CVPR Nashville JUNE 11-15, 2025

Few-shot Personalized Scanpath Prediction


Ruoyu Xue


Jingyi Xu


Sounak Mondal


Hieu Le


Gregory Zelinsky


Minh Hoai


Dimitris Samaras

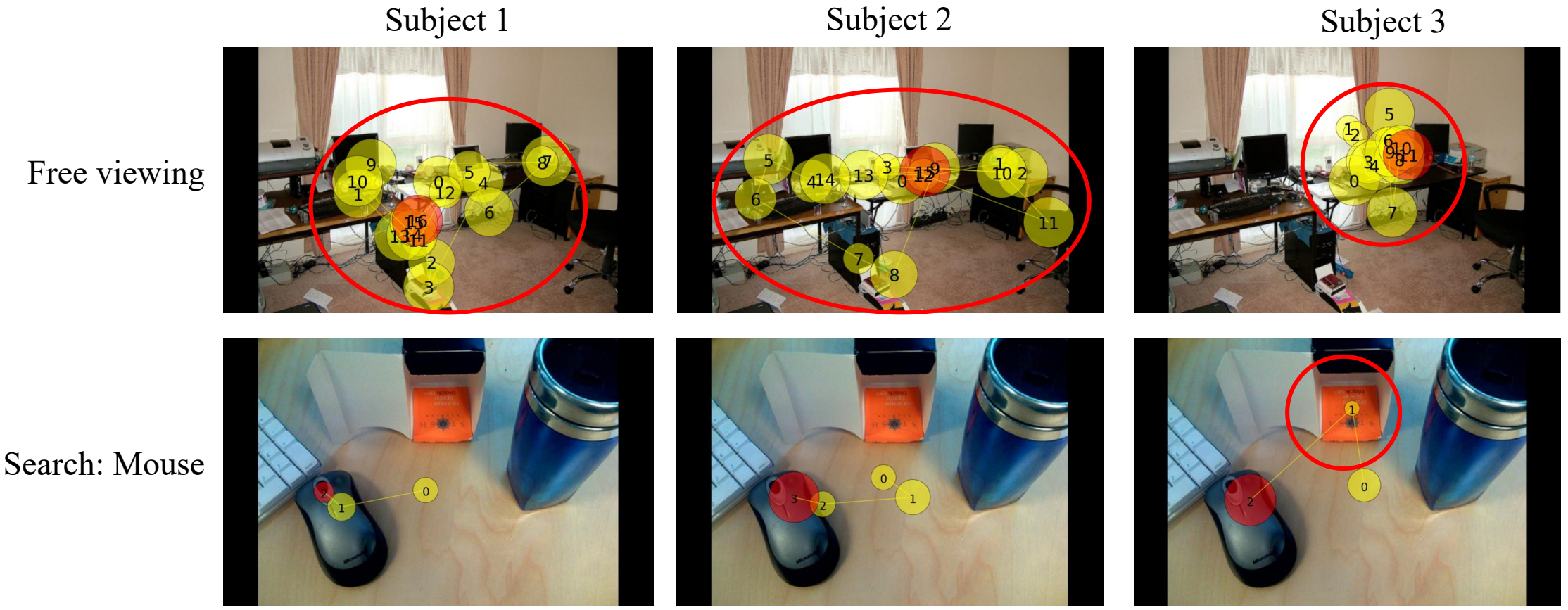
Stony Brook University

EPFL

THE UNIVERSITY of ADELAIDE

Personalized Scanpath Prediction

- Predicting individual-specific sequence of eye movements.
  - Individuals are different.
- Human attention various under different viewing tasks.



