Modeling Human Gaze Behavior with Diffusion Models for Unified Scanpath Prediction

Supplementary Material

In the following, we provide additional results on the analysis of scanpath variability, demonstrating that `ScanDiff` outperforms existing methods in capturing the diversity of human gazes, along with achieving state-of-the-art results in traditional scanpath prediction metrics.

A. Additional Quantitative Results

Additional Comparison Details. ChenLSTM-ISP [14] is originally designed to generate user-specific scanpaths. In contrast, our model accounts for population diversity. To adapt ChenLSTM-ISP to our setting, we make it predict a single scanpath conditioned on an image, a task, and a user identifier. By simulating this process across different user identifiers, we obtain a population of scanpaths representing diverse subjects. For TPP-Gaze [23], we re-train it on our splits, as training data used in the original model may include samples from our test sets.

Zero-shot Evaluation. As a complement of Table 1 of the main paper, we report in Table 6 the results obtained on the OSIE dataset [62], which is used exclusively for zero-shot evaluation and not included in the training process. This analysis allows us to assess the generalization capability of `ScanDiff` when applied to unseen data. Our model achieves the best overall results when compared with models that does not use OSIE as training set. This highlights the ability of the model to generate plausible and diverse scanpaths without requiring dataset-specific fine-tuning, demonstrating its robustness also in out-of-distribution scenarios.

Additional Metrics on COCO-Search18. Table 7 presents a detailed breakdown of `ScanDiff`'s performance on the COCO-Search18 dataset across individual MM metrics for both target-present and target-absent scenarios. Extending the results reported in Table 2 of the main paper, `ScanDiff` consistently outperforms existing methods in modeling spatial characteristics of scanpaths with substantial margins. In the target-present condition, our approach achieves KL-divergence reductions of $68-78\%$ for shape, $79-81\%$ for length, $77-88\%$ for direction, and $50-99\%$ for position compared to models trained under identical settings (highlighted in gray). Similar improvements are observed in the target-absent condition. While TPP-Gaze shows slightly better performance in duration modeling for target-absent cases, `ScanDiff` maintains competitive performance in duration metrics for target-present scenarios (*i.e.*, 0.033). These results further validate the effectiveness of our model in capturing the complex dynamics of task-driven visual behavior across different search conditions.

	OSIE									
	MM ↓						SM ↓		SS ↓	
	Sh	Len	Dir	Pos	Dur	Avg	w/ Dur	w/o Dur	w/ Dur	w/o Dur
Itti-Koch [33]	1.886	1.293	0.482	2.910	-	1.643	-	4.467	-	4.086
CLE (Itti) [7, 33]	0.070	0.049	0.294	1.007	-	0.355	-	2.560	-	3.185
CLE (DG) [7, 40]	0.105	0.025	0.215	0.906	-	0.313	-	2.021	-	3.766
PathGAN [2]	0.070	0.108	0.575	2.148	0.199	0.620	3.504	3.155	2.061	1.960
G-Eymol [70]	1.531	0.782	0.238	2.159	0.324	1.007	14.068	7.125	9.468	3.341
DeepGazeIII [41]	0.058	0.025	0.143	0.200	-	0.107	-	0.333	-	2.465
ChenLSTM [13]	0.723	0.477	0.122	0.420	**0.026**	0.353	0.781	0.638	0.402	0.350
HAT [69]	2.793	1.248	0.207	2.236	-	1.621	-	3.371	-	1.548
TPP-Gaze [23]	0.070	0.070	0.085	0.288	0.067	0.116	1.058	0.648	0.779	0.365
`ScanDiff` (Ours)	**0.036**	**0.024**	**0.040**	**0.150**	0.035	**0.057**	**0.219**	**0.305**	**0.253**	**0.226**

Table 6. Performance comparison of different models on the OSIE [62] dataset for zero-shot prediction. Models with the highest performance for each metric is marked in **bold**.

	Target-Present					Target-Absent				
	MM ↓					MM ↓				
	Sh	Len	Dir	Pos	Dur	Sh	Len	Dir	Pos	Dur
PathGAN [2]	0.594	0.365	0.937	0.333	0.336	0.030	0.103	0.167	0.153	0.172
ChenLSTM [13]	0.253	0.276	0.337	0.054	0.066	0.075	0.044	0.111	0.052	0.092
Gazeformer [48]	0.581	0.301	0.316	0.056	0.150	0.067	0.047	0.054	0.044	0.233
HAT [69]	0.161	0.089	0.115	0.108	-	0.108	0.040	0.024	0.034	-
ChenLSTM-ISP [14]	0.272	0.232	0.302	0.019	0.044	0.114	0.087	0.132	0.016	0.060
GazeXplain [15]	0.238	0.255	0.245	0.038	0.052	0.019	0.016	0.015	0.051	0.128
TPP-Gaze [23]	0.676	0.250	0.845	0.825	0.025	0.051	0.017	0.117	0.287	0.018
Gazeformer [48]	0.436	0.326	0.313	0.031	0.147	0.081	0.418	0.269	1.613	0.250
GazeXplain [15]	0.244	0.224	0.295	0.030	0.045	0.009	0.023	0.024	0.035	0.095
TPP-Gaze [23]	0.445	0.254	0.581	1.216	0.037	0.023	0.032	0.052	0.547	**0.023**
`ScanDiff` (Ours)	**0.077**	**0.048**	**0.067**	**0.015**	**0.033**	**0.008**	**0.010**	**0.007**	**0.008**	0.067

Table 7. Additional results on COCO-Search18 dataset [16] for both target-present and target-absent settings. Models trained using identical settings and training splits to `ScanDiff` are highlighted in **gray**. Among these, the highest performance for each metric is marked in **bold**. Underlined values denote the top overall performance across all models and metrics.

