Unifying Top-down and Bottom-up Scanpath Prediction Using Transformers

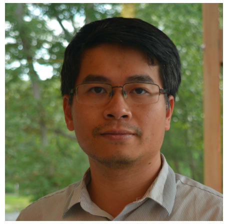

Zhibo Yang


Sounak Mondal


Seoyoung Ahn


Ruoyu Xue


Gregory Zelinsky


Minh Hoai

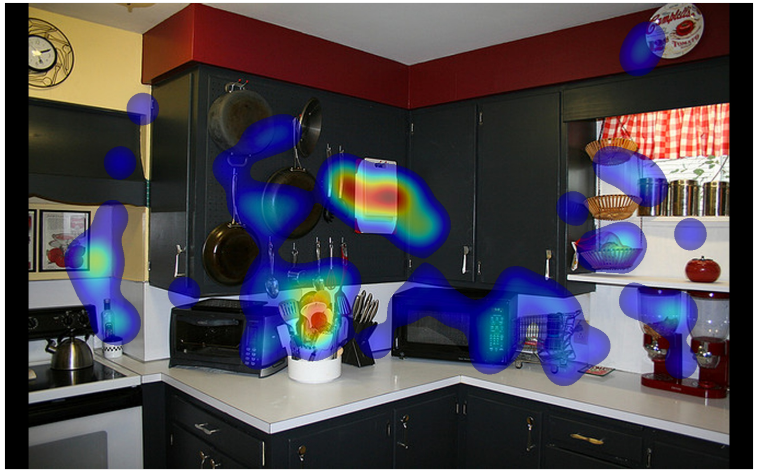

Dimitris Samaras

CVPR
JUNE 17-21, 2024
SEATTLE, WA

Stony Brook University
The State University of New York

Human attention: bottom-up vs top-down

Free viewing



Visual search

Microwave search

Clock search

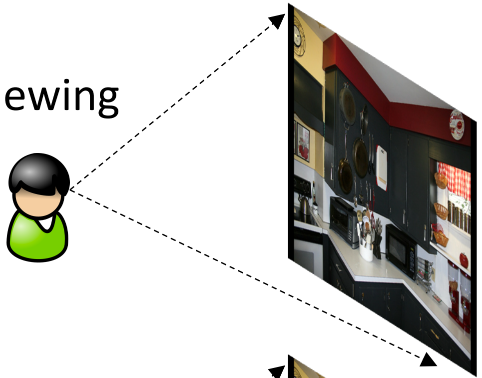
Bottom-up attention:
- Free viewing (“taskless ”)
- Attention prioritization (saliency) solely based on information in the image input (e.g., feature contrast)

Top-down attention:
- Visual search (goal-directed)
- Attention prioritization based on an **external goal** and the image put

Scanpath prediction

Bottom-up

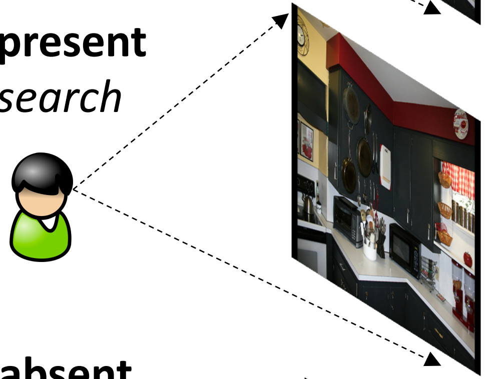
Free viewing



Saltinet [M. Assens et al., ICCV Workshops, 2017]
PathGAN [M. Assens et al., ECCV Workshops, 2018]
IOR-ROI-LSTM [W. Sun et al., PAMI, 2019]
DeepGaze III [M. Kummerer et al.,JoV, 2022]
…

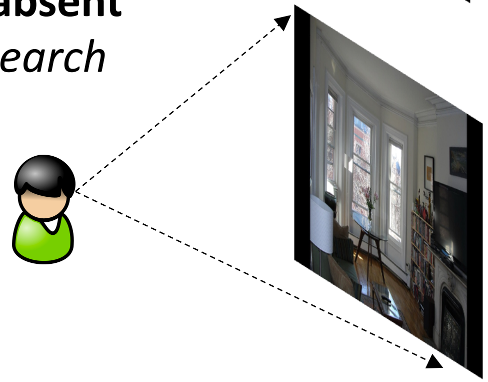
Top-down

Target-**present**
clock *search*

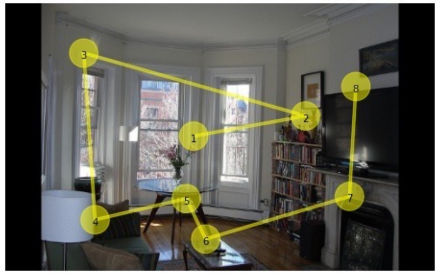


CFI [Zelinsky et al., CVPR Workshops, 2019]
IRL [Z. Yang et al., CVPR, 2020]
VQA [X. Chen et al., CVPR, 2022]
Gazeformer [S. Mondal et al., CVPR, 2023]
…