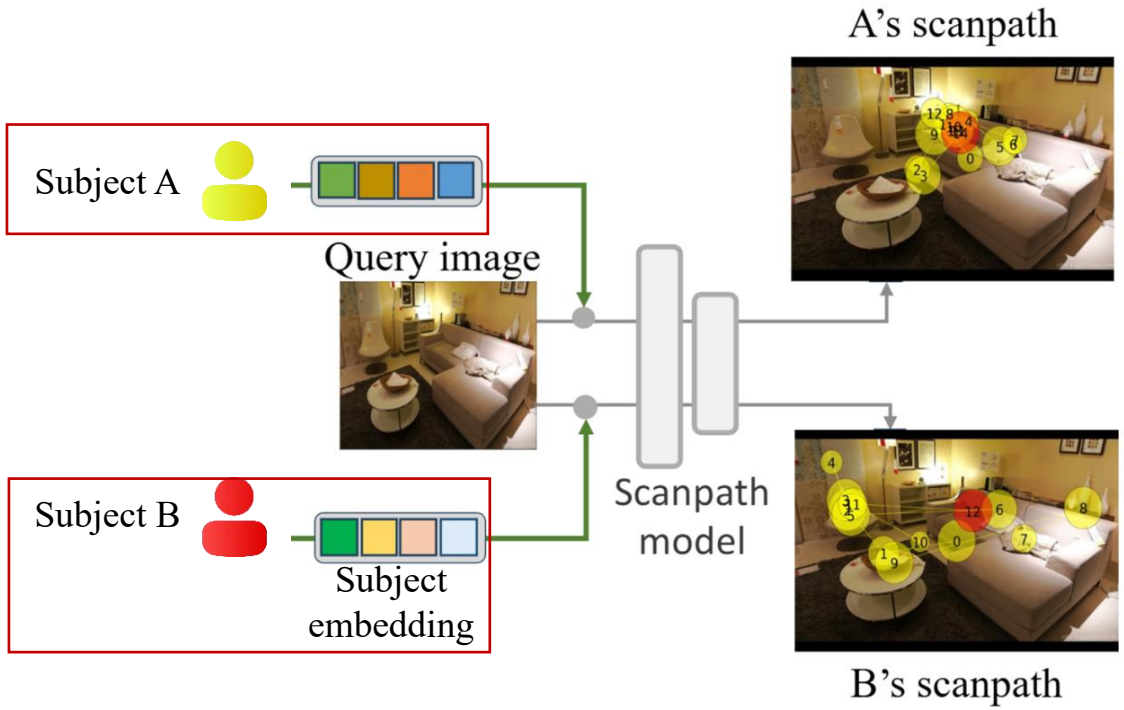
Few-shot personalized scanpath prediction

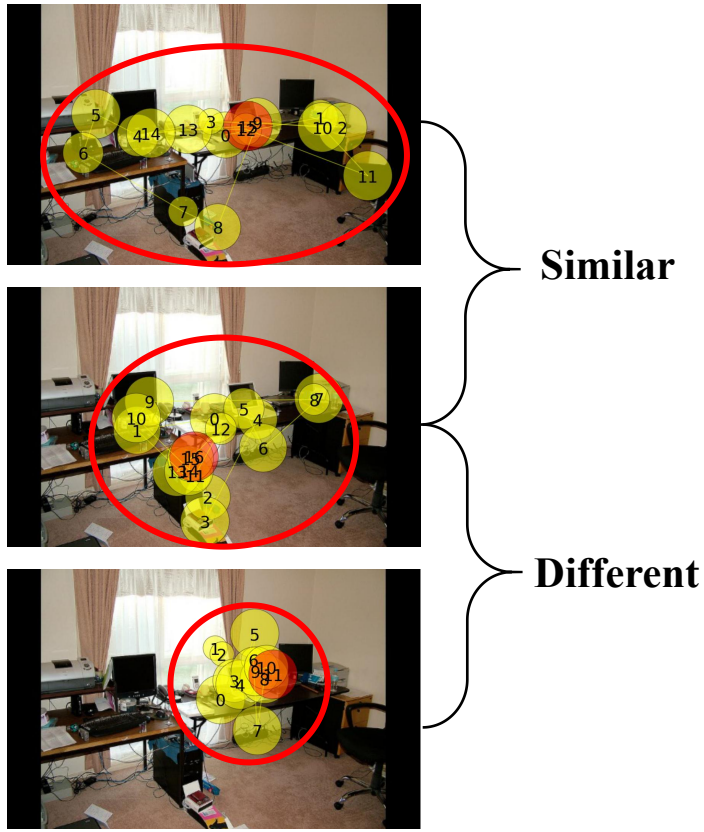
- Scanpath prediction models are data-hungry and parameter-heavy.
  - Collecting scanpath data requires up to **12** hours per subject.
  - Impractical in **real-life** application such as advertisement and recommendation.

- Few-shot Personalized Scanpath Prediction
  - Predict scanpath of **unseen** subject with **a few** labeled examples (image-scanpath).