Additional Ablation Studies. Previous diffusion-based approaches [36, 60] perform the conditioning by directly concatenating the noisy gaze sequence with the image embedding. On a different line, we condition the denoising process through the cross-attention layer. Results in Table 8 demonstrate that the proposed approach allows for a better semantic alignment between the scanpath and the multimodal features compared to the rigid concatenation of the input. Additionally, in Table 8, we also report an ablation study on the maximum scanpath length. We set this value to 16, matching the highest median across all datasets used in our experiments and the value used in [13, 15]. This hyperparameter serves as an upper bound, though the model can predict variable lengths.

		COCO-FreeView			COCO-Search18 (TP)			
	Len	MM ↓	SM ↓	SS ↓	MM ↓	SM ↓	SS ↓	SemSS ↓
w/o cross-attention	16	0.108	0.173	0.111	0.439	1.839	0.862	0.962
	32	**0.052**	0.204	0.113	0.084	0.069	0.030	**0.056**
	24	0.058	0.046	0.029	0.075	0.084	0.044	0.092
ScanDiff	16	0.078	**0.015**	**0.013**	**0.048**	**0.037**	**0.019**	0.072

Table 8. Ablation study on the effect of cross-attention compared to input concatenation and on the maximum scanpath length.

B. Additional Qualitative Results

Additional qualitative results are depicted from Fig. 4 to Fig. 7 on COCO-FreeView [66], MIT1003 [37], and OSIE [62] for free-viewing and COCO-Search18 [16] for the visual search task, respectively. The qualitative results support the findings of the main paper, highlighting the accuracy of ScanDiff in predicting human-like scanpaths. Other methods, however, demonstrate limitations by either focusing excessively on specific elements or producing shorter scanpaths than the ones exhibited by humans.

The qualitative results further support the findings of our scanpath variability analysis, highlighting the ability of ScanDiff to generate diverse yet human-like scanpaths across different viewing tasks. As shown in the comparison evaluation from Fig. 9 to Fig. 12 existing methods often produce scanpaths that are either overly deterministic – failing to capture the natural variability of human gaze behavior – or overly stochastic, resulting in implausible trajectories. This is particularly true for Gazeformer that, as shown in Fig. 11 and Fig. 12, produces scanpaths that are identical to each other over the simulations, essentially generating the same fixation pattern repeatedly regardless of the inherent variability present in human visual attention processes. While Gazeformer achieves reasonable performance on task-oriented datasets as shown in the quantitative results, its deterministic nature fundamentally limits its ability to model the stochastic aspects of human gaze behavior that our approach successfully captures.