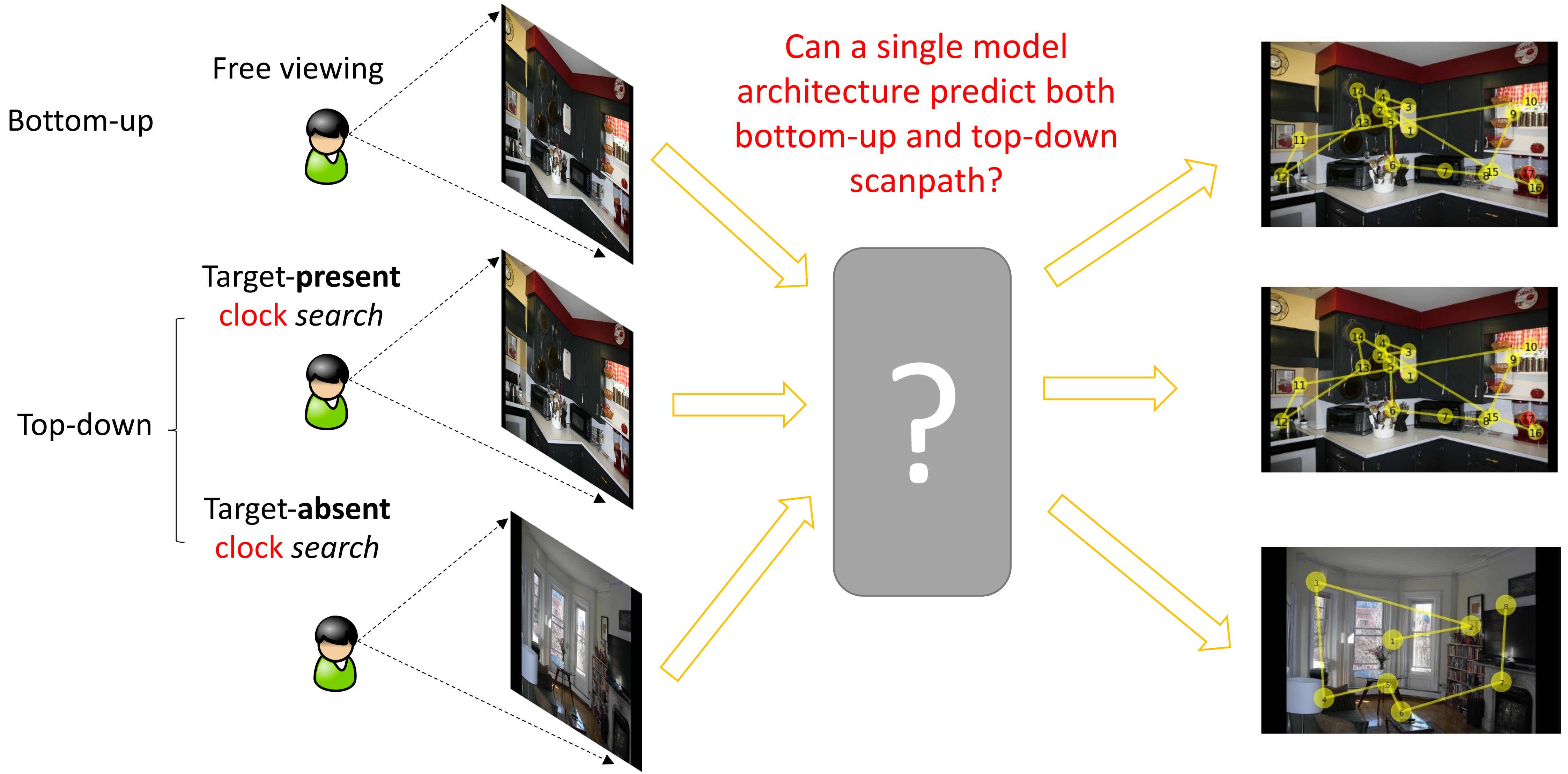
Target-**absent**
clock *search*



Detectability map [Rashidi et al., NeurIPS 2020]
FFMs [Z. Yang et al., ECCV, 2022]
…