Limitation of existing approachs

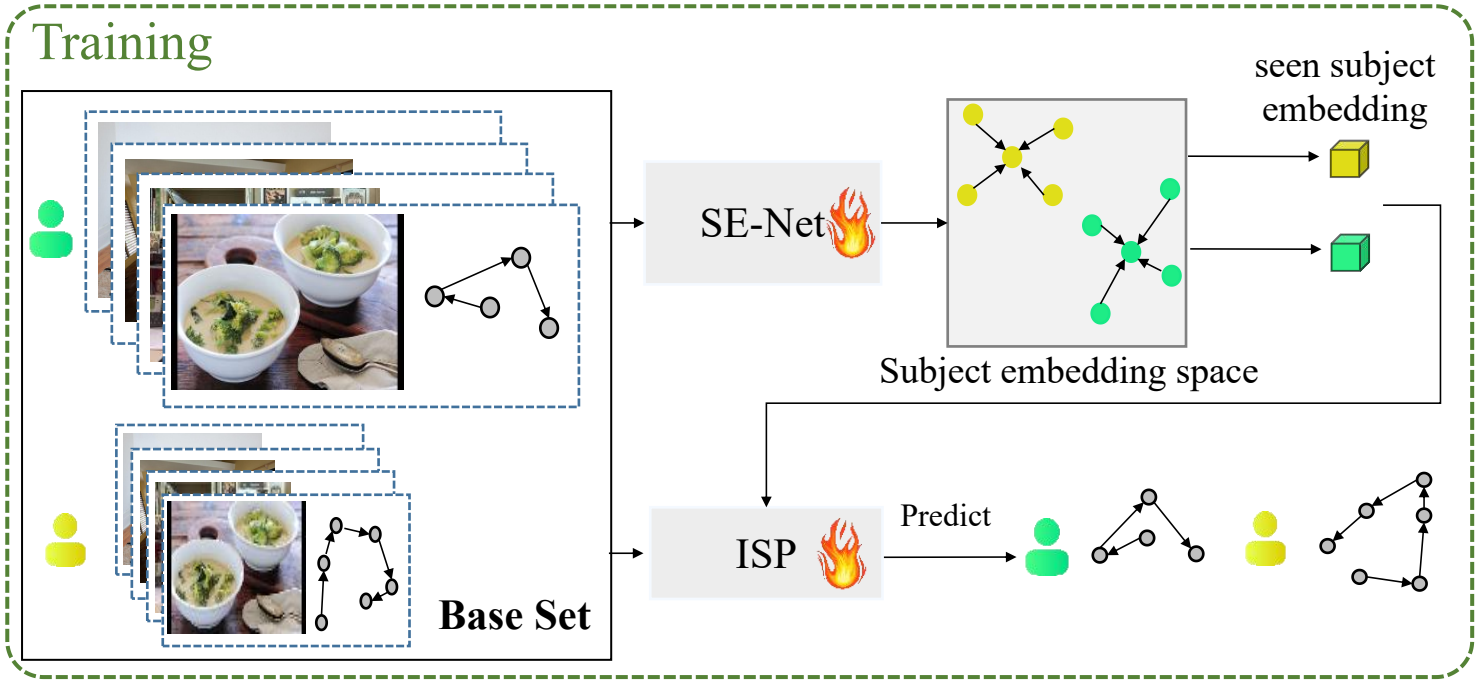
- ISP [Chen et al., CVPR 2024], EyeFormer [Jiang et al., USIT 2024]
  - Overfitting: Fine-tuning the model for **unseen subject** with minimal support data.
  - Failed to utilize subject similarity and difference.





