Few-shot Personalized Scanpath Prediction

Supplementary Material

1. Overview

This supplementary meterial is arranged as:

- Sec. 2 shows the implementation details of ISP-SENet.
- Sec. 3 shows statistics of Tab. 1 in the main paper.
- Sec. 4 shows the supplementary evaluation of ISP-SENet.
- Sec. 5 shows ablation study on more parameters and modules.
- Sec. 6 shows more qualitative results.
  In the experiments, unless specified otherwise, we sample the 10-shot support set for 10 times.

2. Implementation Details

2.1. Feature Extractor F

Humans perceive images through a high-resolution foveal region and a low-resolution peripheral region [3], creating distinct focal and contextual areas. This principle also guides scanpath prediction models [5, 7, 8]. Following HAT [8], we encode images and scanpaths using hierarchical feature maps from the image encoder and decoder. The image encoder produces multi-scale feature maps based on ResNet [4]. To better align with human object-centric attention [6], deformable attention [9] pre-trained on segmentation tasks is utilized to generate hierarchical feature maps that capture semantic object information. The output of image decoder is four-scale hierarchical feature maps, where we utilize two feature maps with lowest and highest resolution ($P_l \in \mathbb{R}^{(\frac{H}{32}\cdot\frac{W}{32})\times C}$ and $P_h \in \mathbb{R}^{(\frac{H}{4}\cdot\frac{W}{4})\times C}$, respectively). P_l is flattened and directly used as image tokens F_I, simulating the peripheral region of human attention. We select corresponding location of all fixations from P_h and obtain $F_S \in \mathbb{R}^{L\times C}$, where L is the length of scanpath, resembling the foveated regions of human attention.

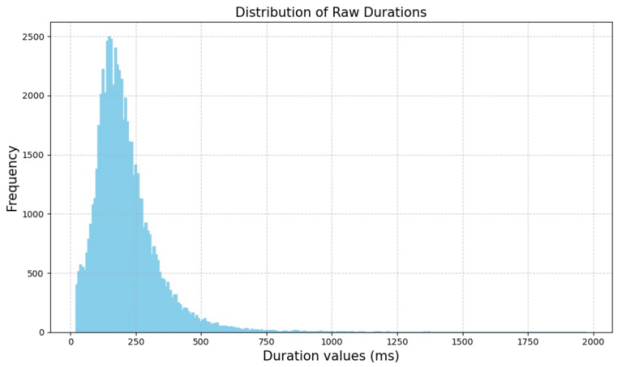
2.2. Duration

It should be noted that the duration strategy is exclusively implemented in SE-Net, whereas ISP-SENet utilizes raw durations.

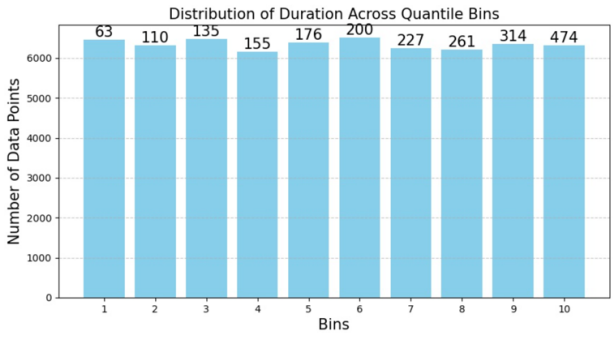
This strategy involves categorizing each fixation duration into one of ten bins, ensuring that each bin contains approximately the same number of fixation durations, a method known as quantile-based intervals. This approach is motivated by two main reasons:

1. **Significance of Duration:** Fixation duration is indicative of the importance attributed to a point in an image, as longer durations generally reflect greater interest by the viewer. By grouping durations into bins, we aim to quantitatively represent the significance, or the underlying importance, of each fixation.
2. **Distribution Characteristics:** The fixation durations exhibit a long-tail distribution, as evidenced in Fig. 1 (we collect all fixation durations values across all fixations). Employing a quantile strategy prevents the highly frequent shorter durations from clustering excessively in the initial bins. Instead, it ensures a more balanced distribution across the bins, with larger values being more evenly dispersed among them.

The number of bins is set to 10, and the visualization of each bin's statistics is shown in Fig. 1. The ablation of duration



(a)



(b)

Figure 1. **Duration distribution.** Figure (a) shows the long-tail distribution of all fixation durations in the base set of seen subjects. Figure (b) visualizes the number of points in each bin. The mean value of each bin is shown on top of each bar. In SE-Net, the bin index replaces the raw duration and is encoded using 1D sinusoidal positional encoding.

strategy is discussed in Sec. 5 and Tab. 7.