Scanpath prediction



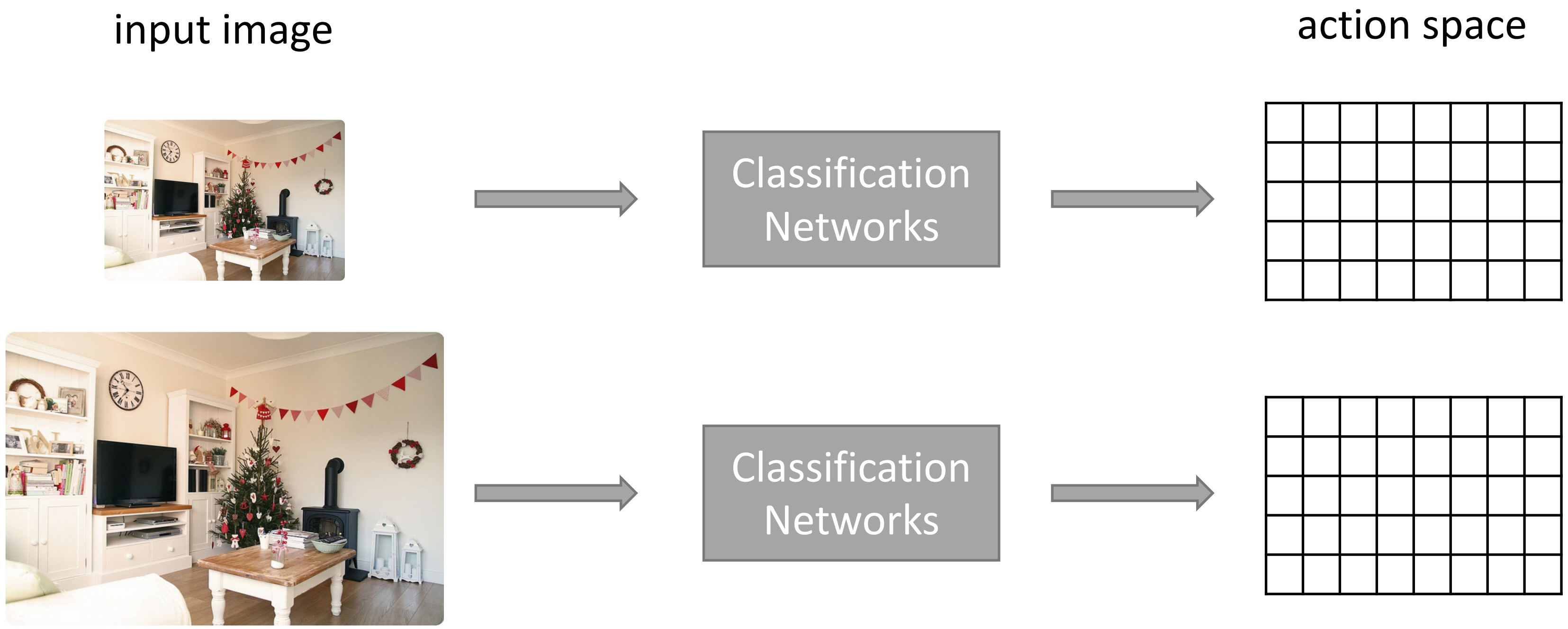
Limitation of existing approaches

- Traditional approaches have leaned on recurrent neural networks (RNNs) to uphold a dynamically updated hidden vector conveying information across fixations
    - PathGAN [M. Assens et al., ECCV Workshops, 2018]
    - IOR-ROI-LSTM [W. Sun et al., PAMI, 2019]
    - VQA [X. Chen et al., CVPR, 2022]
- Alternatively, simulations of a foveated retina have combined multi-resolution information at pixel, feature, or semantic levels
    - CFI [Zelinsky et al., CVPR Workshops, 2019]
    - IRL [Z. Yang et al., CVPR, 2020]
    - FFMs [Z. Yang et al., ECCV, 2022]

- Drawbacks
    - RNNs sacrifice interpretability
    - Multi-resolution simulations fall short in capturing crucial temporal and spatial information integration

