Beyond Average: Individualized Visual Scanpath Prediction

Supplementary Material

6. Introduction

In the main paper, we have introduced a new method for predicting individual scanpaths (ISP) with three novel components, *i.e.*, observer encoding, observer-centric feature integration, and adaptive fixation prioritization, which can accurately model saccadic eye movements in various tasks, such as free-viewing, visual search, and visual question answering. The experimental results have demonstrated that our method performs competitively and is highly generalizable. This supplementary material supports our main findings with further results and introduces additional implementation details of the proposed method:

1) Section 7 presents detailed descriptions of how ISP is applied to the Gazeformer architecture and its adaptation for predicting scanpaths in different tasks.
2) Section 8 presents supplementary ablation studies on three other eye-tracking datasets (*i.e.*, OSIE [79], COCO-Search18 [83] and AiR-D [12]) to demonstrate the effectiveness of the three novel components of our method. Furthermore, we conduct experiments on the OSIE-ASD [71] dataset to investigate the impacts of hyperparameters used in adaptive fixation prioritization. Lastly, we present supplementary experiments to investigate the generalizability of our ISP models on new subjects.
3) Section 9 presents additional quantitative results on the saliency map evaluation, comparing our ISP models with different state-of-the-art scanpath prediction approaches and baselines.
4) Section 10 presents a quantitative comparison with new baseline models conditioned on the one-hot observer identity, along with statistical analyses to demonstrate our ISP models' individualization ability.
5) Section 11 presents additional qualitative results, comparing our method with state-of-the-art scanpath prediction approaches. These results highlight the superior performance of our method across three datasets: OSIE-ASD [79] (free-viewing), COCO-Search18 [83] (visual search), and AiR-D [12] (VQA), confirming its adaptability to diverse real-life visual tasks.

7. Implementation Details of Gazeformer-ISP

We have implemented our ISP on two baseline architectures, ChenLSTM and Gazeformer. The ISP method designed for ChenLSTM has been introduced in Section 3 of the main paper. In this section, we elaborate detailed implementation of ISP on the Gazeformer model. There are three distinctions compared with the ChenLSTM architecture:

Task Encoder. Different from ChenLSTM that uses the machine attention of a VQA model [12] or object detection outputs to initially guide the first fixation to direct the eye movement based on the given visual task, we first extract the target feature or question feature using the language model RoBERTa [48] as $\boldsymbol{v}_\ell$. To model the interaction between the visual feature $\boldsymbol{E}$ and the language feature $\boldsymbol{v}_\ell$, a task guidance map can be computed through a linear combination:

$$\boldsymbol{m}_0 = \mathrm{softmax}\big(\boldsymbol{w}_{e\ell}^T \tanh(\boldsymbol{W}_{e\ell}\boldsymbol{E} + \boldsymbol{W}_{m\ell}\boldsymbol{v}_\ell)\big), \quad (10)$$

where $\boldsymbol{w}_{e\ell}$, $\boldsymbol{W}_{e\ell}$, $\boldsymbol{W}_{m\ell}$ are learnable parameters. This observer guidance localizes salient image regions of specific interest to the observer.

Observer-Centric Feature Integration. Different from the LSTM architecture, Transformers process input sequences in parallel, relying on positional encodings to impart positional information. As a result, the temporal subscript t becomes less relevant in the Transformer context. Therefore, we always set $\boldsymbol{m}_{t-1}$ as $\boldsymbol{m}_0$ from Equation (10) and eliminate all the t in the subscript of each variables, such as $\boldsymbol{X}_t$, $\boldsymbol{X}_{ut}$, $\boldsymbol{u}_t^s$, $\boldsymbol{u}_t^c$ and $\boldsymbol{R}_t$. This simplification maintains consistency within the Transformer architecture, facilitating efficient parallelization and computation.

Adaptive Fixation Prioritization. The original Gazeformer [51] model comprises two variants: the Gazeformer-Reg produces fixation coordinates and is trained using a regression objective, while the Gazeformer-noReg generates a fixation map and is trained through grid-based classification. Our ISP is based on the latter implementation by applying the proposed adaptive fixation prioritization model, using a cross-attention mechanism [20, 51] to compute the semantic feature maps $\boldsymbol{A}_t$.

8. Supplementary Ablation Studies

In this section, we conduct comprehensive ablation studies on the OSIE [79], COCO-Search18 [83] and AiR-D [12] datasets, as well as the ablation study of the number of output feature channels for the proposed adaptive fixation prioritization. Lastly, we discuss how the ISP models can generalize on new subjects and how much gaze data of the new subjects is needed to achieve comparable performance.