3. Statistics of Main Results

3.1. Margin of Error

To show the stabilization of our method, we show the margin of error at 95% confidence level in Tab. 1. It is obtained by sampling the support set 10 times, and ensuring each sampling set is exclusive. From the results, ISP-SENet experienced more stable performance across different support set sampling, while the performance of baselines experienced more variance, suffering from the different image content in the support set.

3.2. Second Seen-Unseen Split

To ensure the result is not biased on subjects due to the model's ability may various on different subjects, we conduct one more split of seen-unseen subjects, which still follows the rule of 70% seen and 30% unseen. To specify, 10 seen subjects and 5 unseen subjects for OSIE, 7 seen and 3 unseen subjects for both COCO-Search18 and COCO-FreeView. We ensure this second split contains different unseen subjects compared with the split shown in the main paper. The result is shown in Tab. 2. From the results in Tab. 1 and Tab. 2, we observe that ISP-SENet demonstrates stability across different seen-unseen splits, as indicated by relatively consistent performance metrics. In contrast, the performance variations in the two baselines are significant. This difference is largely attributed to the fine-tuning process, where performance is heavily dependent on the small support set. With only 10 images from each subject, substantial variation arises due to biases in image content and the specific human attention related to individual scenes. These observations further underscore the limitations of existing methods in few-shot settings.

4. Supplementary Evaluation

In this section, we evaluate the performance of ISP-SENet in three aspect:
In Sec. 4.1, we compare the performance of ISP-SENet and baselines on seen subject.
In Sec. 4.2, we use subject embeddings learned from SE-Net to replace the subject embedding of ChenLSTM-ISP, and compare with baseline fine-tuned on full training set of unseen subjects.
In Sec. 4.3, we develop a new evaluation method to validate that ISP-SENet can distinct different subject embeddings.

4.1. Results on Seen Set

In Tab. 3, we evaluate the performance of ISP-SENet and baselines on seen subjects, same as the split defined in the main paper. Notably, although the subject embeddings generated by SE-Net are frozen during the training process of ISP-SENet, indicating that they are not tailored for scanpath prediction, the performance on COCO-Search18 is significantly better. Moreover, it achieved comparable results with OSIE and COCO-FreeView. This suggests that the seen subject embeddings learned by SE-Net, despite being optimized for distinguishing different subjects rather than specifically for scanpath prediction, effectively retain individual attention traits and excel in personalized scanpath prediction.

4.2. ISP-SENet with Different Scanpath Prediction Models

To demonstrate the adaptability of the subject embeddings learned from SE-Net across different scanpath prediction models, we substituted the original subject embeddings in ChenLSTM-ISP with those learned from SE-Net. The performance of this configuration, referred to as ISP-SENet(ChenLSTM-ISP), is shown in the fourth row of Tab. 4.
Further, to compare the performance of ISP-SENet with baselines, we fine-tuned both Gazeformer-ISP and ChenLSTM-ISP on the complete training set for unseen subjects. The results, labeled as Gazeformer-ISP-FT and ChenLSTM-ISP-FT, are presented in the first and third rows of Tab. 4.
These results highlight that, without any fine-tuning on unseen subjects, ISP-SENet achieves comparable results compared with baseline models, which are fully fine-tuned on full training set of unseen subjects.

4.3. Cross-subject embedding Evaluation

To confirm that our unseen subject embeddings capture unique attention patterns rather than a global optimum applicable to all subjects, we implement a cross-subject embedding evaluation. For each unseen subject u_k, we first calculate the SM, MM, and SED metrics using the subject's own embedding e_k for predicting scanpaths on the query set. Then we replace e_k with embeddings e_i from m different subjects u_i, where $u_i \in U_{\text{unseen}, i \neq k}$, and compute the average SM, MM, and SED. The differences in these metrics underscore the uniqueness of each embedding. For simplicity, we define $m = 3$ and randomly sampled 5 different support sets.

In Tab. 5, the symbol $\times$ represents that the subject embedding and prediction correspond to the same subject, while ✓ indicates they belong to different subjects. To better understand the model performance of cross-subject embedding evaluation, we include comparisons with ISP(Seen) and ISP-SENet(Seen). **ISP(Seen)** evaluates Gazeformer-ISP's ability to differentiate among embeddings of seen subjects. As ISP-SENet is built upon Gazeformer-ISP, the distinction ability of these two models will not have significant differences. **ISP-SENet(Seen)** as-