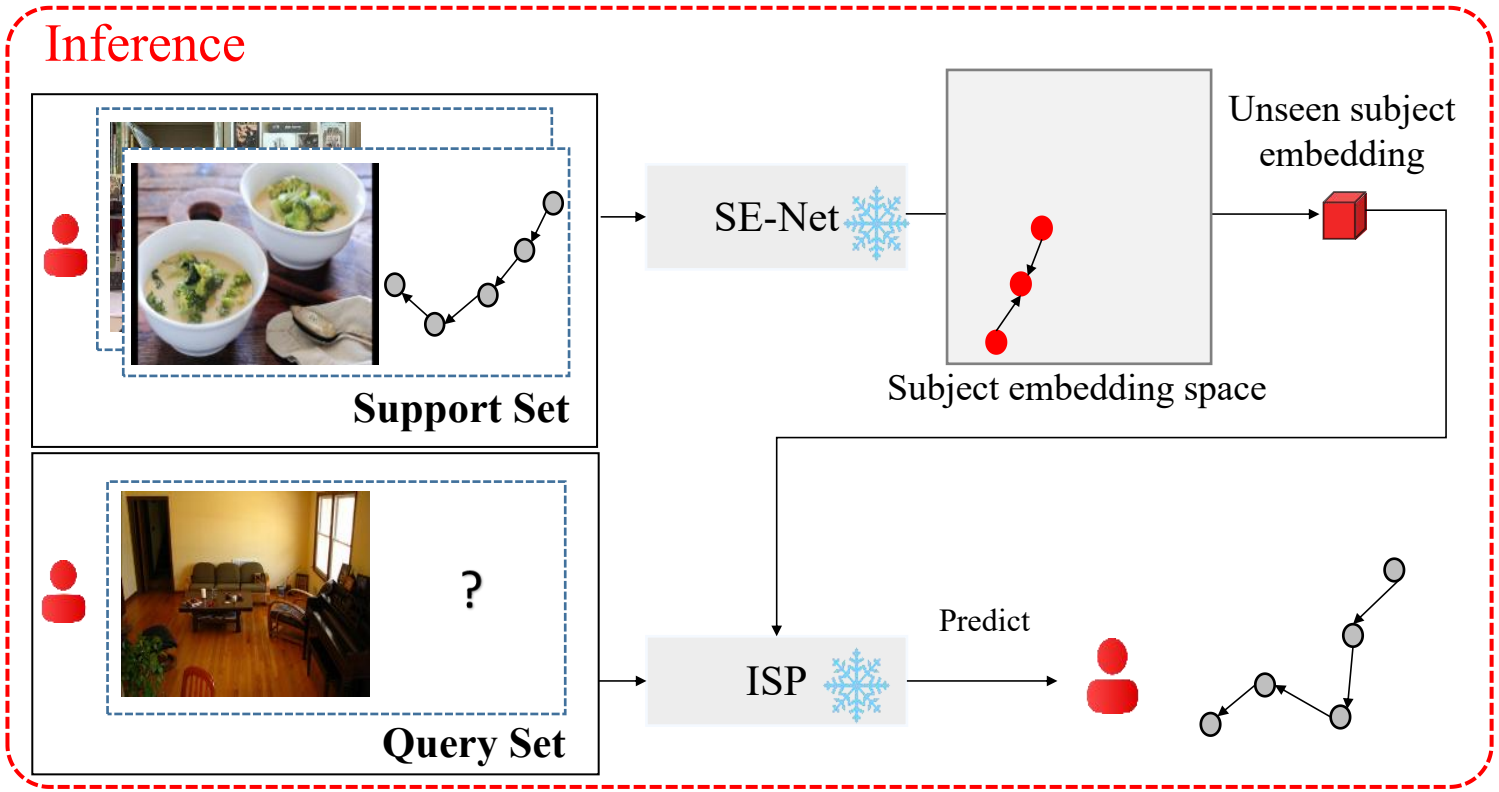
Method – ISP-SENet

- Separately train two networks.
- SE-Net to learn subject embedding to represent unique attention traits and capture inter-subject variability.
- Individual Scanpath Predictor (ISP) [Chen et al., CVPR 2024] to predict scanpath.