8.1. Ablation Study on OSIE, COCO-Search18 and AiR-D Datasets

In Table 3 of the main paper, we have presented a comprehensive ablation study on the OSIE-ASD [71] dataset to

Modules			ChenLSTM						Gazeformer					
OE	FI	FP	SM ↑	MM ↑	SED ↓	MRR ↑	R@1 ↑	R@5 ↑	SM ↑	MM ↑	SED ↓	MRR ↑	R@1 ↑	R@5 ↑
			0.373	0.804	7.309	0.222	7.048	32.952	0.372	0.809	7.298	0.223	7.048	32.476
✓	✓		0.376	0.806	**7.271**	0.282	11.333	43.143	0.389	0.810	7.164	0.264	9.619	41.238
✓		✓	0.377	0.807	7.299	0.229	7.238	33.143	0.384	0.810	7.186	0.241	7.905	37.524
✓	✓	✓	**0.377**	**0.810**	7.284	**0.291**	**12.667**	**44.095**	**0.390**	**0.813**	**7.163**	**0.268**	**10.095**	**41.905**

Table 5. Ablation study for the proposed technical components: observer encoder (OE), observer-centric feature integration (FI) and adaptive fixation prioritization (FP) for OSIE [79] dataset.

Modules			ChenLSTM						Gazeformer					
OE	FI	FP	SM ↑	MM ↑	SED ↓	MRR ↑	R@1 ↑	R@5 ↑	SM ↑	MM ↑	SED ↓	MRR ↑	R@1 ↑	R@5 ↑
			0.454	0.799	1.932	0.296	10.199	50.719	0.432	0.796	2.023	0.292	9.873	50.114
✓	✓		0.471	0.809	1.896	0.365	15.266	61.360	0.450	0.804	2.021	0.347	14.678	59.170
✓		✓	0.474	0.807	1.871	0.329	12.994	55.574	0.450	0.799	2.014	0.321	12.096	55.149
✓	✓	✓	**0.480**	**0.811**	**1.862**	**0.369**	**16.639**	**61.769**	**0.455**	**0.806**	**1.997**	**0.353**	**15.299**	**60.020**

Table 6. Ablation study for the proposed technical components: observer encoder (OE), observer-centric feature integration (FI) and adaptive fixation prioritization (FP) for COCO-Search18 [83] dataset.

evaluate the significance of the proposed technical components: object encoder (OE), observer-centric feature integration (FI), and adaptive fixation prioritization (FP). In this section, to demonstrate the generalizability of our proposed ISP, we further provide more insights on the effectiveness of these proposed technical components in the other eye-tracking datasets (*i.e.*, OSIE [79], COCO-Search18 [83] and AiR-D [12]), see Tables 5, 6 and 7. Similar to Table 3 demonstrating that the FI and FP play a complementary role in leading to the most significant overall performance improvements on different neural network architectures, These supplementary results emphasize the consistent importance of these technical components across diverse scenarios and tasks. This ablation study reinforces the robustness and effectiveness of these components in our proposed ISP method.

8.2. Ablation Study of Output Feature Maps

As outlined in the main paper, our adaptive fixation prioritization dynamically integrates a number of output feature maps for predicting the next fixation position. Thus, the number of feature maps, denoted as L, is cruicial to the prediction performance. Renowned for its extensive participant pool and diverse demographics, OSIE-ASD [71] provides an ideal testbed for assessing individualized visual scanpath prediction due to its capacity to differentiate among various individuals. Table 8 reports model performances across diverse configurations on the OSIE-ASD [71] dataset. On the one hand, a lower L (*e.g.*, $L = 1$ or $L = 2$) leads to feature maps with limited semantic variations, making it challenging to discern distinct individual fixation patterns, resulting in suboptimal outcomes for individual scanpath prediction. On the other hand, a higher L introduces higher dimensionality, leading to overfitting on the training fixation data (*e.g.*, $L = 6$), causing a notable decline in evaluation metric scores. Setting $L = 4$ results in a reasonable balance, facilitating the learning of diverse fixation semantic feature maps.

8.3. Cross-Subject Generalizability

To explore the generalizability of our ISP models on new subjects, we conduct fine-tuning experiments. ISP models pre-trained on the OSIE-ASD [71] dataset are fine-tuned on the OSIE [79] dataset using varying amounts of gaze data ($N = 20,\ 50,\ 100,\ 200,\ 560$ samples per subject). These fine-tuned models are compared to models trained entirely on OSIE data (*i.e.*, Full OSIE). As shown in Table 9, even with a small number of training samples of new subjects (*e.g.*, $N = 100$), the evaluation results closely resemble those of the Full OSIE model (*e.g.*, Gazeformer-ISP, SM=0.372 on 100 images compared to SM=0.390 from Full OSIE models), demonstrating that models pre-trained on different subjects provide common knowledge for generalizing to new ones. On the other hand, with an increase in the number of images, performance gradually improves and approaches that of the Full OSIE model, indicating that increasing the data can enhance the model's capabilities.

9. Saliency Evaluations

In this section, we investigate the spatial accuracy of predicted fixations with a comprehensive saliency comparison across a variety of state-of-the-art models and baselines on the OSIE-ASD [71] dataset.