(a) OSIE

n-shot	Method	SM ↑	MM ↑	SED ↓
n = 1	ChenLSTM-ISP	0.282 ± 0.009	0.763 ± 0.006	7.832 ± 0.181
	Gazeformer-ISP	0.327 ± 0.007	0.792 ± 0.003	7.873 ± 0.134
	ChenLSTM-ISP-S	0.328 ± 0.001	0.793 ± 0.001	7.601 ± 0.039
	Gazeformer-ISP-S	0.354 ± 0.000	0.801 ± 0.000	7.503 ± 0.003
	ISP-SENet	**0.368 ± 0.003**	**0.805 ± 0.002**	**7.413 ± 0.033**
n = 5	ChenLSTM-ISP	0.319 ± 0.005	0.773 ± 0.004	7.855 ± 0.116
	Gazeformer-ISP	0.340 ± 0.003	0.791 ± 0.002	7.920 ± 0.082
	ChenLSTM-ISP-S	0.329 ± 0.001	0.801 ± 0.000	7.499 ± 0.028
	Gazeformer-ISP-S	0.354 ± 0.000	0.791 ± 0.001	7.699 ± 0.003
	ISP-SENet	**0.376 ± 0.002**	**0.803 ± 0.001**	**7.649 ± 0.028**
n = 10	ChenLSTM-ISP	0.322 ± 0.005	0.777 ± 0.002	7.740 ± 0.079
	Gazeformer-ISP	0.345 ± 0.003	0.794 ± 0.002	7.916 ± 0.054
	ChenLSTM-ISP-S	0.328 ± 0.005	0.791 ± 0.001	7.637 ± 0.060
	Gazeformer-ISP-S	0.354 ± 0.000	0.802 ± 0.000	7.505 ± 0.003
	ISP-SENet	**0.375 ± 0.001**	**0.803 ± 0.001**	**7.318 ± 0.017**

(b) COCO-FreeView

n-shot	Method	SM ↑	MM ↑	SED ↓
n = 1	ChenLSTM-ISP	0.287 ± 0.014	0.805 ± 0.003	13.307 ± 0.195
	Gazeformer-ISP	0.244 ± 0.021	0.787 ± 0.011	15.118 ± 0.510
	ChenLSTM-ISP-S	0.339 ± 0.000	0.814 ± 0.000	12.523 ± 0.029
	Gazeformer-ISP-S	0.333 ± 0.000	0.817 ± 0.000	12.538 ± 0.012
	ISP-SENet	**0.369 ± 0.002**	**0.832 ± 0.001**	**12.227 ± 0.134**
n = 5	ChenLSTM-ISP	0.320 ± 0.009	0.815 ± 0.005	12.950 ± 0.190
	Gazeformer-ISP	0.286 ± 0.012	0.800 ± 0.005	14.630 ± 0.310
	ChenLSTM-ISP-S	0.338 ± 0.000	0.814 ± 0.000	12.540 ± 0.023
	Gazeformer-ISP-S	0.333 ± 0.000	0.817 ± 0.000	12.539 ± 0.008
	ISP-SENet	**0.368 ± 0.001**	**0.829 ± 0.001**	**12.017 ± 0.058**
n = 10	ChenLSTM-ISP	0.323 ± 0.010	0.819 ± 0.005	12.541 ± 0.114
	Gazeformer-ISP	0.317 ± 0.002	0.805 ± 0.002	14.224 ± 0.207
	ChenLSTM-ISP-S	0.340 ± 0.000	0.814 ± 0.000	12.532 ± 0.025
	Gazeformer-ISP-S	0.333 ± 0.000	0.816 ± 0.000	12.545 ± 0.006
	ISP-SENet	**0.367 ± 0.001**	**0.828 ± 0.001**	**11.956 ± 0.010**

(c) COCO-Search18

n-shot	Method	SM ↑	MM ↑	SED ↓
n = 1	ChenLSTM-ISP	0.371 ± 0.024	0.760 ± 0.029	2.756 ± 0.464
	Gazeformer-ISP	0.342 ± 0.018	0.770 ± 0.008	2.818 ± 0.216
	ChenLSTM-ISP-S	0.448 ± 0.000	0.803 ± 0.001	2.394 ± 0.013
	Gazeformer-ISP-S	0.446 ± 0.001	0.802 ± 0.001	2.463 ± 0.002
	ISP-SENet	**0.475 ± 0.007**	**0.814 ± 0.001**	**2.333 ± 0.063**
n = 5	ChenLSTM-ISP	0.386 ± 0.015	0.773 ± 0.008	2.489 ± 0.058
	Gazeformer-ISP	0.353 ± 0.028	0.774 ± 0.011	2.980 ± 0.292
	ChenLSTM-ISP-S	0.449 ± 0.001	0.803 ± 0.001	2.380 ± 0.014
	Gazeformer-ISP-S	0.445 ± 0.001	0.803 ± 0.001	2.457 ± 0.002
	ISP-SENet	**0.484 ± 0.005**	**0.815 ± 0.001**	**2.354 ± 0.044**
n = 10	ChenLSTM-ISP	0.393 ± 0.006	0.781 ± 0.004	2.394 ± 0.038
	Gazeformer-ISP	0.370 ± 0.007	0.785 ± 0.006	2.765 ± 0.128
	ChenLSTM-ISP-S	0.449 ± 0.000	0.803 ± 0.001	2.379 ± 0.019
	Gazeformer-ISP-S	0.446 ± 0.001	0.802 ± 0.001	2.464 ± 0.002
	ISP-SENet	**0.482 ± 0.002**	**0.815 ± 0.001**	**2.359 ± 0.019**