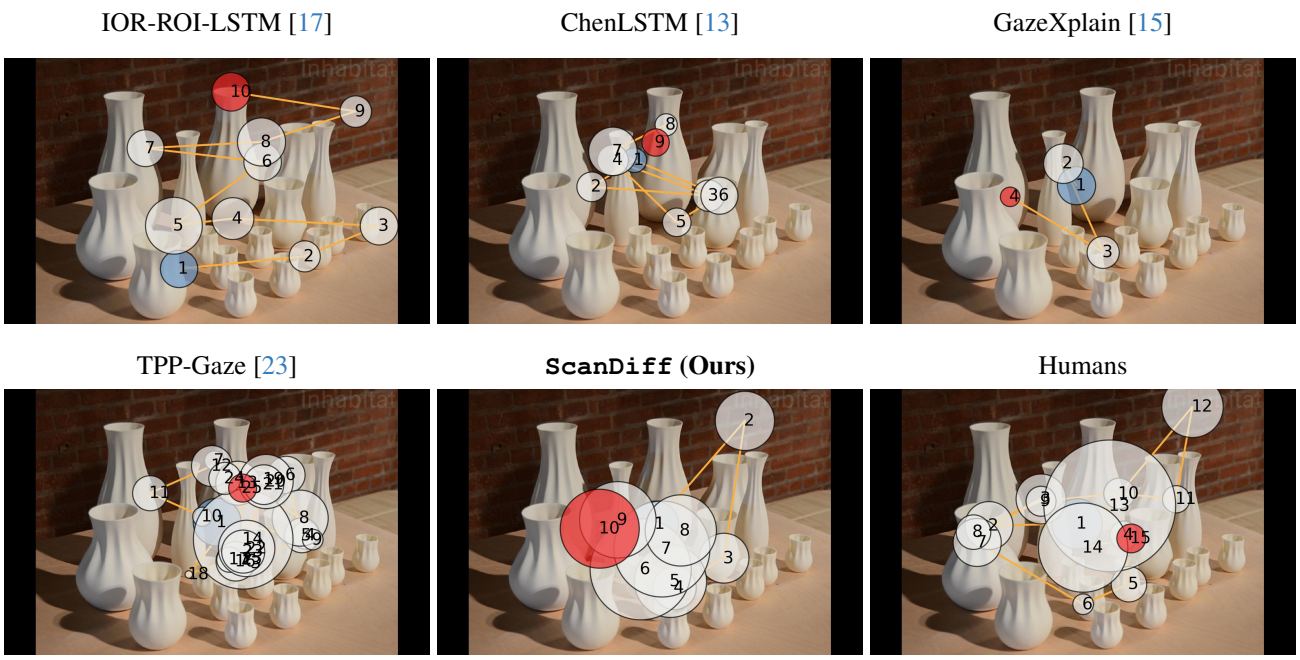

Figure 4. Qualitative comparison of simulated and human scanpaths on the COCO-FreeView dataset.

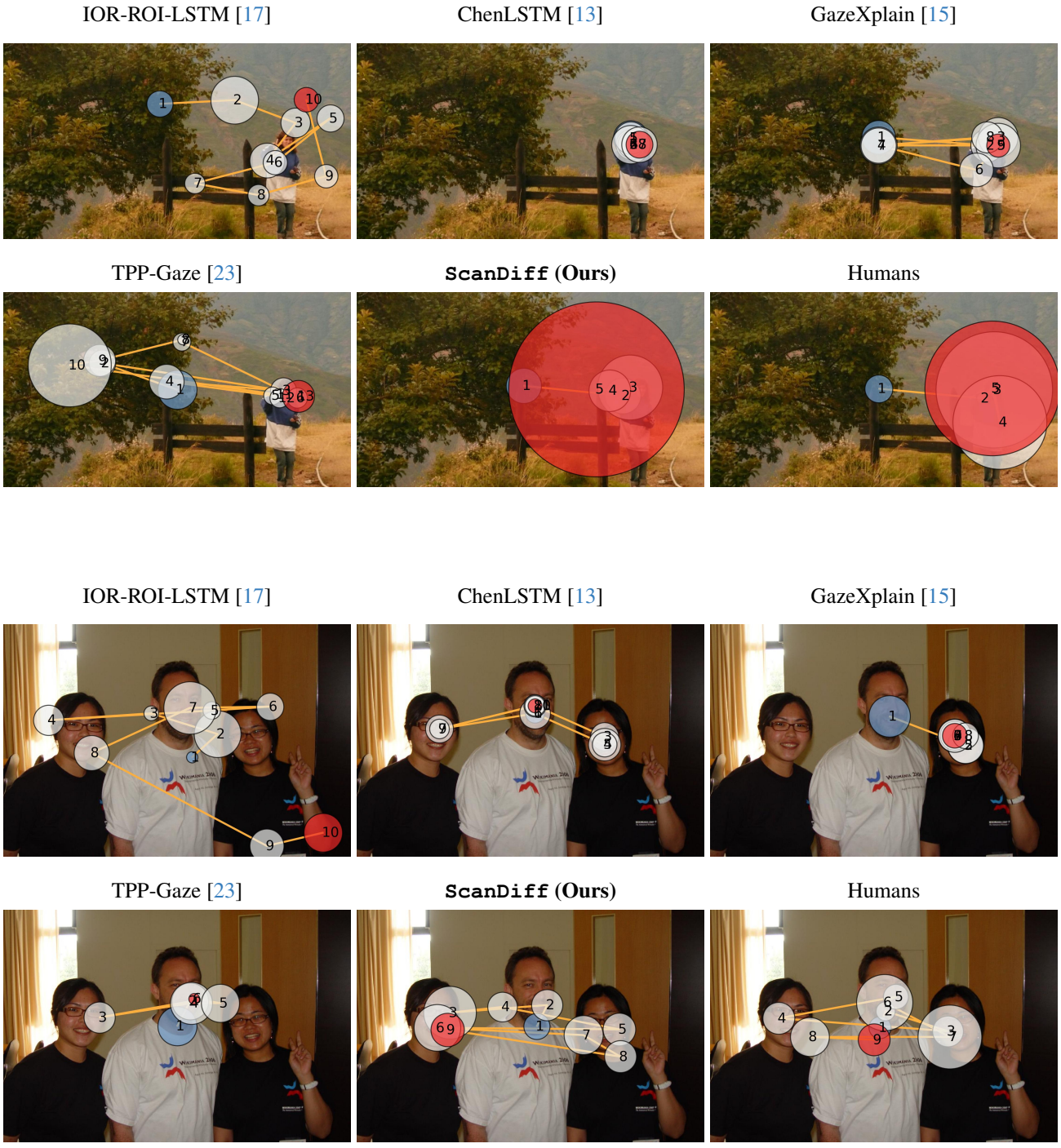

Figure 5. Qualitative comparison of simulated and human scanpaths on the MIT1003 dataset.