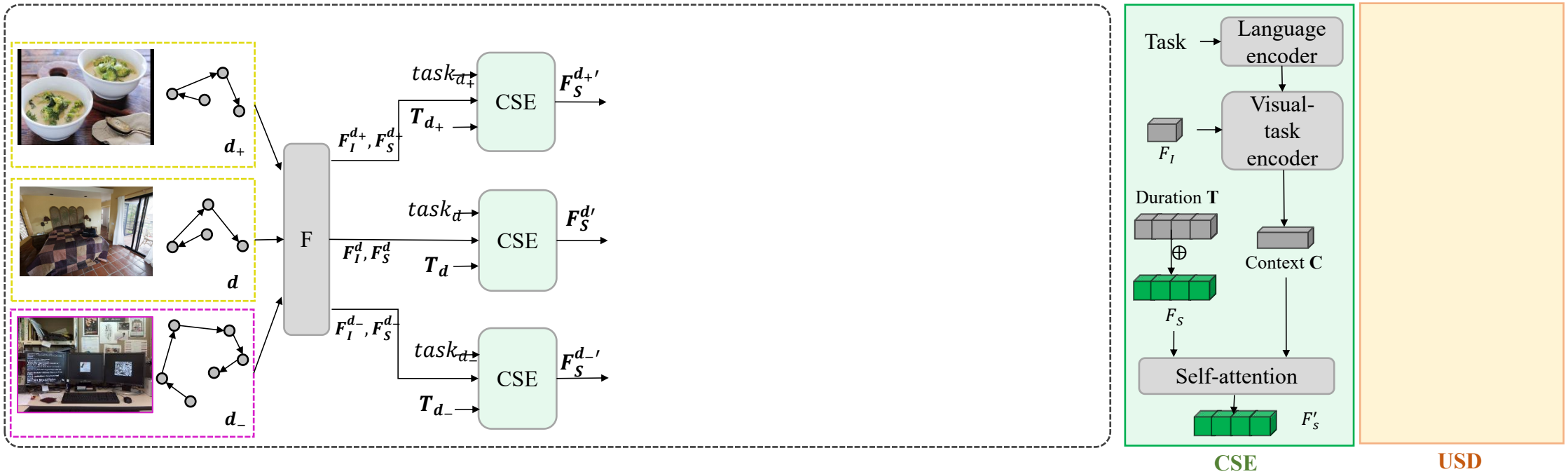
Method – ISP-SENet

- Extract unseen subject embedding → guide ISP to predict scanpath of unseen subject on unseen images.
  - No fine-tuning on unseen subjects.
  - Reduce overfitting and benefit real-time applications.



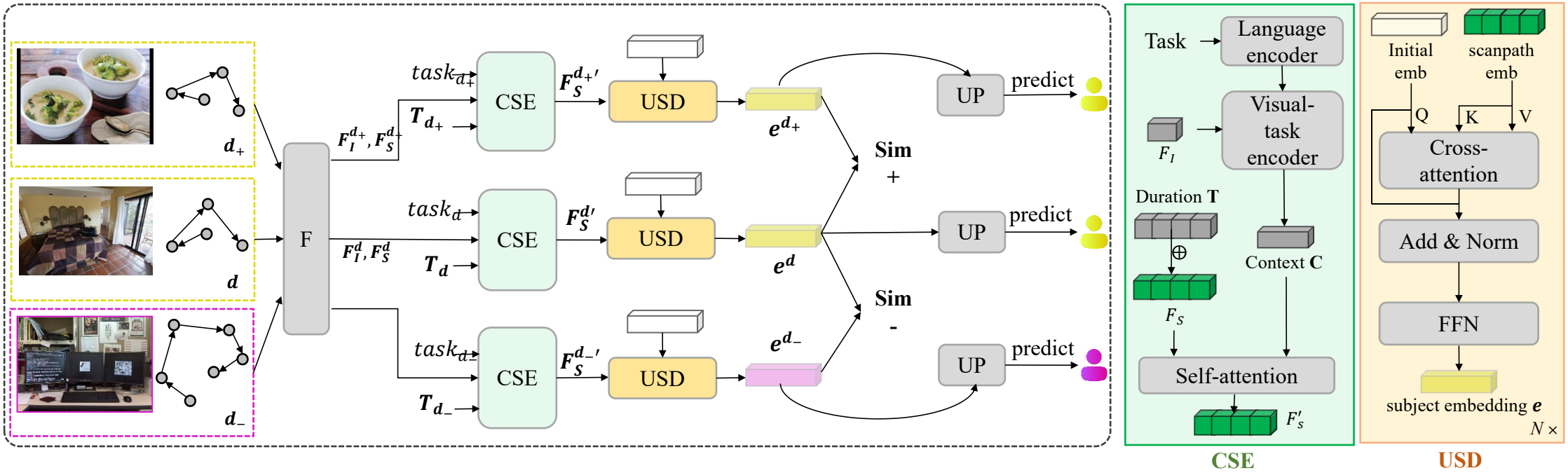
Method – Subject Embedding Network(SE-Net)

- Context-Scanpath Encoder (CSE): Obtain refined scanpath features $\boldsymbol{F_S^{d\prime}}$ from image $\boldsymbol{F_I^d}$, fixations position $\boldsymbol{F_S^d}$ and duration $\boldsymbol{T_d}$, viewing task $\overline{task_d}$



Method – Subject Embedding Network(SE-Net)

- User-Scanpath Decoder (USD): Extract subject embedding $\boldsymbol{e^d}$ from refined scanpath features $\boldsymbol{F_S^{d\prime}}$.
- User-Predictor (UP): Subject classification.
- Contrastive Learning: capture inter-subject variability.



