Few-shot Personalized Scanpath Prediction

Ruoyu Xue, Jingyi Xu, Sounak Mondal, Hieu Le, Gregory Zelinsky, Minh Hoai, Dimitris Samaras

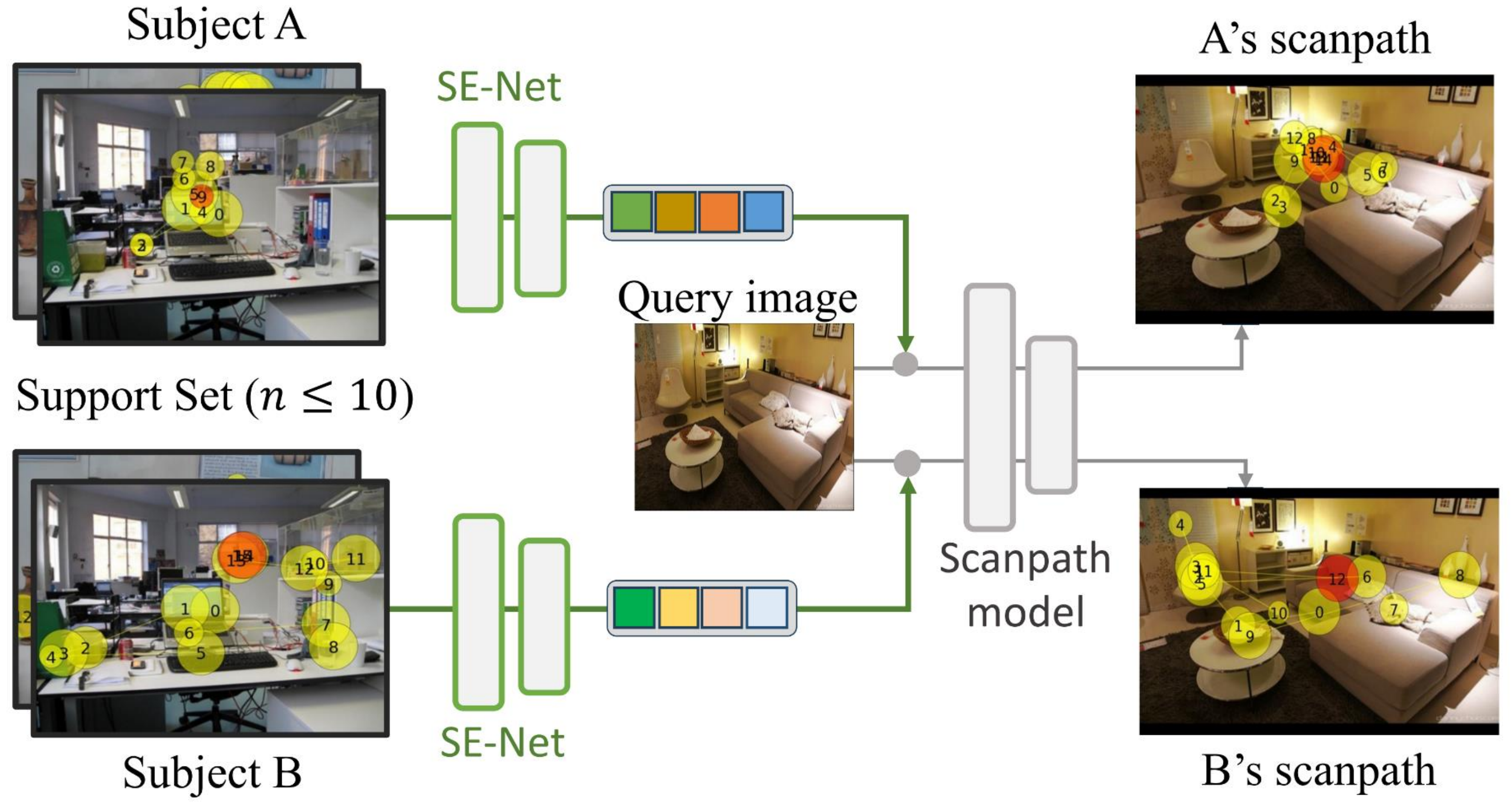
Stony Brook University

CVPR Nashville JUNE 11-15, 2025

Introduction

Motivation

- ➢ Personalized scanpath prediction requires extensive data.
  1. COCO-Search18 requires each subject spend 10-12 hours.
  2. In practice, PSP should be trainable on less data.
- ➢ Essential to develop a model that can quickly adapt to new viewer (subject) with minimal support data.