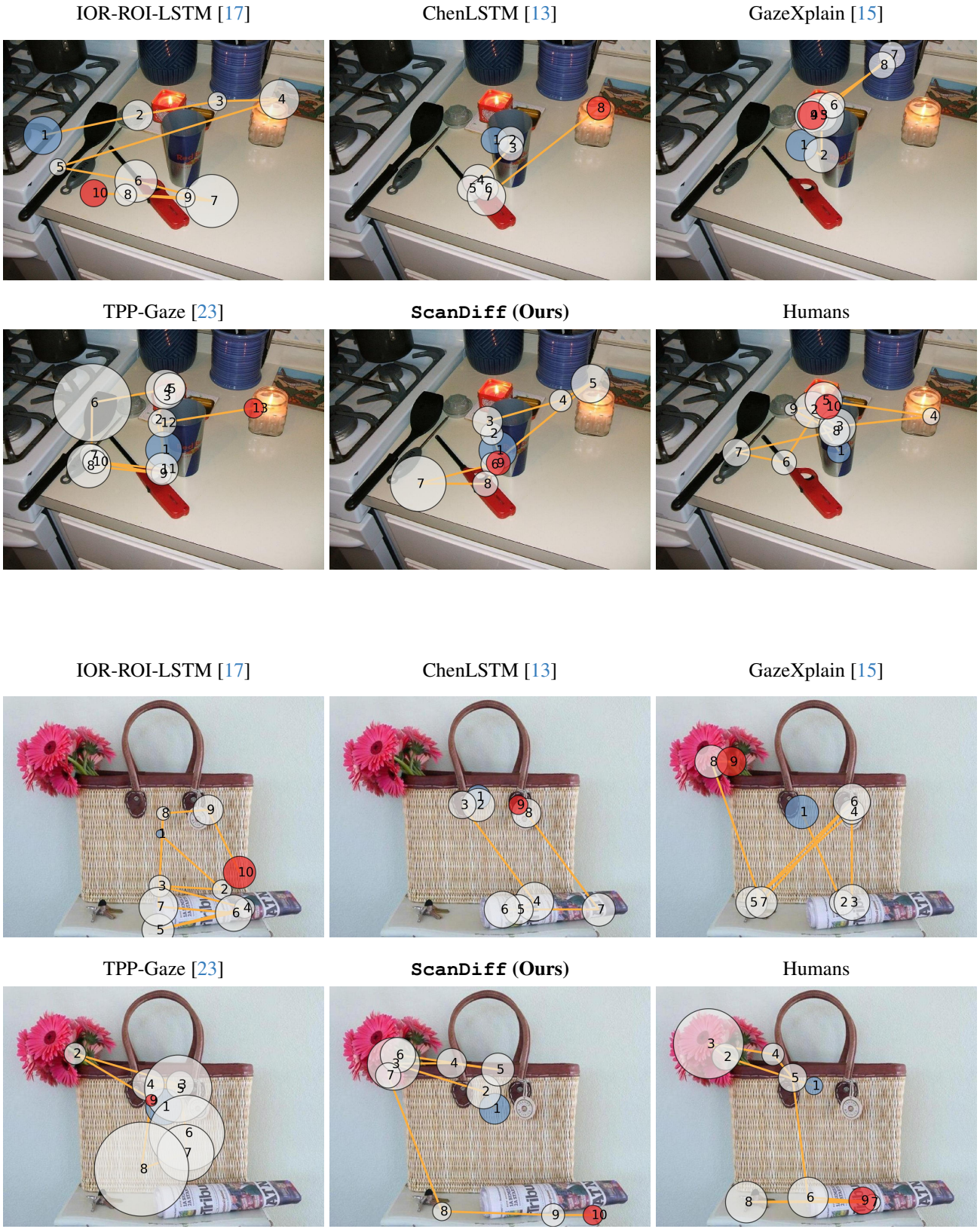

Figure 6. Qualitative comparison of simulated and human scanpaths on the OSIE dataset.