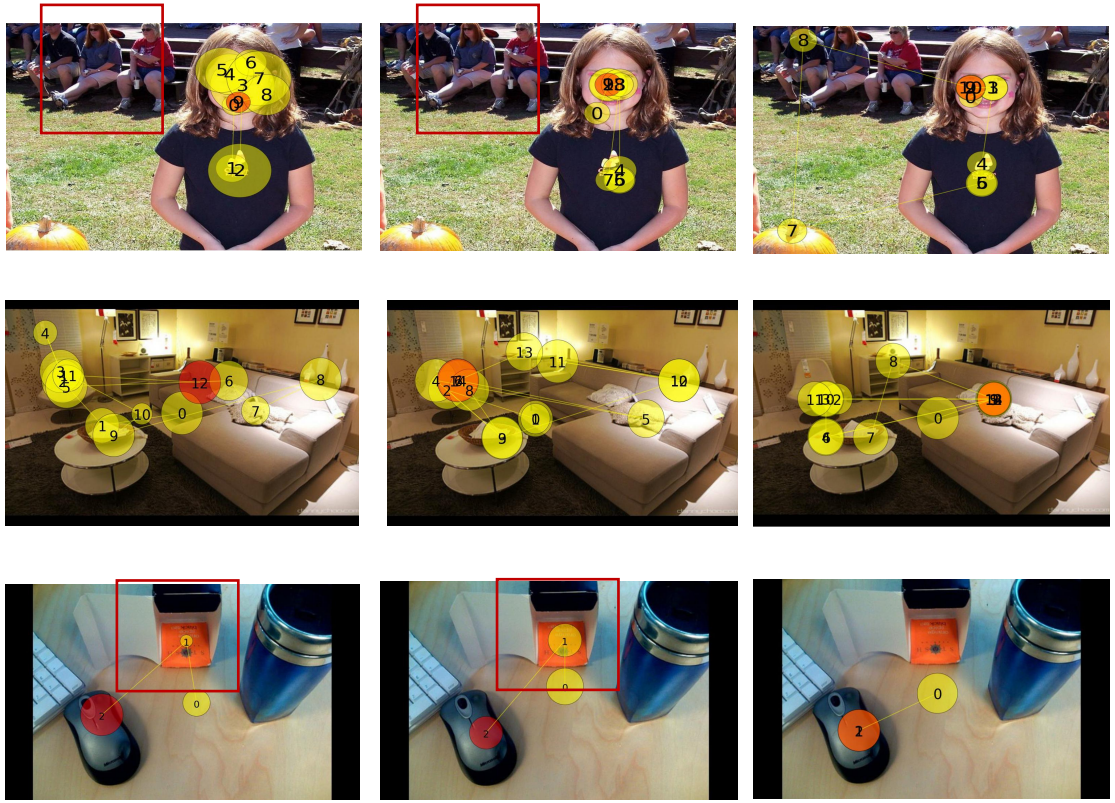
Quantitative Results

n-shot	Method	OSIE			COCO-FreeView			COCO-Search18		
		SM ↑	MM ↑	SED ↓	SM ↑	MM ↑	SED ↓	SM ↑	MM ↑	SED ↓
$n = 1$	ChenLSTM-ISP	0.282	0.763	7.832	0.287	0.805	13.307	0.371	0.760	2.756
	Gazeformer-ISP	0.327	0.792	7.873	0.244	0.787	15.118	0.342	0.770	2.818
	ChenLSTM-ISP-S	0.328	0.793	7.601	0.339	0.814	12.523	0.448	0.803	2.394
	Gazeformer-ISP-S	0.354	0.801	7.503	0.333	0.817	12.538	0.446	0.802	2.463
	ISP-SENet	**0.368**	**0.805**	**7.413**	**0.369**	**0.832**	**12.227**	**0.475**	**0.814**	**2.333**
$n = 5$	ChenLSTM-ISP	0.319	0.773	7.855	0.320	0.815	12.950	0.386	0.773	2.489
	Gazeformer-ISP	0.340	0.791	7.920	0.286	0.800	14.630	0.353	0.774	2.980
	ChenLSTM-ISP-S	0.329	0.791	7.649	0.338	0.814	12.540	0.449	0.803	2.380
	Gazeformer-ISP-S	0.354	0.801	7.499	0.333	0.817	12.539	0.445	0.803	2.457
	ISP-SENet	**0.376**	**0.803**	**7.337**	**0.368**	**0.829**	**12.017**	**0.484**	**0.815**	**2.354**
$n = 10$	ChenLSTM-ISP	0.322	0.777	7.740	0.323	0.819	12.541	0.393	0.781	2.394
	Gazeformer-ISP	0.345	0.794	7.916	0.317	0.805	14.224	0.370	0.785	2.765
	ChenLSTM-ISP-S	0.328	0.791	7.637	0.340	0.814	12.532	0.449	0.803	2.379
	Gazeformer-ISP-S	0.354	0.802	7.505	0.333	0.816	12.545	0.446	0.802	2.464
	ISP-SENet	**0.375**	**0.803**	**7.318**	**0.367**	**0.828**	**11.956**	**0.482**	**0.815**	**2.359**