Challenge

- ➢ Minimal support data ($n \leq 10$) cause severe overfitting on unseen subjects.
- ➢ Existing methods do not fully utilize the learned information on seen subjects.
  1. Subject embedding is a by-product of scanpath prediction.
  2. Failed to capture subject similarity and difference.

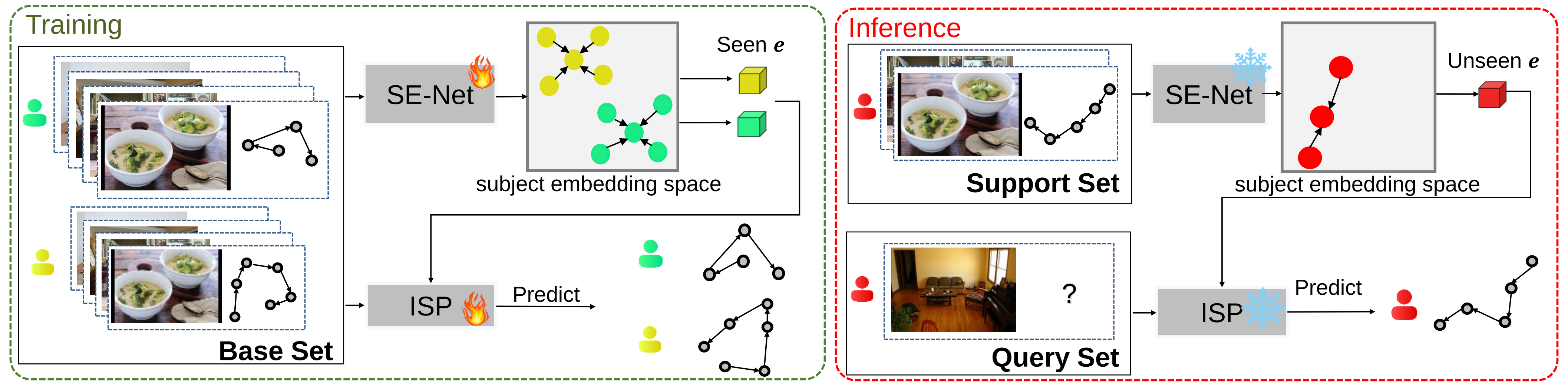
Contribution

- ➢ We propose a new task: few-shot personalized scanpath prediction (FS-PSP).
- ➢ We tackle this problem by separately train a subject embedding network and scanpath prediction network.
- ➢ We achieve SOTA on FS-PSP over three datasets under different viewing tasks including free-viewing and search.