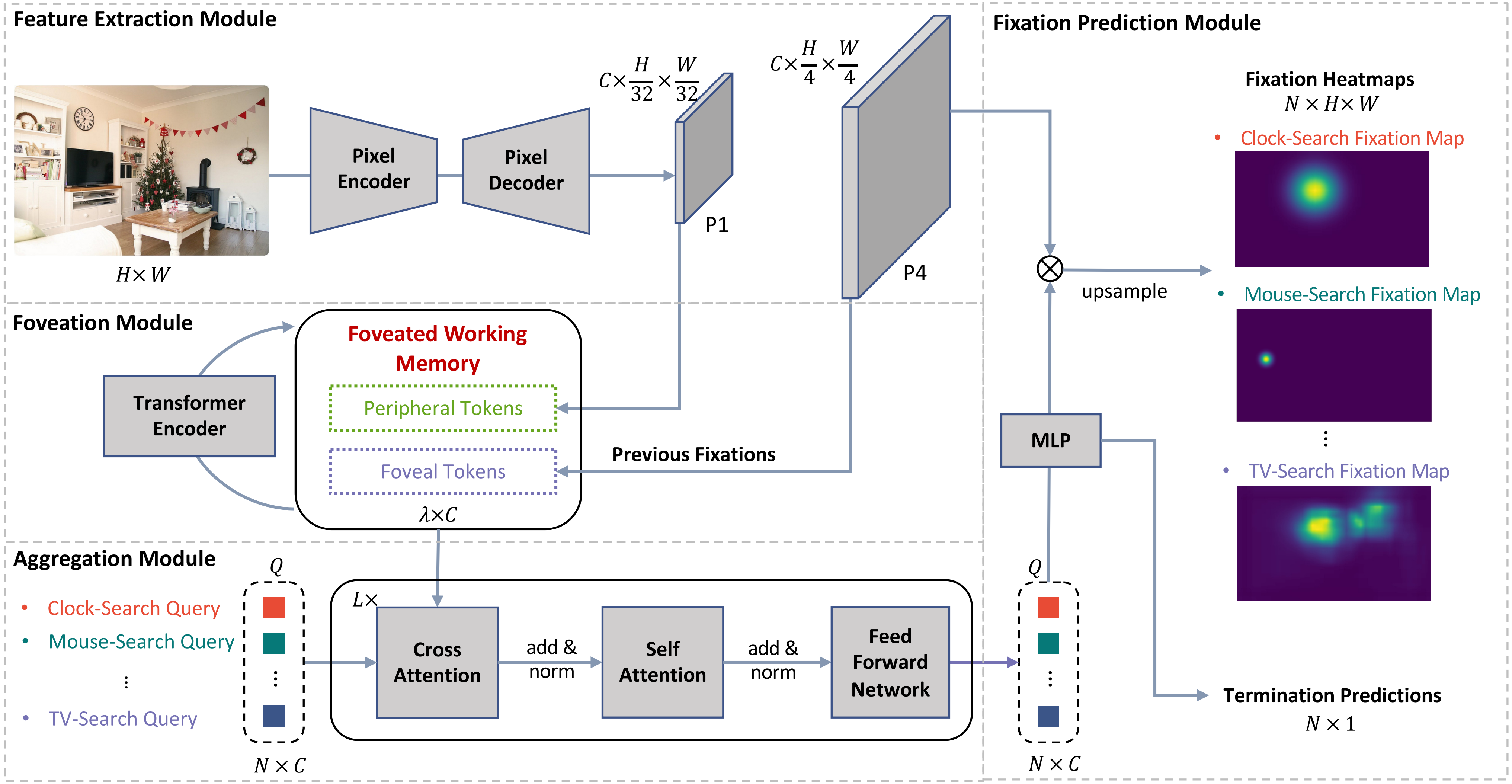
Limitation of existing approaches

- Existing methods use classification networks that discretize the space of all possible fixation locations as a coarse grid, which is invariant to input resolution and hence compromises accuracy.



- Cannot model fixations within the same cell, which occurs more often for high-res inputs:
  - For a 320x512 image with a 10x16 action space: a cell = 32x32 pixels
  - For a 1050x1680 image with a 10x16 action space: a cell = 105x105 pixels

Human Attention Transformer (HAT)