Table 1. Margin of error for Tab. 1 in the main paper.

sesses ISP-SENet's cross-subject embedding performance on seen subjects, indicative of the potential upper limit of our model's discriminative capability. **ISP-SENet(Unseen)** represents our cross-subject embedding evaluation for unseen subjects. The results demonstrate that ISP-SENet's capacity to distinguish unseen subjects exceeds the baseline's

(a) OSIE

Method	SM ↑	MM ↑	SED ↓
ChenLSTM-ISP	0.288 ± 0.009	0.780 ± 0.004	**7.350 ± 0.127**
Gazeformer-ISP	0.318 ± 0.006	0.789 ± 0.002	8.363 ± 0.155
ISP-SENet	**0.384 ± 0.001**	**0.813 ± 0.001**	7.460 ± 0.022

(b) COCO-FreeView

Method	SM ↑	MM ↑	SED ↓
ChenLSTM-ISP	0.296 ± 0.007	0.823 ± 0.001	12.534 ± 0.030
Gazeformer-ISP	0.275 ± 0.008	0.801 ± 0.006	14.266 ± 0.286
ISP-SENet	**0.364 ± 0.001**	**0.835 ± 0.001**	**12.342 ± 0.019**

(c) COCO-Search18

Method	SM ↑	MM ↑	SED ↓
ChenLSTM-ISP	0.333 ± 0.006	0.766 ± 0.006	2.712 ± 0.042
Gazeformer-ISP	0.403 ± 0.010	0.803 ± 0.005	2.734 ± 0.116
ISP-SENet	**0.465 ± 0.001**	**0.812 ± 0.001**	**2.286 ± 0.020**

Table 2. Results from the second seen-unseen split under the 10-shot setting. The unseen subject set in this split is distinct from the unseen set used in the main paper's split.

Methods	OSIE			COCO-FreeView			COCO-Search18		
	SM ↑	MM ↑	SED ↓	SM ↑	MM ↑	SED ↓	SM ↑	MM ↑	SED ↓
ChenLSTM-ISP	0.373	0.814	7.171	0.373	0.828	12.126	0.475	0.820	2.128
Gazeformer-ISP	0.382	0.813	**7.077**	**0.380**	**0.835**	**11.707**	0.480	0.815	2.204
ISP-SENet	**0.382**	**0.816**	7.127	0.375	0.833	11.872	**0.517**	**0.825**	**2.086**

Table 3. Performance Comparison of methods on seen subjects. All methods are trained on all training data of seen subjects, and test on the test set of seen subjects.

Methods	OSIE			COCO-FreeView			COCO-Search18		
	SM ↑	MM ↑	SED ↓	SM ↑	MM ↑	SED ↓	SM ↑	MM ↑	SED ↓
Gazeformer-ISP-FT	0.372	0.803	7.614	0.383	0.834	11.443	0.479	0.815	2.330
ISP-SENet (Gazeformer-ISP)	0.375	0.803	7.318	0.367	0.828	11.956	0.482	0.815	2.359
ChenLSTM-ISP-FT	0.371	0.801	7.449	0.387	0.832	11.422	0.475	0.813	2.159
ISP-SENet (ChenLSTM-ISP)	0.369	0.800	7.574	0.366	0.824	12.241	0.467	0.810	2.272

Table 4. ISP-SENet with Different Scanpath Prediction Models, and comparsion between ISP-SENet without fine-tuning on unseen subjects, with ISP[2] fine-tuned on full training set of unseen subjects.

performance with seen subjects and is comparable with ISP-SENet's performance on seen subjects.

4.4. Quantatitive results on visual-task encoder

To demonstrate that the visual-task encoder effectively captures the alignment between the task and image content, we evaluate the similarity between its cross-attention maps and the ground truth bounding boxes using Correlation Coefficient (CC) and AUC. For SE-Net, we achieve a CC of 0.31 and an AUC of 0.76. While the CC is sensitive to false positives—such as attention allocated to relevant peripheral objects—our model still outperforms ChenLSTM-ISP's task-guidance map m_0(CC = 0.07, AUC = 0.63), averaged across channels. This indicates stronger target understanding, despite our model not being explicitly designed for object detection.