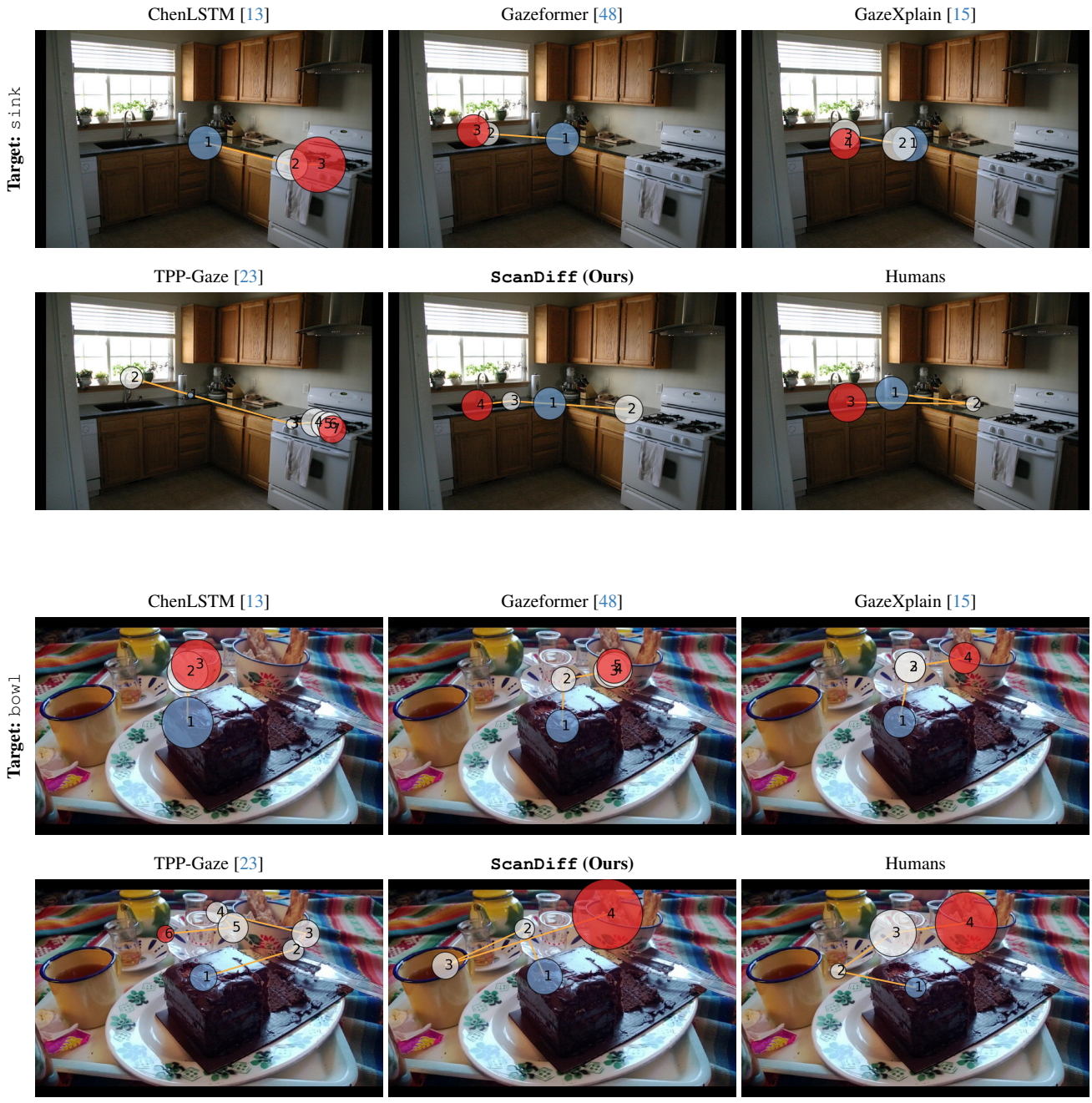

Figure 7. Qualitative comparison of simulated and human scanpaths on the COCO-Search18 (TP) dataset for the visual search task.