Supplementing the individualized scanpath evaluation

Modules			ChenLSTM						Gazeformer					
OE	FI	FP	SM ↑	MM ↑	SED ↓	MRR ↑	R@1 ↑	R@5 ↑	SM ↑	MM ↑	SED ↓	MRR ↑	R@1 ↑	R@5 ↑
			0.356	0.808	7.845	0.297	9.957	51.433	0.349	0.810	8.004	0.299	10.459	51.361
✓	✓		0.357	0.808	7.824	0.330	12.607	57.020	0.353	0.814	7.992	0.316	12.536	53.224
✓		✓	0.370	0.812	7.768	0.306	11.175	51.146	0.355	0.811	7.956	0.299	10.888	51.576
✓	✓	✓	**0.371**	**0.813**	**7.651**	**0.338**	**13.610**	**57.235**	**0.362**	**0.814**	**7.911**	**0.334**	**13.539**	**57.450**

Table 7. Ablation study for the proposed technical components: observer encoder (OE), observer-centric feature integration (FI) and adaptive fixation prioritization (FP) for AiR-D [12] dataset.

	ChenLSTM						Gazeformer					
L	SM ↑	MM ↑	SED ↓	MRR ↑	R@1 ↑	R@5 ↑	SM ↑	MM ↑	SED ↓	MRR ↑	R@1 ↑	R@5 ↑
1	0.389	0.795	7.064	0.122	3.150	15.238	0.398	0.796	6.982	0.134	3.810	17.509
2	0.393	0.796	6.885	0.147	4.835	**19.414**	0.406	**0.798**	6.992	0.138	3.626	18.608
4	**0.401**	**0.798**	**6.599**	**0.147**	**4.835**	19.194	**0.406**	0.797	**6.823**	**0.141**	**4.286**	**18.571**
6	0.385	0.796	7.043	0.140	3.883	18.352	0.400	0.793	6.849	0.135	3.626	18.242

Table 8. Ablation study of different values of hyperparameter L on the OSIE-ASD [71] dataset. We select the best hyperparameter based on the ScanMatch scores.

results in Table 1, we evaluate the spatial accuracy of predicted fixations through population-level saliency maps. In particular, we aggregate the fixations from the generated scanpaths of all observers, and post-process the fixations to obtain smoothed saliency maps [68]. This analysis compares ISP to state-of-the-art methods and baselines on their ability to capture the overall distribution of fixations across observers. Besides the models presented in Table 1, we also add the Le Meur *et al*. [50], Li *et al*. [46], and DeepGaze III [44] models into the comparison. Table 10 presents the evaluation results. Notably, ISP models outperform both baselines and the previous state-of-the-art method (with the highest average ranking $\bar{R}$) on the OSIE-ASD [71] dataset. This notable improvement confirms the effectiveness of our proposed ISP method in capturing the overall prediction of fixation distributions for the population.

10. Supplementary Baseline Models

In this section, we introduce supplementary baseline models in addition to the fine-tuned (FT) baseline in the main paper. We present a comprehensive comparison of value-based and ranking-based evaluations with the state-of-the-art methods and the baselines on the OSIE-ASD [71] dataset. We also conduct a comprehensive statistical analysis of the attention traits at the population level among the baselines.

10.1. Implementation Details

While fine-tuning the general models on individual observer data (*i.e*., ChenLSTM-FT, Gazeformer-FT) provides a baseline for assessing the impact of explicitly incorporating observer-specific characteristics, we additionally incorporate supplementary baseline models (*i.e*., ChenLSTM-onehot, Gazeformer-onehot) with the scanpath prediction conditioned on the one-hot observer identity (*e.g*., concatenating one-hot encoding of the identity with the feature maps to the scanpath decoder). On the one hand, for the fine-tuned models, we fine-tune the pre-trained model on the population gaze data with a reduced learning rate 10^{-5} in 2 epochs. On the other hand, for the one-hot baseline models, we train the model similarly to the ISP models mentioned in the main paper, with 15 epochs of supervised learning and 10 epochs of self-critical sequence training (SCST) [14, 60] with learning rate 10^{-4}.

10.2. Quantitative Results

In Table 1 and Table 2 of the main paper, we have conducted a comprehensive comparison of value-based and ranking-based evaluations on the OSIE [79], OSIE-ASD [71] COCO-Search [83] and AiR-D [12] datasets to demonstrate the effectiveness of the proposed ISP. Here, we further include the performance on the one-hot baseline models on OSIE-ASD [71] dataset in Table 11. Interestingly, the one-hot baseline outperforms the fine-tuned Gazeformer model (Gazeformer-FT) but underperforms the fine-tuned ChenLSTM model (ChenLSTM-FT), yet both baseline methods underperform the proposed ISP approach. Overall, these results highlight ISP's ability to achieve robust and effective individualization by explicitly capturing observer-specific attention patterns.