Method

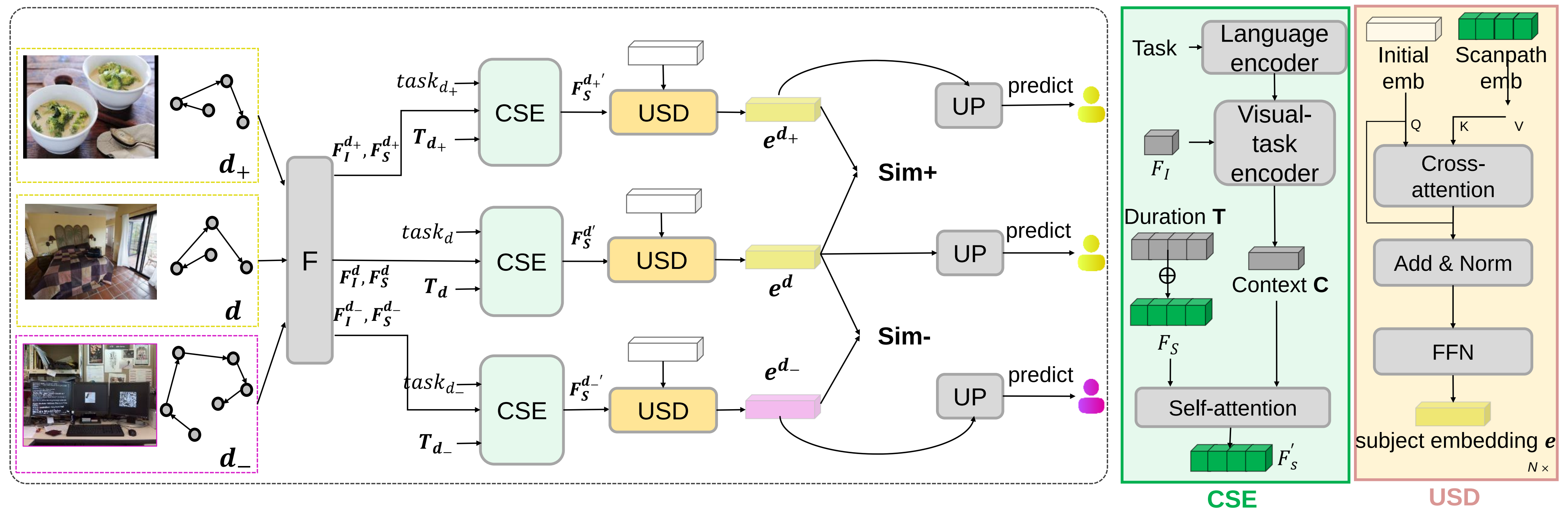
Problem Formulation

- ➢ Given: base set (large, seen subjects), support set ($n \leq 10$, unseen subjects).
- ➢ Task: predict scanpath of unseen subjects on query set (unseen images).



ISP-SENet

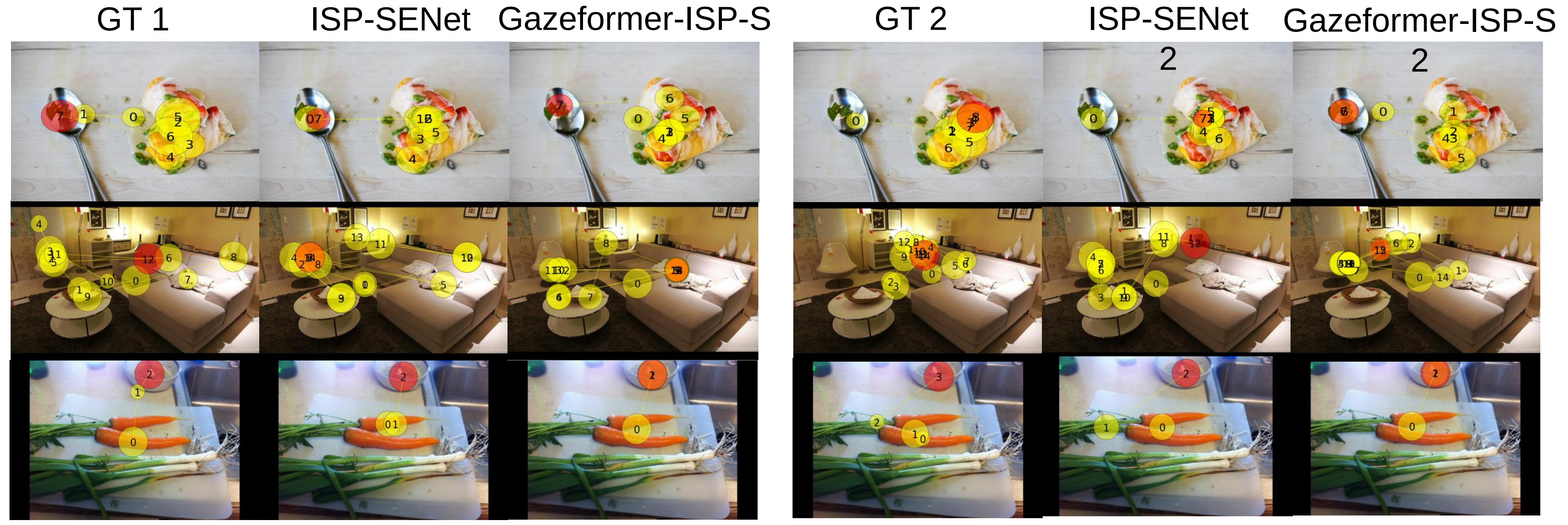
- ➢ Train SE-Net to generate seen subject embedding $\boldsymbol{e}$.
- ➢ Train ISP[Chen et al. CVPR 2024] with $\boldsymbol{e}$ to predict scanpaths.
- ➢ Use support set to obtain unseen $\boldsymbol{e}$ from SE-Net.
- ➢ Predict scanpath of unseen subject on query set with unseen $\boldsymbol{e}$.