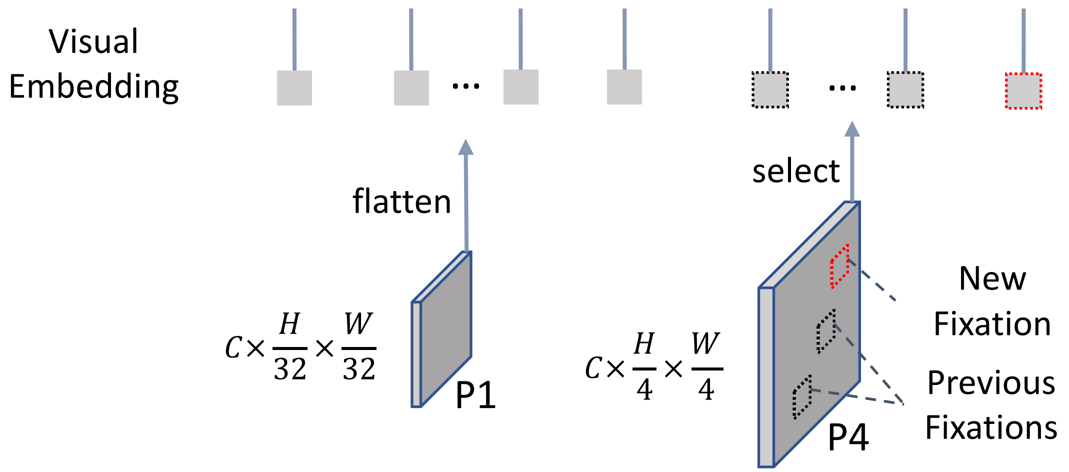


Foveated working memory

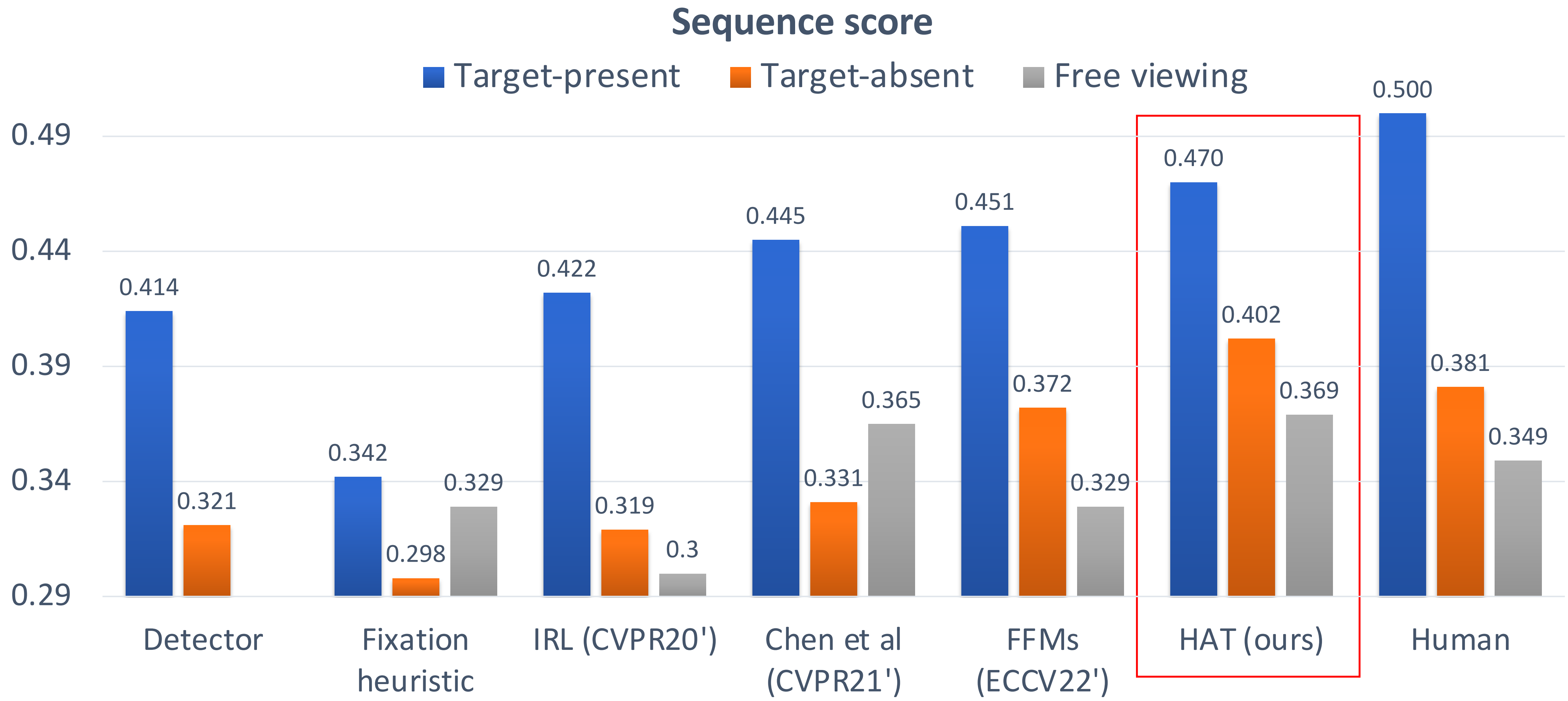
- We construct the working memory by starting with the **visual** embeddings ("what") flattened from P_1 over the spatial axes and selected from P_4 at previous fixation locations.

- **Scale** embedding is introduced to capture scale information.

- **Spatial** embeddings and **temporal** embeddings are further added to the tokens to enhance the "where" and "when" signals.