10.3. Population-Level Comparison

We extend the statistical comparison between the predicted fixations for the ASD and Control groups by incorporating one-hot baseline models (ChenLSTM-onehot, Gazeformer-

	ChenLSTM						Gazeformer					
Num. of Img.	SM ↑	MM ↑	SED ↓	MRR ↑	R@1 ↑	R@5 ↑	SM ↑	MM ↑	SED ↓	MRR ↑	R@1 ↑	R@5 ↑
Full OSIE	**0.377**	**0.810**	7.284	**0.291**	**12.667**	44.095	**0.390**	**0.813**	**7.163**	0.268	10.095	41.905
20 images	0.344	0.800	7.571	0.240	7.333	39.238	0.369	0.809	7.583	0.253	8.857	40.191
50 images	0.358	0.805	7.409	0.259	9.143	26.191	0.365	0.809	7.370	0.264	10.571	40.667
100 images	0.361	0.805	7.257	0.266	10.095	40.667	0.372	0.810	7.268	0.274	11.333	40.667
200 images	0.365	0.807	7.260	0.270	10.095	42.286	0.374	**0.813**	7.233	0.271	10.286	43.429
560 images	0.373	0.806	**7.218**	0.279	10.762	**44.476**	0.376	0.811	7.325	**0.290**	**12.571**	**45.714**

Table 9. Ablation study of the number of gaze data from new subjects.

Method	CC ↑	AUC ↑	NSS ↑	sAUC ↑	KLD ↓	SIM ↑	$\bar{R}$ ↓
SaltiNet [2]	0.224	0.661	0.541	0.587	1.496	0.414	10.33
PathGAN [3]	0.239	0.599	0.560	0.505	10.703	0.225	11.67
IOR-ROI [69]	0.572	0.803	1.562	0.669	1.841	0.534	9.00
Le Meur *et al*. [50]	0.621	0.819	1.768	**0.752**	0.702	0.542	5.42
Li *et al*. [46]	0.538	0.854	1.767	0.686	1.409	0.542	6.92
DeepGaze III [44]	0.786	**0.877**	2.050	0.749	**0.359**	**0.694**	2.33
ChenLSTM [14]	0.728	0.810	2.348	0.720	3.310	0.549	6.83
Gazeformer [51]	0.689	0.790	2.179	0.708	4.135	0.520	8.67
ChenLSTM-FT	0.782	0.830	2.495	0.736	2.151	0.599	4.33
Gazeformer-FT	0.733	0.816	2.303	0.721	2.762	0.565	6.17
ChenLSTM-ISP	**0.807**	0.855	**2.529**	0.748	1.442	0.642	**2.17**
Gazeformer-ISP	0.779	0.842	2.396	0.733	1.771	0.620	4.17

Table 10. Comparison of saliency evaluation results of state-of-the-art models and baselines. $\bar{R}$ indicates the average ranking across all the saliency evaluations.

onehot) into the analysis alongside ISP models and ground truth fixations (Figure 7). As shown in Figure 7, the one-hot baselines exhibit significant limitations. ChenLSTM-onehot completely fails to differentiate gaze statistics between subject groups, resulting in no statistically significant differences. Gazeformer-onehot, while partially successful, misses four key statistical differences and even produces one incorrect result, suggesting a higher fixation proportion for individuals with ASD on social semantics, which is a finding contradicted by the data. In contrast, both ChenLSTM-ISP and Gazeformer-ISP models closely resemble the ground truth, accurately capturing the population-level gaze patterns. This success demonstrates the effectiveness of ISP methods in generalizing and representing population-level characteristics.

11. Supplementary Qualitative Results

In addition to the qualitative examples shown in Figure 3 of our main paper, we offer an expanded exploration of the qualitative results derived from our ISP method. This extended presentation involves a thorough comparison among ISP models, fine-tuned models, and human ground truth, covering a spectrum of visual tasks based on the OSIE-ASD [71], COCO-Search18 [83], and AiR-D [12]. Our ISP models consistently align accurately with the scanpaths of individual observers. These qualitative examples demonstrate their potential as a promising and interpretable tool in unraveling visual perception and decision-making processes.

OSIE-ASD. Figure 8 illustrates a qualitative comparison on the OSIE-ASD [71] dataset. Examples (a), (c), and (e) show the scanpaths of observers with ASD, while examples (b), (d), and (f) show the scanpaths of controls. In examples (a) and (b), both Gazeformer-ISP and ChenLSTM-ISP effectively distinguish observers with ASD (repeatedly focusing on the center) from the control (exploring the dog and two children). Similarly, in panels (c) and (d), both models discern observers with ASD (repeatedly focusing on the center or the same object for an extended period) from the control (exploring the boys and ball). The same distinction is observed in panels (e) and (f). In contrast, fine-tuned baselines like Gazeformer-FT and ChenLSTM-FT lack this ability to discern such distinctions in gazing patterns across observers.