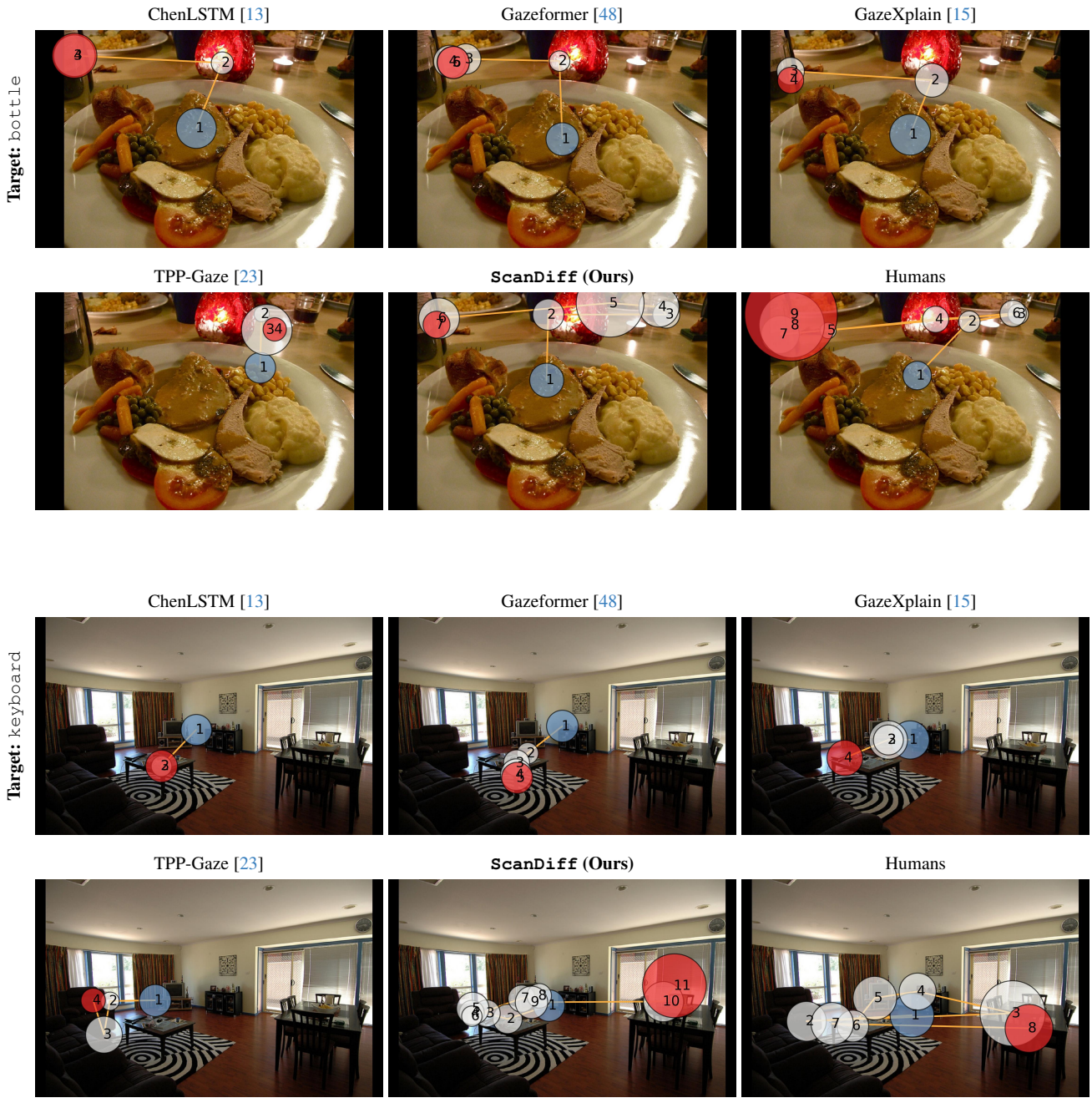

Figure 8. Qualitative comparison of simulated and human scanpaths on the COCO-Search18 (TA) dataset for the visual search task.

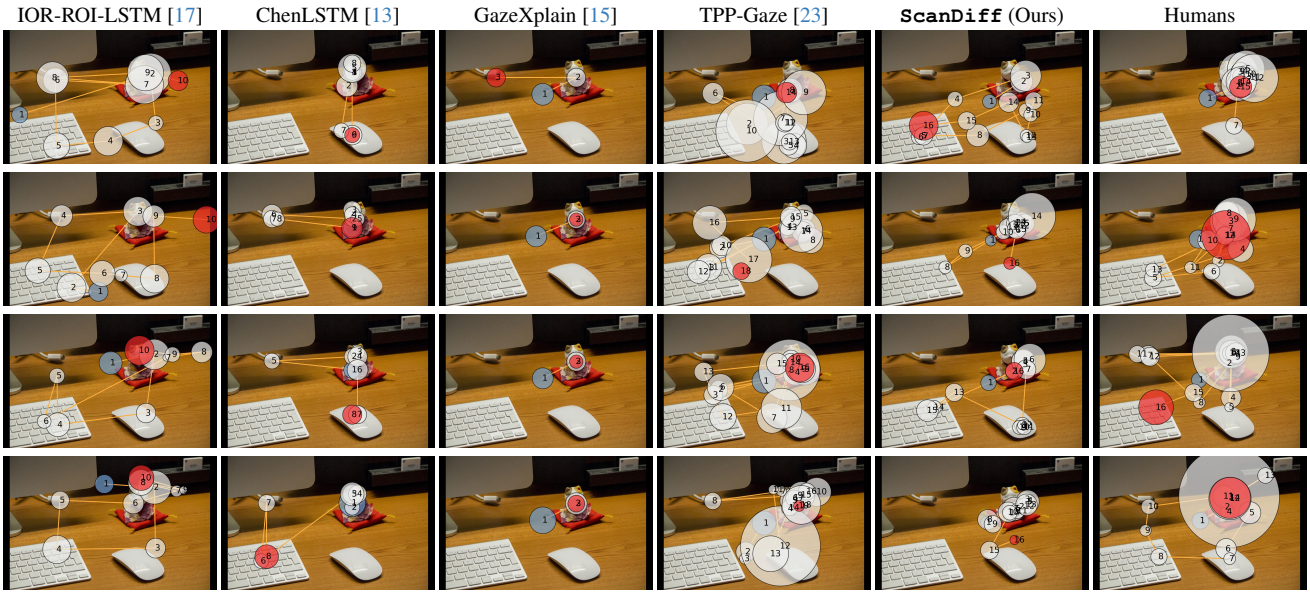

Figure 9. Qualitative comparison of the variability in simulated and human scanpaths on the COCOFreeView dataset. Each row corresponds to a different simulation or a different human observer.