Method	cross-subject embedding	OSIE		
		SM ↑	MM ↑	SED ↓
ISP(Seen)	×	0.386	0.814	7.003
	✓	0.379	0.812	7.163
ISP-SENet(Seen)	×	0.387	0.815	7.009
	✓	0.373	0.810	7.360
ISP-SENet(Unseen)	×	0.376	0.802	7.286
	✓	0.361	0.800	7.340

Table 5. Cross-embedding evaluation on the distinction ability between subject embeddings. The symbol × denotes that the subject embedding and prediction correspond to the same subject, while ✓ indicates that the subject embeddings and prediction belong to different subjects.

4.5. More analysis on size of base set

In Tab. 6, we analyze the impact of varying the number of seen subjects in the base set during training on COCO-Search18, supplementing the results in Table 4 of the main paper. We consistently select one subject as unseen and vary the number of seen subjects selected from the remaining ones. With fewer seen subjects, SE-Net struggles to infer the attention traits of new subjects based on its learned experience. Performance improves when increasing the number of seen subjects from 7 (as in the main paper) to 9, suggesting that ISP-SE-Net benefits from additional subjects.

num seen	SM ↑	MM ↑	SED ↑
4(40%)	0.472	0.819	2.542
9(90%)	**0.489**	**0.826**	**2.145**
Ours(70%)	0.487	0.823	2.333

Table 6. Performance comparison of 10-shot setting on COCO-Search18 with different numbers of seen subjects in training stage.

5. More Ablation Results

5.1. Duration

In Tab. 7, we ablate the effect of duration encoding strategy on OSIE in three settings: (1) No duration encoding in scanpath embeddings. (2) Encoding raw duration without assigning bin index. (3) Assigning durations to 10 bins of equal width, without employing the quantile strategy. (4) Assigning durations to 100 bins using the quantile strategy. (5) Assigning durations to 300 bins using the quantile strategy. (6) Assigning durations to 10 bins using the quantile strategy as in the main paper.

The result indicates that: (1) Duration is crucial for understanding subject attention traits. (2) Raw durations offer limited information as they introduce redundancy. For instance, 200 ms and 201 ms are treated as distinct durations, despite their negligible difference, which does not accurately reflect varying importance levels between two fixations. (3) Without the quantile strategy, the bins fail to manage the long-tail effect effectively, resulting in sparse distribution of higher durations across most bins while smaller durations crowd into a few bins due to their higher frequency. (4) and (5) demonstrate how varying the number of bins impacts performance.

Duration Strategy	SM ↑	MM ↑	SED ↓
w/o Duration	0.365	0.797	7.634
Raw duration	0.367	0.800	7.534
Uniform bin width	0.369	0.799	7.431
100 bins	0.374	0.801	7.377
300 bins	0.370	0.801	7.474
ISP-SENet (10 bins)	**0.375**	**0.803**	**7.318**

Table 7. Ablation on performance with different duration encoding strategy on OSIE.

5.2. Margin in Contrastive Loss

We ablate the effect of different margins m in the contrastive loss. The margin is a predefined threshold that specifies how much farther the negative example should be from the anchor compared to the positive example. As shown in Tab. 8, lower or higher margins decrease the prediction performance. A possible reason is lower margin prevents SE-Net from distinguishing different subjects.

The performance decreases associated with a higher margin can be attributed to the characteristics of human attention. Despite differences among subjects, their scanpaths often share similarities, such as a focus on foreground objects like humans. Such similarity is critical for SE-Net to learn the subject embedding, and plays a key role in infering embeddings for unseen subjects. We anticipate that embeddings for unseen subjects will benefit from seen subjects with similar attention patterns. Thus, setting a higher margin may overlook these essential similarities.

Tab. 8 shows that, for COCO-Search18 dataset where viewing patterns between people are more similar (higher Human Consistency(HC)[1]), a smaller margin of 1 performs better. For OSIE dataset with more diverse patterns (lower HC), a larger margin of 5 performs better. Also the effect of margin is more significant for higher HC.

5.3. Embedding Dimension

In Tab. 9 we explore different embedding dimensions of SE-Net and ISP-SENet. All layers in SE-Net and ISP-SENet shares the same embedding dimensions. The results indicate that varying the embedding dimensions does not significantly impact performance.

margin	OSIE(HC=0.39)			COCO-Search18(HC=0.52)		
	SM ↑	MM ↑	SED ↓	SM ↑	MM ↑	SED ↓
1	0.369	0.804	7.506	**0.482**	**0.815**	**2.359**
5	**0.375**	0.803	**7.318**	0.467	0.815	2.455
10	0.367	**0.809**	7.546	0.445	0.813	2.563

Table 8. Ablation on performance with different m in contrastive loss on OSIE and COCO-Search18. In the main paper, we use $m = 5$ for OSIE and $m = 1$ for COCO-Search18.