COCO-Search18. Figure 9 presents a qualitative comparison on the COCO-Search18 [83] dataset. In examples (a) and (b), observers search for the target 'potted plant' with different patterns. Gazeformer-ISP and ChenLSTM-

Method	SM ↑	MM ↑	SED ↓	MRR ↑	R@1 ↑	R@5 ↑
Human	0.370	0.783	7.720	-	-	-
SaltiNet [2]	0.137	0.735	8.688	0.107	2.454	12.454
PathGAN [3]	0.042	0.732	9.342	0.110	2.601	12.894
IOR-ROI [69]	0.301	0.788	7.655	0.109	2.784	12.454
ChenLSTM [14]	0.341	0.791	7.602	0.108	2.418	13.114
Gazeformer [51]	0.388	0.792	7.081	0.107	2.564	11.758
ChenLSTM-FT	0.394	0.796	7.067	0.113	2.711	12.637
Gazeformer-FT	0.387	0.795	7.083	0.108	2.528	13.223
ChenLSTM-onehot	0.366	0.785	7.291	0.106	2.491	11.722
Gazeformer-onehot	0.395	0.797	7.006	0.116	2.894	14.029
ChenLSTM-ISP	0.401	**0.798**	**6.599**	**0.147**	**4.835**	**19.194**
Gazeformer-ISP	**0.406**	0.797	6.823	0.141	4.286	18.571

Table 11. Comparison of value-based evaluation results for models' ability to predict the scanpaths of individual observers and ranking-based evaluation results for models' ability to distinguish different observers.

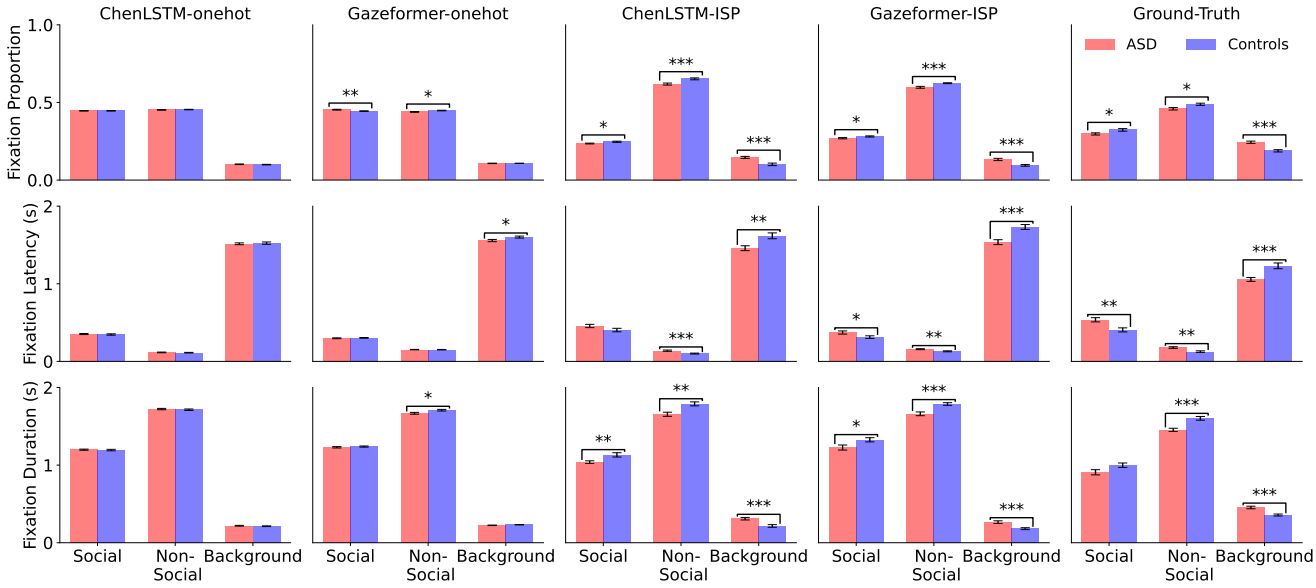

Figure 7. Statistical comparison between the predicted fixations for the ASD and Control groups [71] with the baseline. Error bars indicate the standard error of the mean. Asterisks indicate significant differences (unpaired t-test, *: $p < 0.05$, **: $p < 0.01$, and ***: $p < 0.001$).

ISP reveal that observer (a) successfully finds the plant, while observer (b) fails to find it. Similarly, in examples (c) and (d), observers search for the target 'car' with different patterns. Gazeformer-ISP and ChenLSTM-ISP show that observer (c) successfully finds the car, while observer (d) fails, by recognizing the vehicle on the right-hand side of the image as a car. Lastly, in examples (e) and (f), observers search for the target 'TV' with different patterns. Gazeformer-ISP and ChenLSTM-ISP demonstrate that observer (e) successfully finds the TV, while observer (f) fails, fixating around the table. Fine-tuning methods Gazeformer-FT and ChenLSTM-FT do not clearly indicate success or failure in finding the target object.