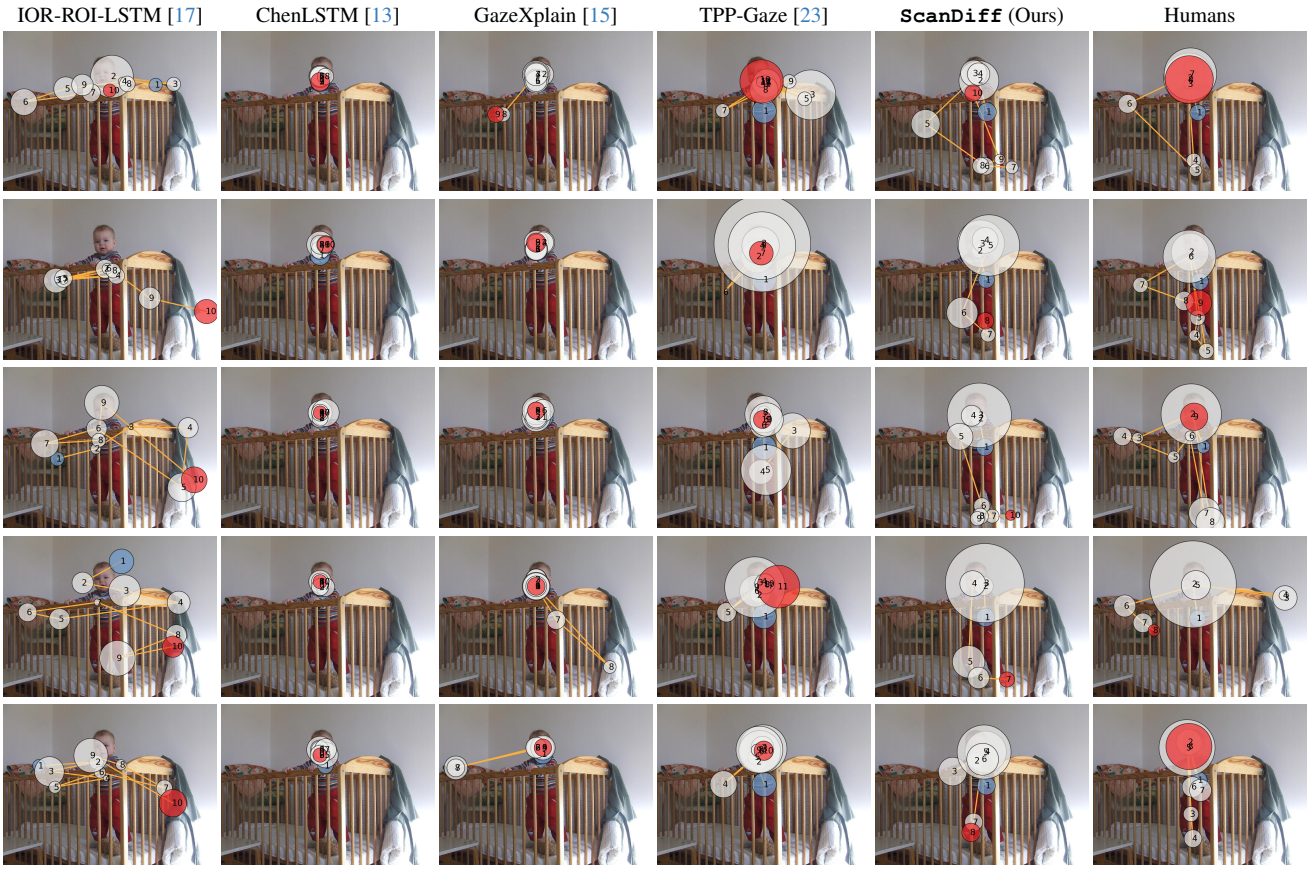

Figure 10. Qualitative comparison of the variability in simulated and human scanpaths on the MIT1003 dataset. Each row corresponds to a different simulation or a different human observer.