Embedding Dimension	SM ↑	MM ↑	SED ↓
128	0.374	**0.804**	7.324
384	**0.375**	0.803	**7.318**

Table 9. Ablation on performance with different embedding dimensions.

6. Qualitative Results

We show more qualitative results of ISP-SENet on OSIE, COCO-FreeView and COCO-Search18 in Fig. 2, Fig. 3, Fig. 4. In most cases, ISP-SE-Net successfully captures the variation in viewed objects across different subjects. Notably, on COCO-FreeView, it learns global attention patterns such as centralized or scattered focus. In the search task, our model also identifies subjects influenced by distractions, outperforming the baselines in capturing such behaviors.

References

[1] Xianyu Chen, Ming Jiang, and Qi Zhao. Predicting human scanpaths in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10876–10885, 2021. 5

[2] Xianyu Chen, Ming Jiang, and Qi Zhao. Beyond average: Individualized visual scanpath prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25420–25431, 2024. 4

[3] Christine A Curcio, Kenneth R Sloan, Robert E Kalina, and Anita E Hendrickson. Human photoreceptor topography. *Journal of comparative neurology*, 292(4):497–523, 1990. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[5] Mengtang Li, Jie Zhu, Zhixin Huang, and Chao Gou. Imitating the human visual system for scanpath predicting. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3745–3749. IEEE, 2024. 1

[6] Brian J Scholl. Objects and attention: The state of the art. *Cognition*, 80(1-2):1–46, 2001. 1

[7] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Target-absent human attention. In *European Conference on Computer Vision*, pages 52–68. Springer, 2022. 1

[8] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Ruoyu Xue, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Unifying top-down and bottom-up scanpath prediction using transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1683–1693, 2024. 1