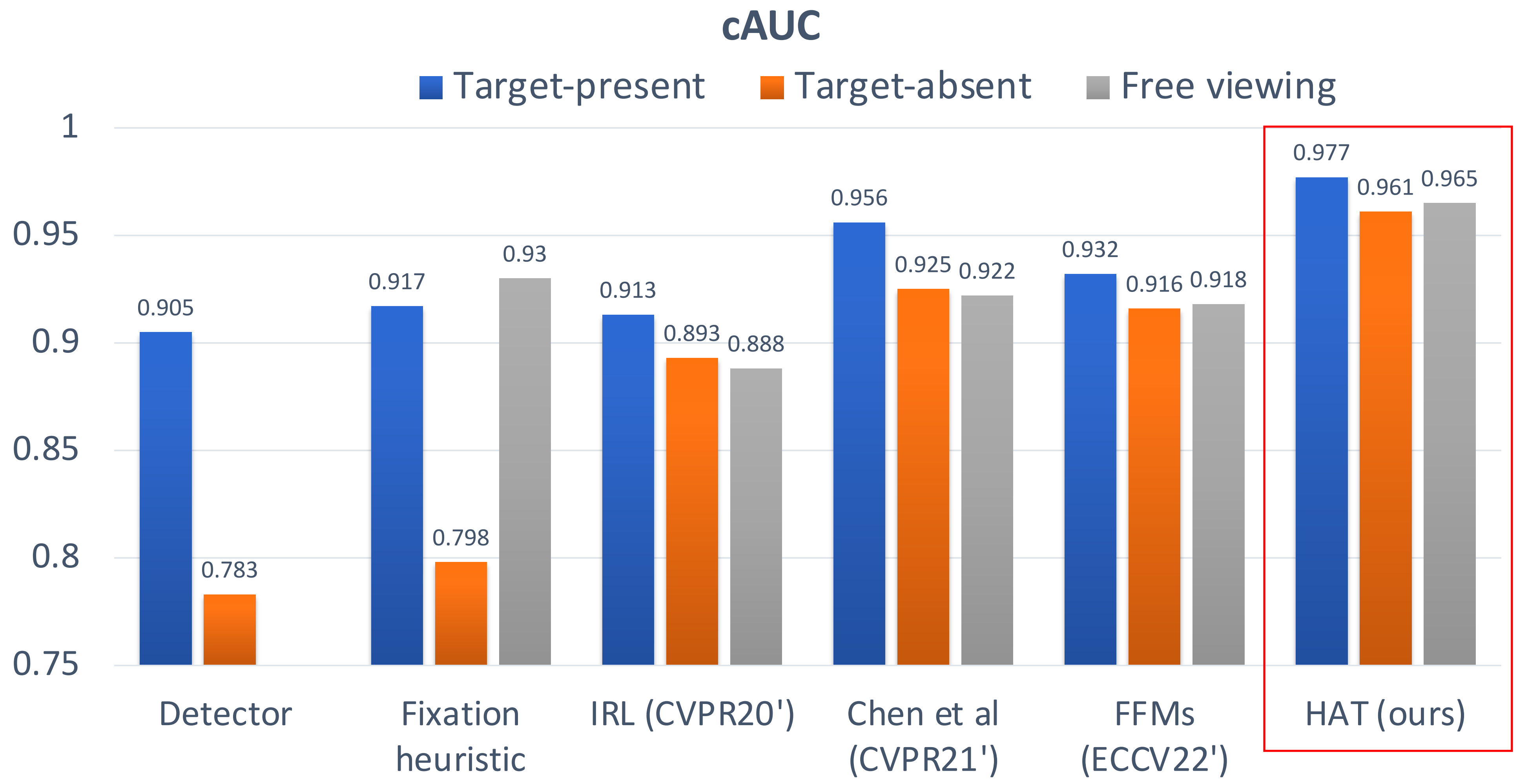


Experiments: quantitative results



- FWM has the best performance overall and surpasses “Human” in the TP and TA settings.

Experiments: quantitative results

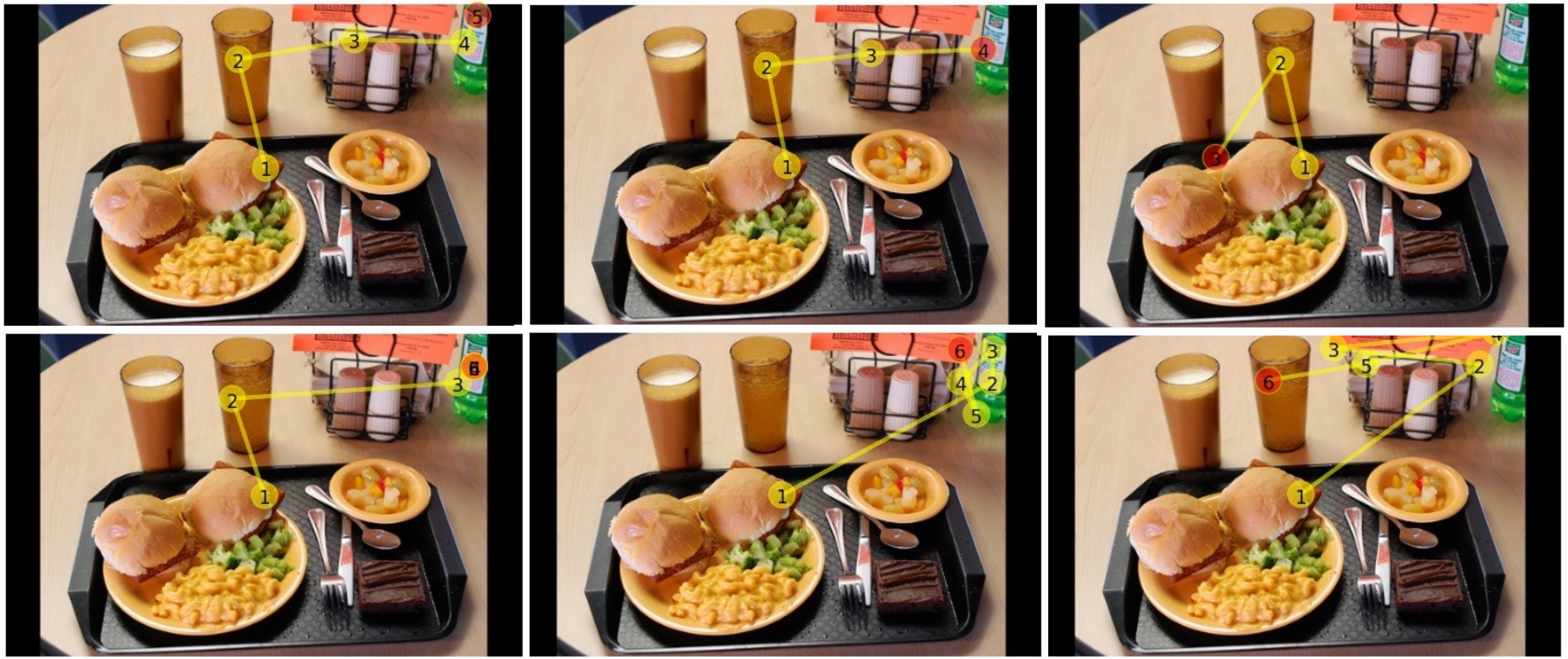


- FWM significantly outperforms all other methods in cAUC over all settings.