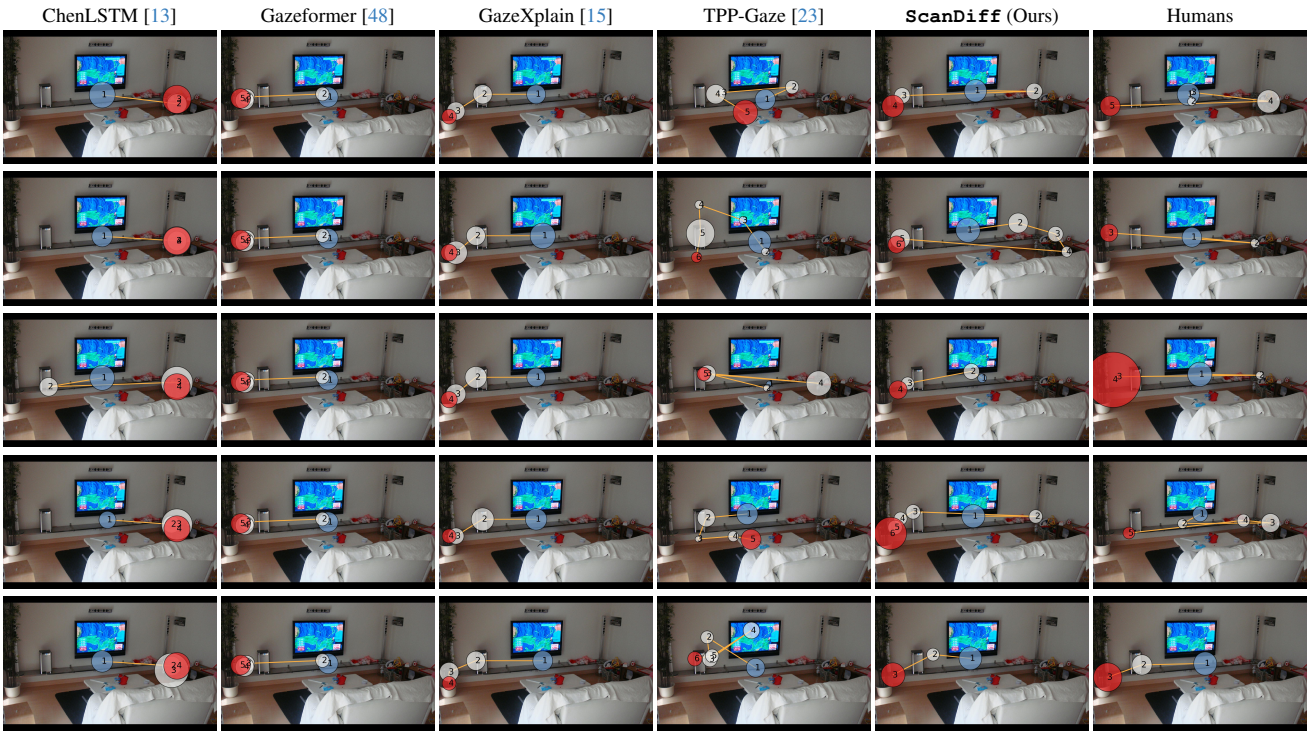

Figure 11. Qualitative comparison of the variability of simulated and human scanpaths on the COCO-Search18 (TP) dataset for the viewing task: `potted plant`. Each row corresponds to a different simulation or a different human observer.

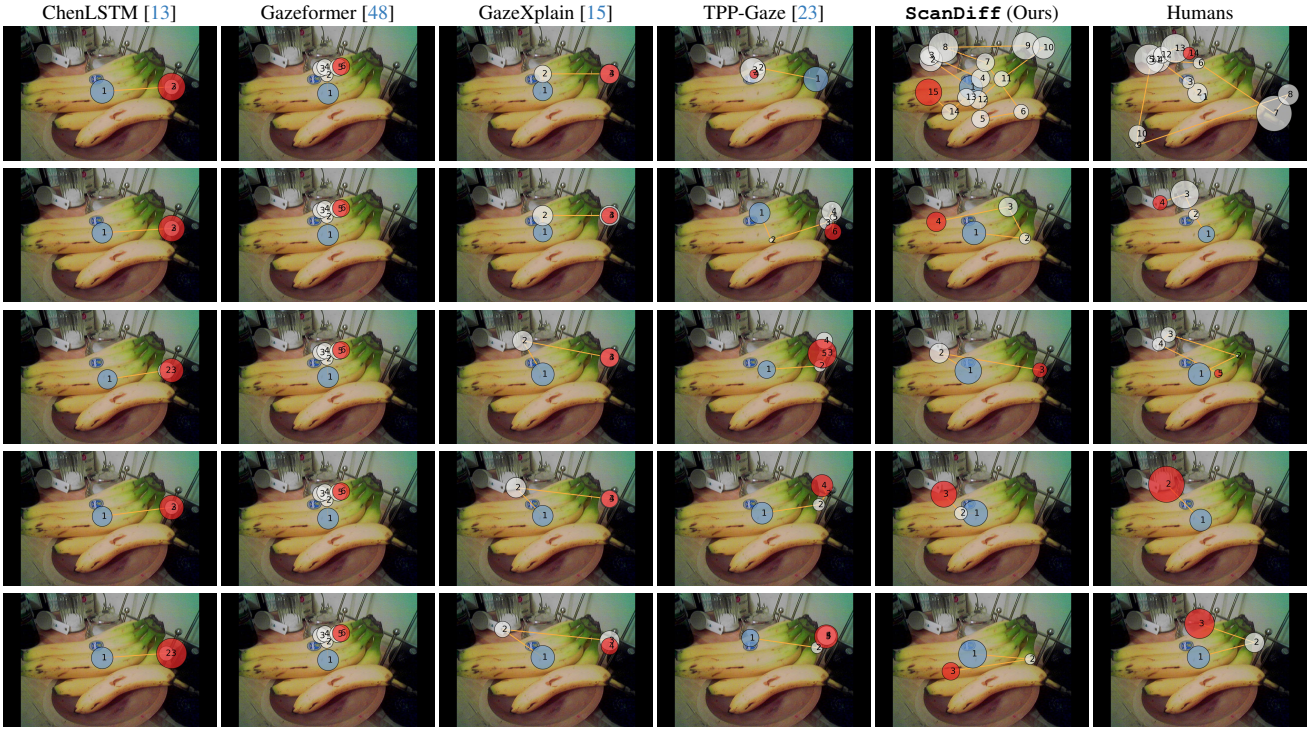

Figure 12. Qualitative comparison of the variability of simulated and human scanpaths on the COCO-Search18 (TA) dataset for the viewing task: `fork`. Each row corresponds to a different simulation or a different human observer.