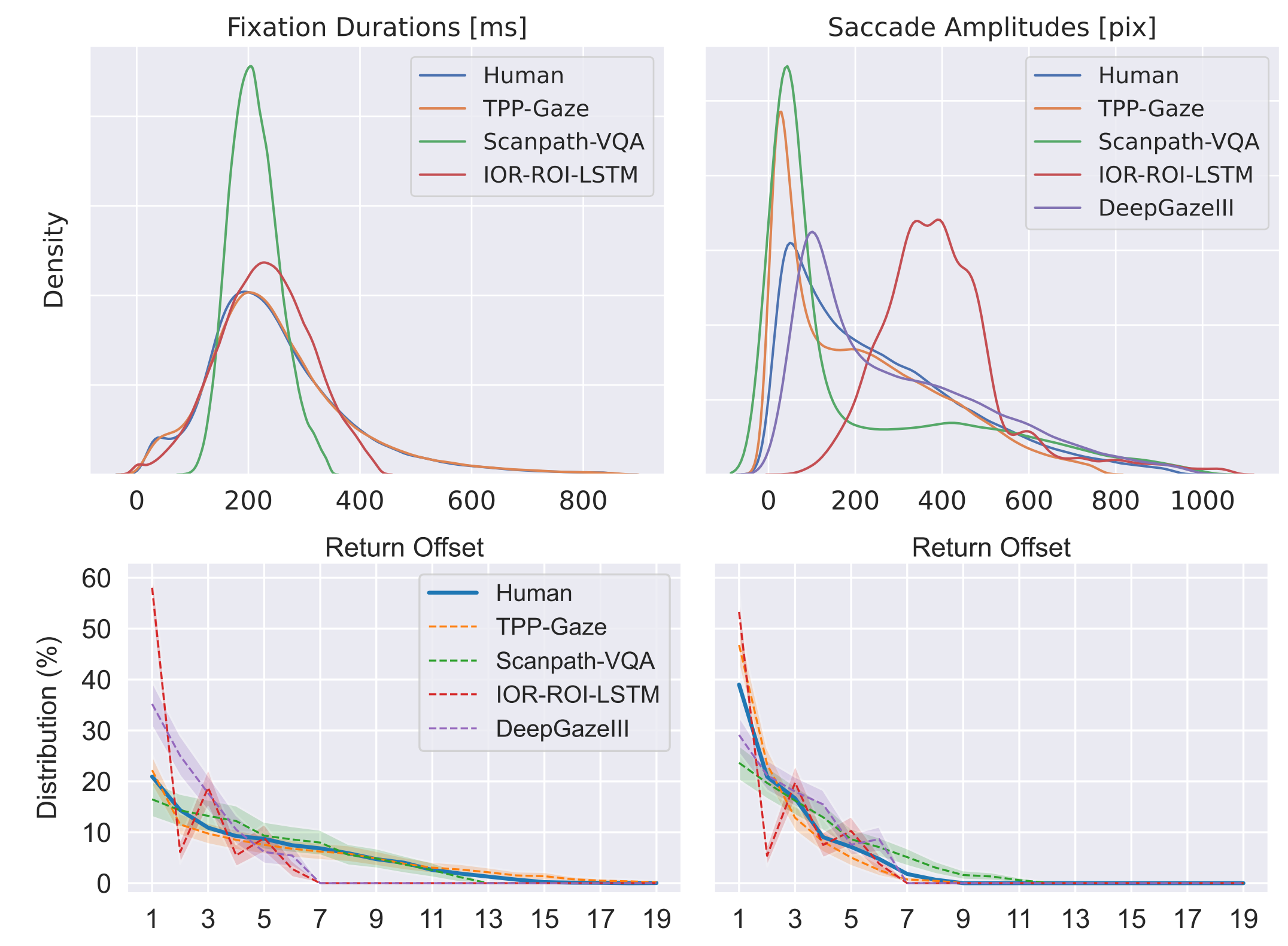
`TPP-Gaze` (with either GRU or Transformer-based history encoding) outperforms all the other approaches on most of the adopted metrics.

	COCO-FreeView			MIT1003			OSIE			NUSEF		
	MM ↓	SM ↓	SS ↓	MM ↓	SM ↓	SS ↓	MM ↓	SM ↓	SS ↓	MM ↓	SM ↓	SS ↓
Itti-Koch [7]	0.51	-	-	0.95	-	-	1.66	-	-	0.45	-	-
CLE (Itti) [8]	0.54	-	-	0.39	-	-	0.28	-	-	0.20	-	-
CLE (DG) [8]	0.44	-	-	-	-	-	0.24	-	-	0.15	-	-
G-Eymol [6]	1.05	9.00	8.75	0.88	15.90	3.32	1.16	16.17	12.28	0.81	1.76.	1.99
IOR-ROI-LSTM [5]	0.38	1.54	0.56	0.31	0.69	5.08	0.69	0.75	0.20	0.36	0.11	0.06
DeepGazeIII [3]	**0.03**	-	-	-	-	-	0.11	-	-	0.07	-	-
Scanpath-VQA [4]	0.12	1.07	0.43	0.07	0.06	0.05	0.08	0.03	0.02	0.06	0.02	0.02
DeepGazeIII [3]	0.04	-	-	0.08	-	-	0.08	-	-	0.08	-	-
Scanpath-VQA [4]	0.23	**0.08**	**0.03**	0.12	0.23	0.14	0.23	0.40	0.29	0.09	0.06	0.06
`TPP-Gaze` (GRU.)	**0.03**	**0.08**	0.05	**0.04**	**0.15**	**0.12**	**0.05**	**0.20**	**0.25**	**0.04**	**0.04**	**0.02**
`TPP-Gaze` (Trans.)	**0.03**	0.10	0.06	**0.04**	0.22	0.14	0.06	0.25	0.29	**0.04**	**0.04**	**0.02**

MM, SM, and SS average values may deliver inconsistent results: models exhibiting less variability w.r.t. humans, can score better than the ground truth model.

Our proposal: Considering a good model as the one that minimises the discrepancy between the target and model-derived score distributions.

7. Statistics of the Generated Scanpaths



Statistical properties of simulated scanpaths closely resemble those from real observers.

Return fixations pattern better in alignment with real ones if compared to SOTA.

References

[1] O. Shchur et al., *Intensity-Free Learning of Temporal Point Processes*, in ICLR, 2020.

[2] O. Shchur et al., *Neural Temporal Point Processes: A Review*, in IJCAI, 2021.

[3] M. Kümmerer et al., *DeepGaze III: Modeling free-viewing human scanpaths with deep learning*, in J. of Vision, 2022.

[4] X. Chen et al., *Predicting Human Scanpaths in Visual Question Answering*, in CVPR, 2021.

[5] Z. Chen et al., *Scanpath Prediction for Visual Attention using IOR-ROI LSTM*, in IJCAI, 2018.

[6] D. Zanca et al., *Gravitational Laws of Focus of Attention*, in TPAMI, 2019.

[7] L. Itti et al., *A model of saliency-based visual attention for rapid scene analysis*, in TPAMI, 1998.

[8] G. Boccignone et al., *Modelling gaze shift as a constrained random walk*, in Physica A, 2004.