AiR-D. Figure 10 offers a qualitative comparison on the AiR-D [12] dataset. In examples (a) and (b), two observers respond differently to the question 'What is the device on top of the nightstand made of wood?' Gazeformer-ISP and ChenLSTM-ISP reveal that observer (a) correctly identifies the answer 'phone' by focusing on the nightstand, while observer (b) fails due to attention on the table, not the nightstand. Similarly, in examples (c) and (d), observers answer 'Is the vase the same color as the scarf?' differently. Gazeformer-ISP and ChenLSTM-ISP show that observer (c) correctly answers 'no' by examining colors, while observer (d) fails to find the scarf. Lastly, in examples (e) and (f), observers respond differently to 'What is the ap-

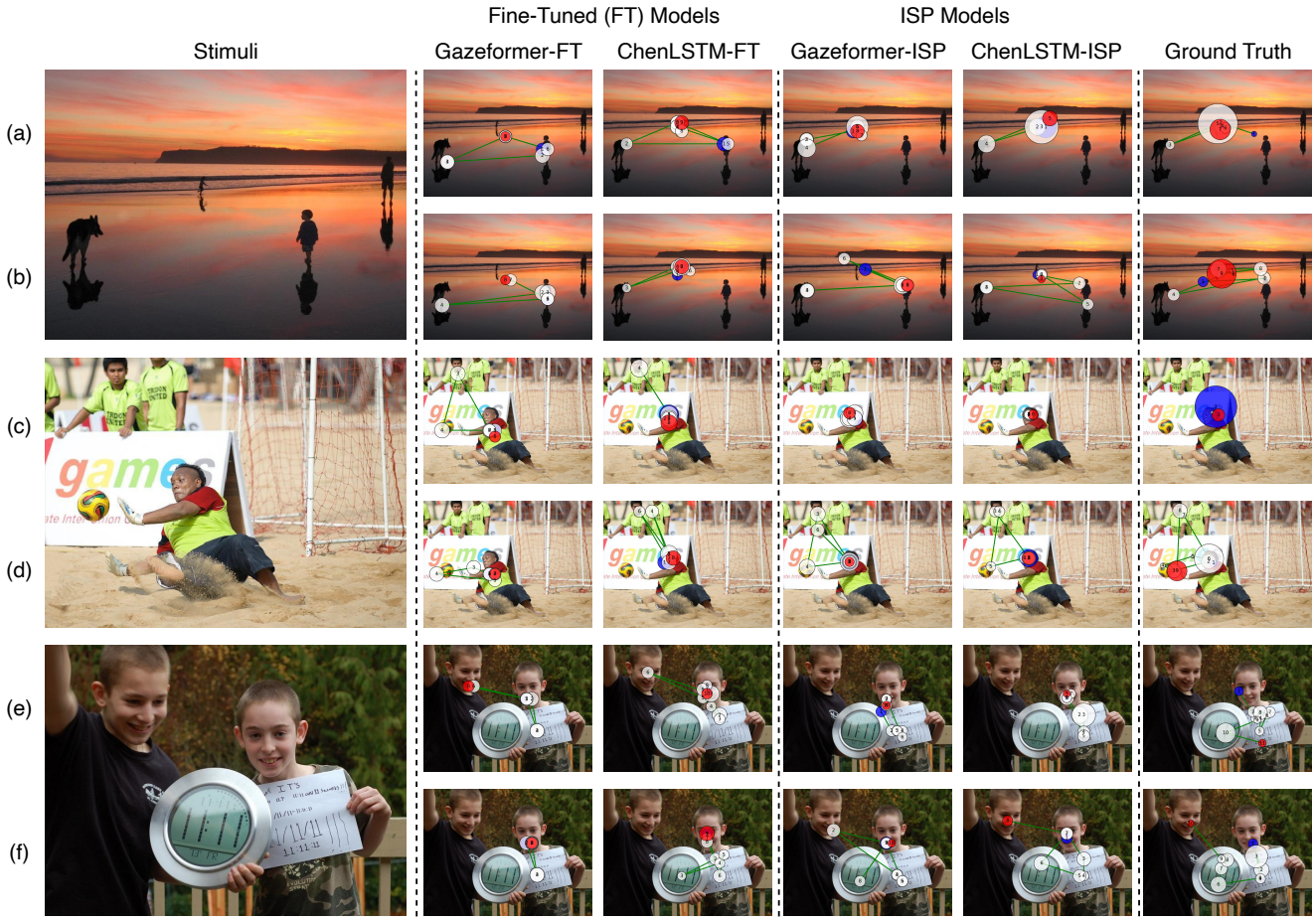

Figure 8. Qualitative scanpath examples from Gazeformer-FT, ChenLSTM-FT, Gazeformer-ISP, ChenLSTM-ISP, and ground truth on OSIE-ASD [71]. The first column shows the stimuli, and each row compares model predictions with one observer's ground truth. Observers with ASD exhibit atypical gaze patterns: (a, c) prolonged fixation on central objects, (e) local exploration, while the controls (b, d, f) explore diverse people and objects. Blue and red dots mark the start and end of scanpaths, respectively.