SE-Net

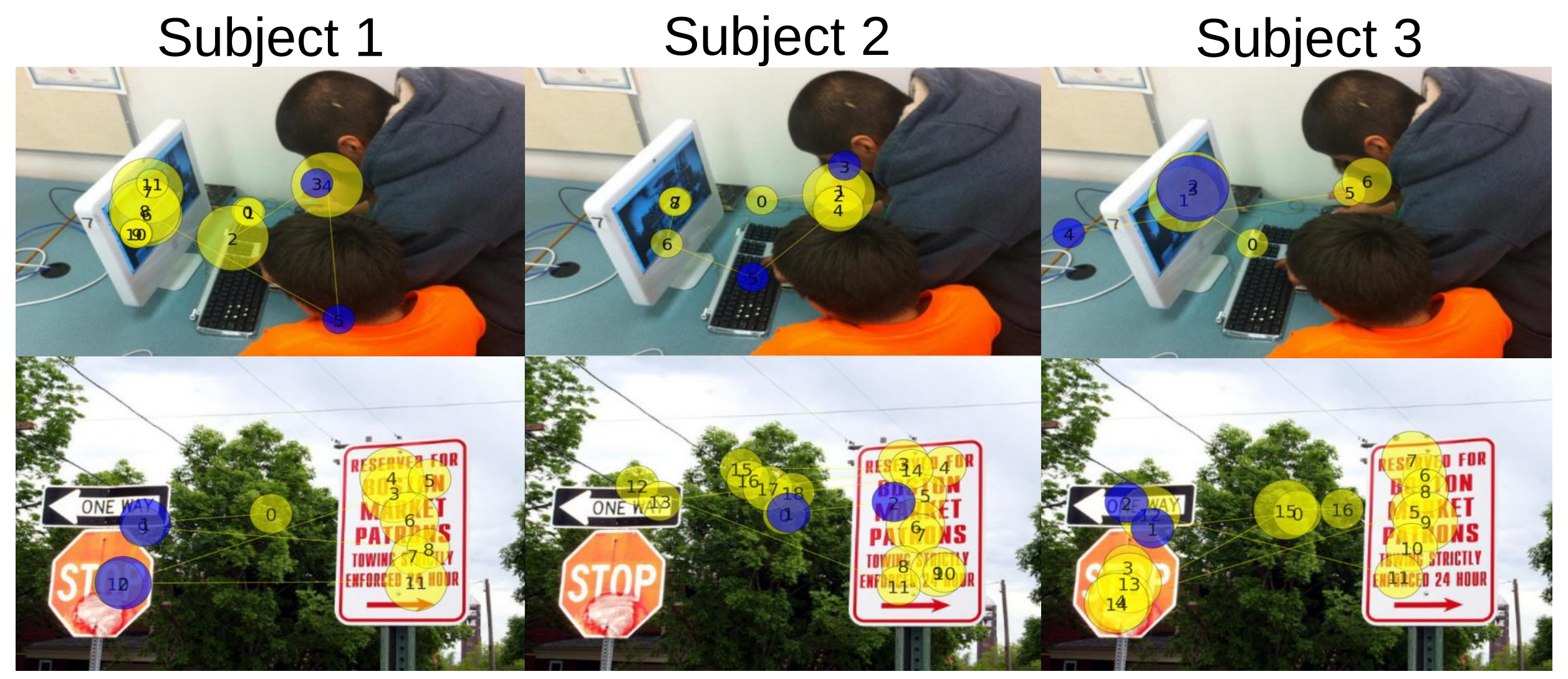
- ➢ We use contrastive learning to capture subject similarity and difference.
- ➢ Context-Scanpath Encoder (CSE): refine scanpath features with image content, viewing task and fixation durations.
- ➢ User-Scanpath Decoder (USD): capture subject attention traits from the refined scanpath features.
- ➢ User-Predictor (UP): predict subject id, accelerate converge speed.