Qualitative results

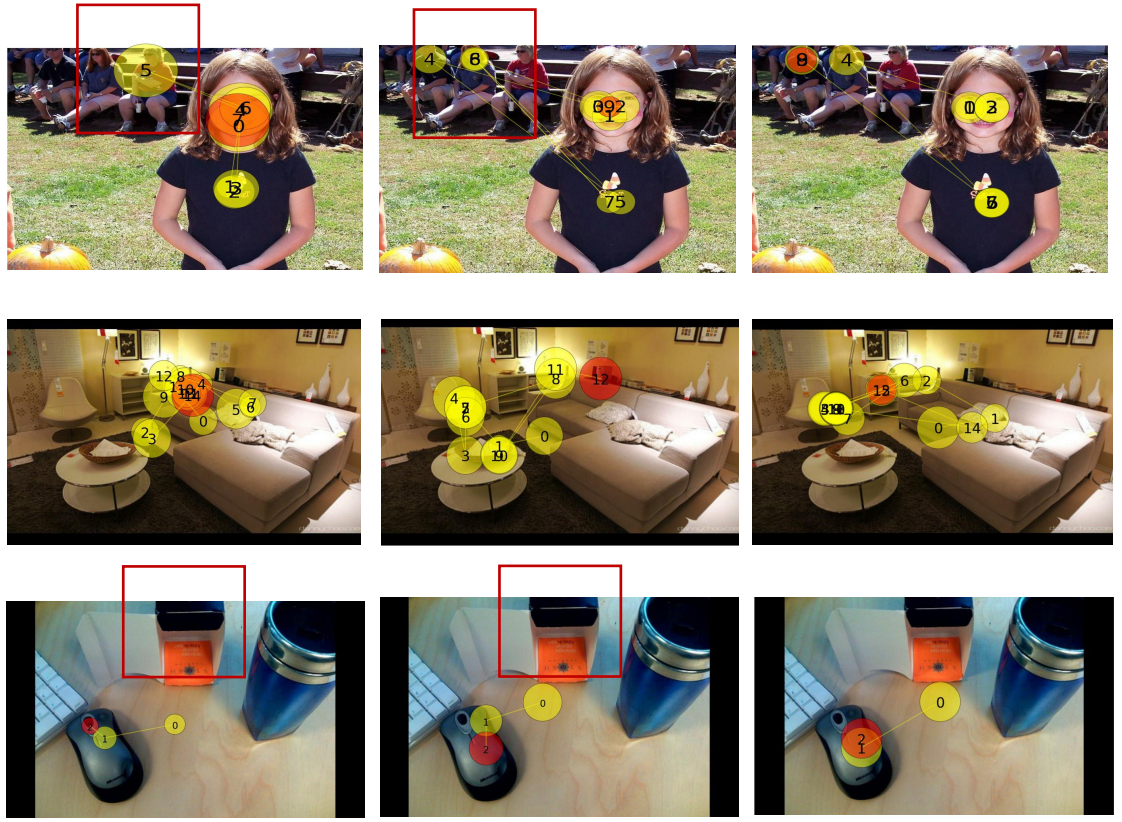
Unseen subject 1

Ground Truth | ISP-SENet | Gazeformer-ISP-S



Unseen subject 2

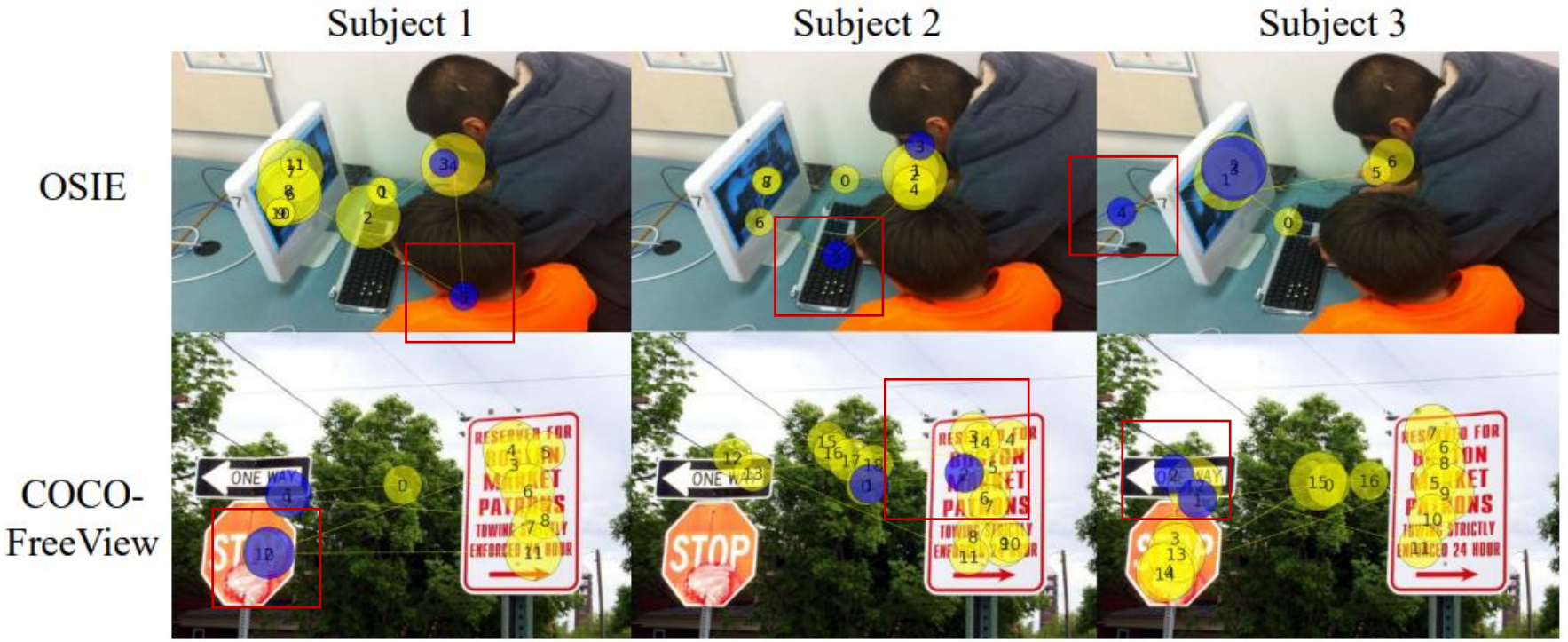
Ground Truth | ISP-SENet | Gazeformer-ISP-S

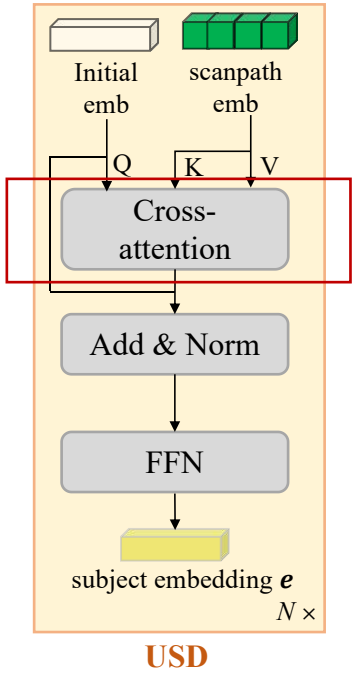


Analysis – Model interpretability

Cross-attention weights: which fixations have greater influence on the subject embedding.

- Highest attention weights on different objects
- Different initial focus





Summary

- We introduce Few-shot personalized scanpath prediction task.
- We create a pipeline that independently learn subject embedding and scanpath prediction.
- ISP-SENet achieves SOTA on various datasets under different viewing tasks.
- Our model is interpretable and encourages future work on personalized attention analysis.

Acknowledgement: This project is supported by US NSF grant IIS-2123920.