pliance in the kitchen?' Gazeformer-ISP and ChenLSTM-ISP show observer (e) correctly finding the 'microwave' in the kitchen, while observer (f) fails, focusing mostly on the living room. Fine-tuned baselines like Gazeformer-FT and ChenLSTM-FT cannot explain reasons for observers' correct or incorrect answers based on predicted scanpaths.

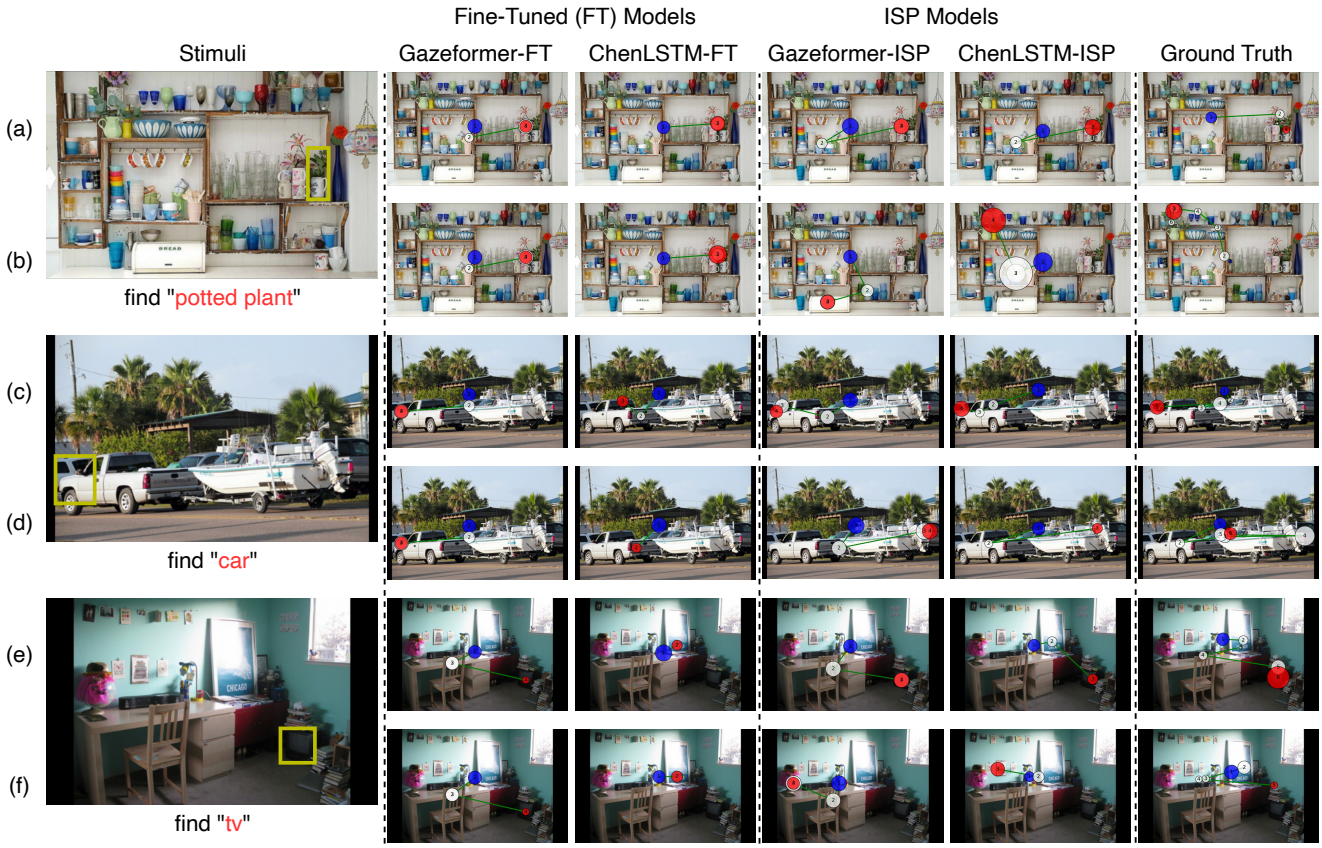

Figure 9. Qualitative scanpath examples from Gazeformer-FT, ChenLSTM-FT, Gazeformer-ISP, ChenLSTM-ISP, and ground truth on COCO-Search18 [83]. The first column shows the stimuli with the target object, and each row compares model predictions with one observer's ground truth. The gaze patterns reveal whether observers successfully find the target object. Some successfully find it, focusing on the (a) potted plant, (c) car, and (e) TV, while others fail because they misrecognize (b) flowers, (d) other vehicles, or (f) searching around the wrong place. Blue and red dots mark the start and end of scanpaths, while yellow boxes represent the search target.