Qualitative and Quantitative Results



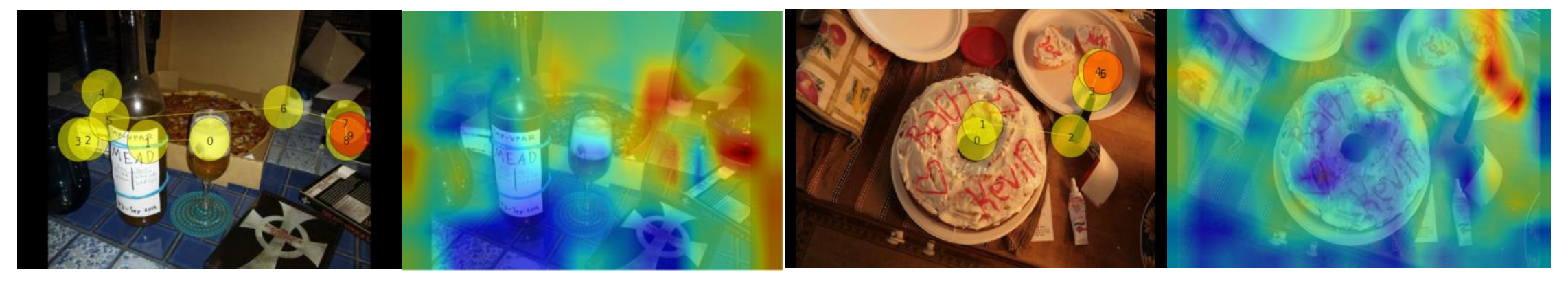
n-shot	Method	OSIE SM ↑	OSIE MM ↑	OSIE SED ↓	COCO-FreeView SM ↑	COCO-FreeView MM ↑	COCO-FreeView SED ↓	COCO-Search18 SM ↑	COCO-Search18 MM ↑	COCO-Search18 SED ↓
$n = 5$	ChenLSTM-ISP	0.319	0.773	7.855	0.320	0.815	12.950	0.386	0.773	2.489
	Gazeformer-ISP	0.340	0.791	7.920	0.286	0.800	14.630	0.353	0.774	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**

- ➢ ISP-SENet is ~44x faster than baselines in inference stage.
- ➢ ISP-SENet achieves 5.6% higher predicted scanpath accuracy, demonstrating that our subject embedding more effectively distinguishes unseen subjects.



Analysis

- ➢ Interpretability
  1. Extract cross-attention weights from USD.
  2. Mark fixations with higher weights in shaping subject embedding.
  3. Better understanding of subject attention traits.



Search “Bowl” Search “Knife”

- ➢ Visual-Task Encoder
  1. Extract attention weights from visual-task encoder.
  2. Awareness of search target.