[9] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1

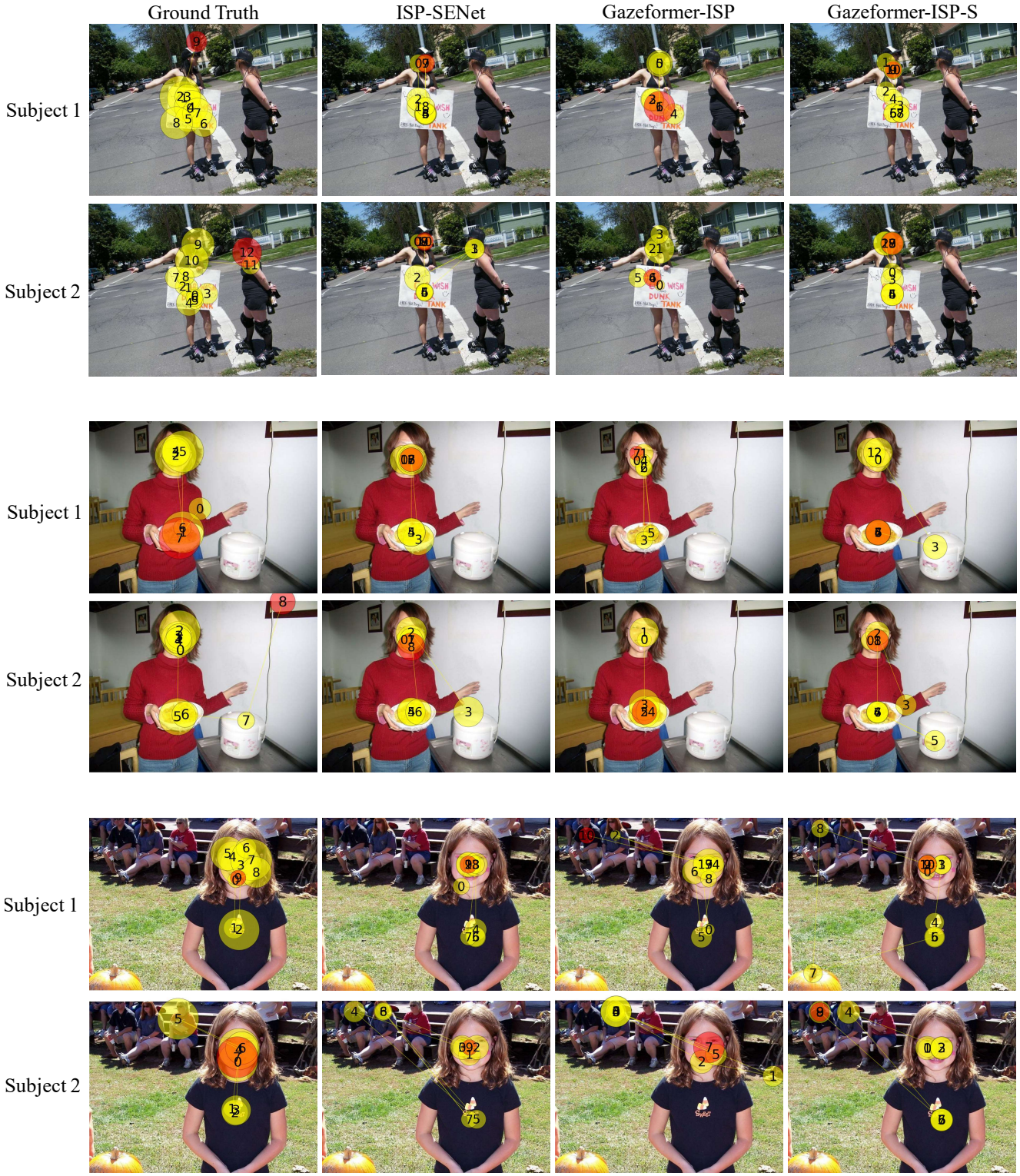


Figure 2. **More Qualitative examples of scanpath prediction for different unseen subjects on OSIE.** GT is the ground truth scanpaths of different unseen subjects. Red circle is the end fixation. Each two rows of the same image are scanpaths belonging to two different unseen subjects.

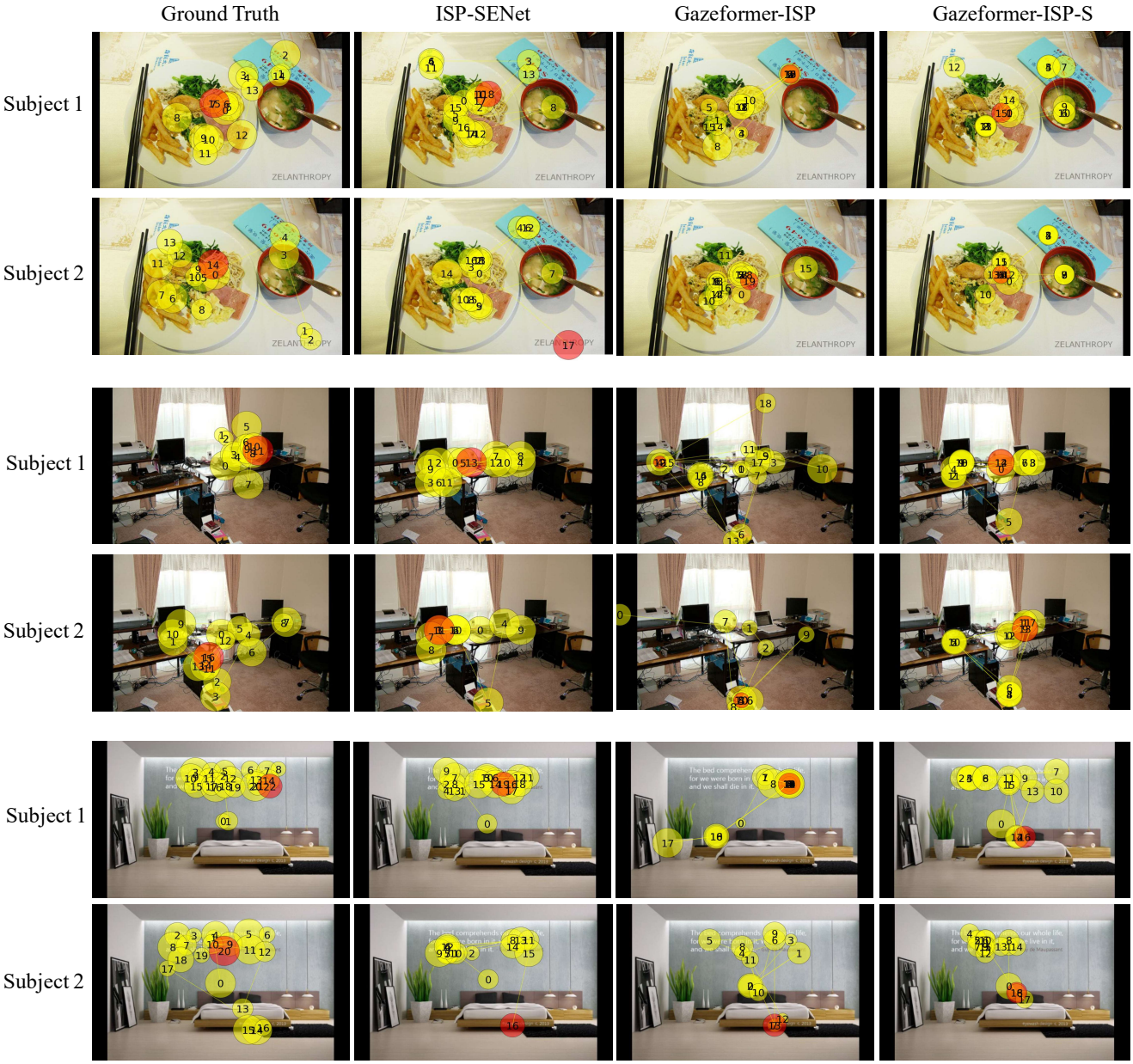


Figure 3. **More Qualitative examples of scanpath prediction for different unseen subjects on COCO-FreeView.** GT is the ground truth scanpaths of different unseen subjects. Red circle is the end fixation. Each two rows of the same image are scanpaths belonging to two different unseen subjects.