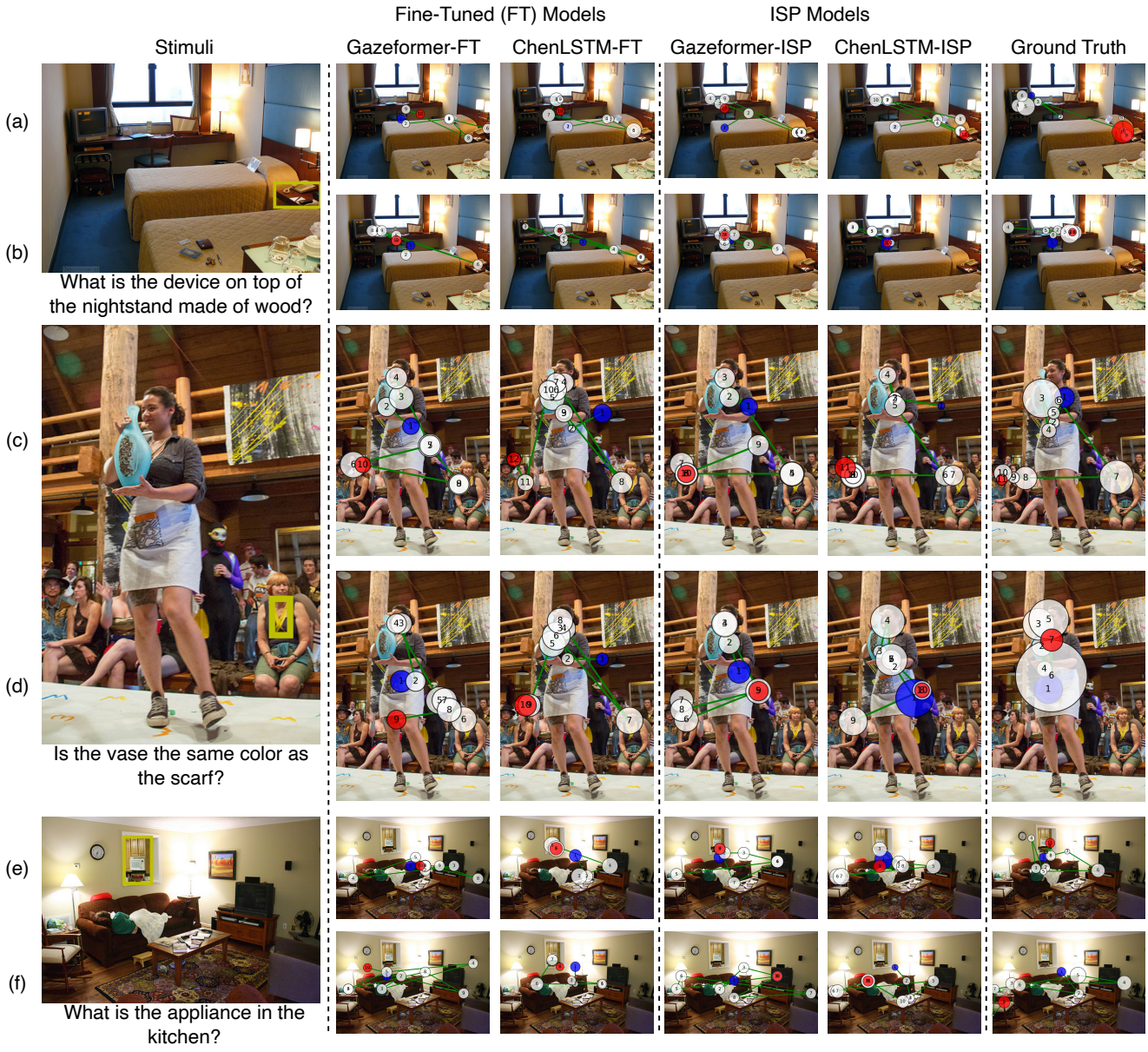

Figure 10. Qualitative scanpath examples from Gazeformer-FT, ChenLSTM-FT, Gazeformer-ISP, ChenLSTM-ISP, and ground truth on AiR-D [12]. The first column shows the stimuli with the corresponding question, and each row compares model predictions with one observer's ground truth. Examining gaze patterns reveals how correct answers correlate with observers focusing on specific regions like (a) the nightstand, (c) the scarf, and (e) the kitchen. In contrast, incorrect answers result from observers overlooking crucial objects: (b) searching the table instead of the nightstand, (d) not seeing the scarf, and (f) only exploring the living room instead of the kitchen. Blue and red dots mark the start and end of scanpaths, while yellow boxes highlight important objects for answering correctly.