Experiments: quantitative results

Target-present bottle search

Human | **HAT (ours)** | FFMs (ECCV22')



Chen *et al* (CVPR21') | IRL (CVPR20') | Detector

Experiments: quantitative results

Target-absent stop sign search

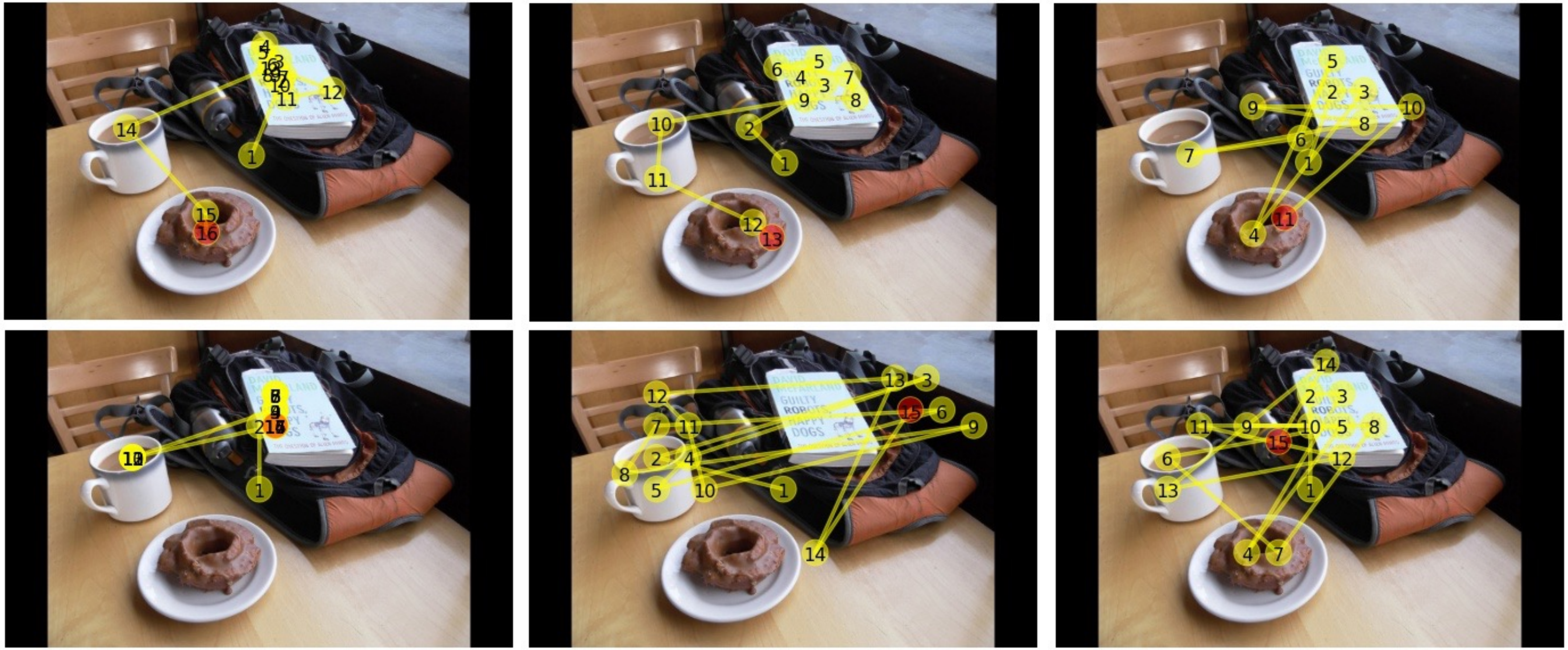
Human | **HAT (ours)** | FFMs (ECCV22')

Chen *et al* (CVPR21') | IRL (CVPR20') | Detector

Experiments: quantitative results

Free viewing

Human | **HAT (ours)** | FFMs (ECCV22')