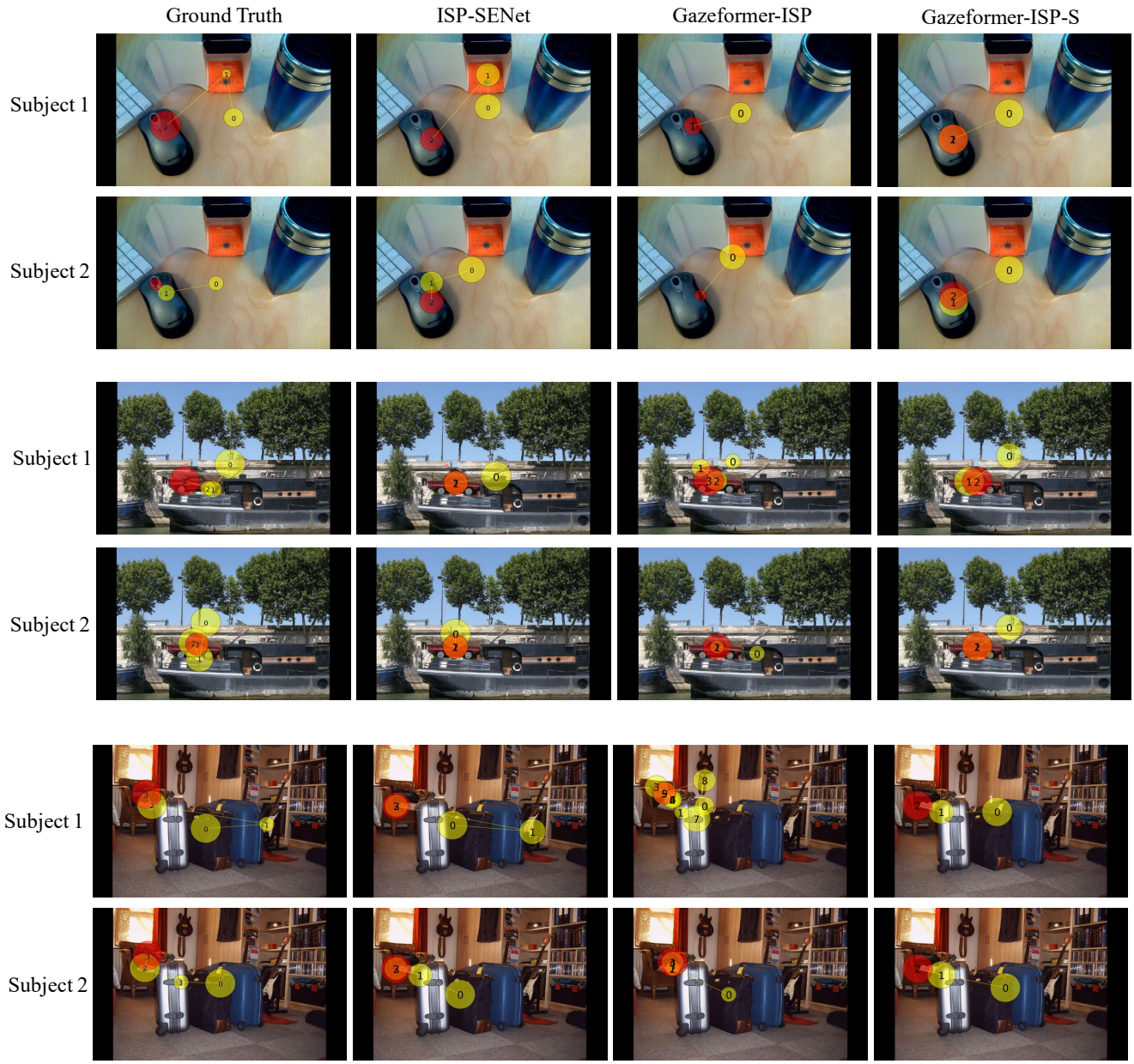


Figure 4. **More Qualitative examples of scanpath prediction for different unseen subjects on COCO-Search18.** GT is the ground truth scanpaths of different unseen subjects. Red circle is the end fixation. Each two rows of the same image are scanpaths belonging to two different unseen subjects. The search targets are **mouse, car, chair**, respectively.