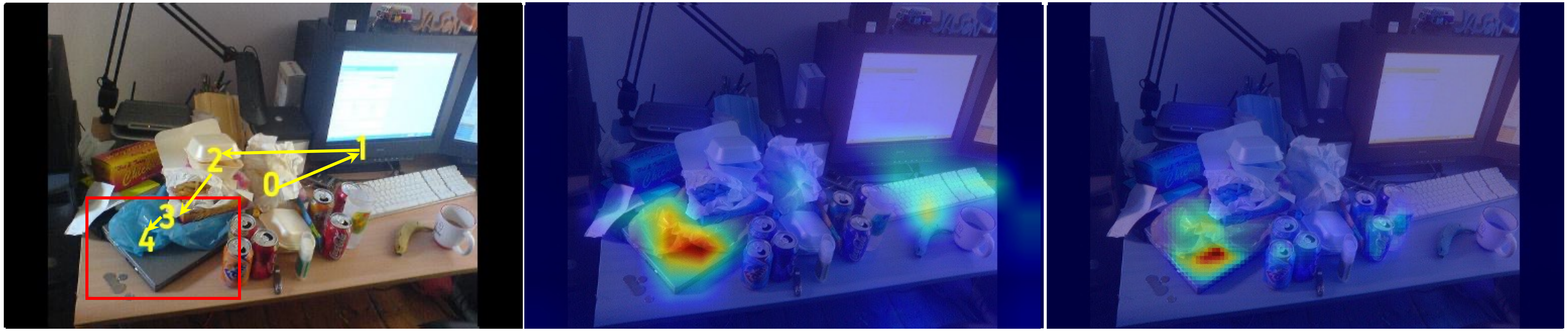


Chen *et al* (CVPR21') | IRL (CVPR20') | Detector

HAT captures contextual cues

Target-present laptop search

Predicted scanpath | Peripheral contribution map | Predicted fixation heatmap



τ is the termination prediction probability.

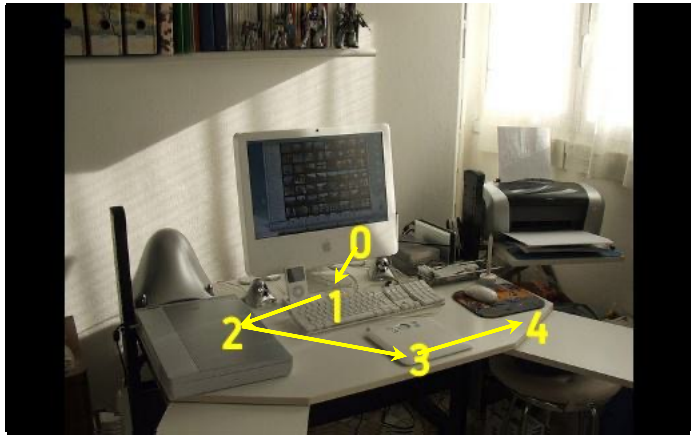
HAT captures contextual cues

Target-absent laptop search

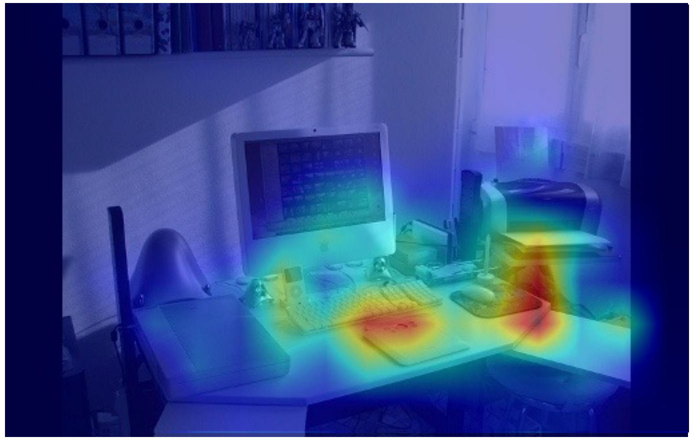
[illegible] fixation

τ = [illegible]

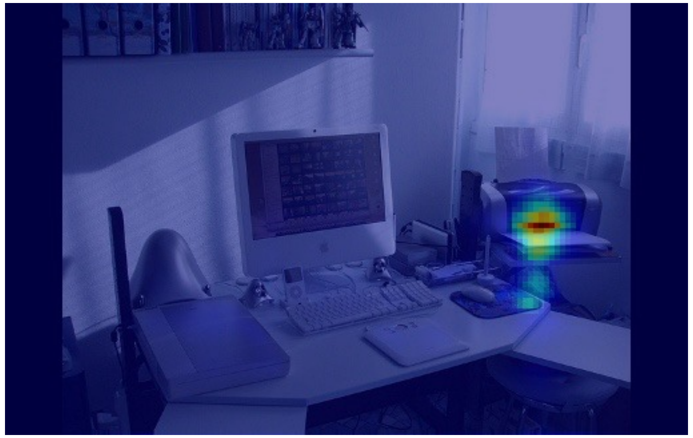
Predicted scanpath



Peripheral contribution map



Predicted fixation heatmap



Summary

- With HAT, our model’s prediction is not only accurate but also interpretable.

- HAT achieves the new SOTA in predicting the scanpath of fixations made during target-present and target-absent search, and reaches or exceeds SOTA in the prediction of “taskless” free-viewing fixation scanpaths.