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

Zhibo Yang[1,2], Sounak Mondal[1], Seoyoung Ahn[1], Ruoyu Xue[1],
Gregory Zelinsky[1], Minh Hoai[1,3], Dimitris Samaras[1]
[1]Stony Brook University [2]Waymo LLC [3]VinAI Research

CVPR 2024

Motivation:

Most models of visual attention aim at predicting either top-down or bottom-up control, as studied using different visual search and free-viewing tasks. In this paper we propose the Human Attention Transformer (HAT), a single model that predicts both forms of attention control.

Contributions:

1.We propose HAT, a novel transformer architecture integrating visual information at two different eccentricities to predict the spatial and temporal allocation of human attention.
2.We formulate scanpath prediction as a sequential dense prediction task without fixation discretization, making HAT applicable to high-resolution input.
3.The HAT architecture can be broadly applied to different attention control tasks, evidenced by the SOTA scanpath predictions in both visual search and free-viewing tasks.
4.HAT's attention predictions offer high interpretability, making it useful for studying gaze behavior.

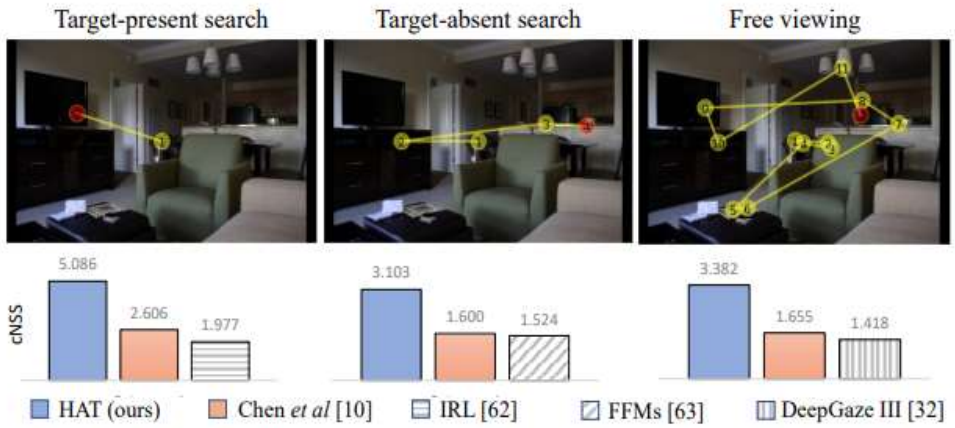


Figure 1. Given an image, the proposed **HAT** is able to predict scanpaths under three settings target-present search for TV; target-absent scanpath for sink; and free viewing. Importantly, HAT outperforms previous state-of-the-art scanpath prediction methods on multiple datasets across three settings: target-present, target-absent visual search and free viewing, that were studied separately.

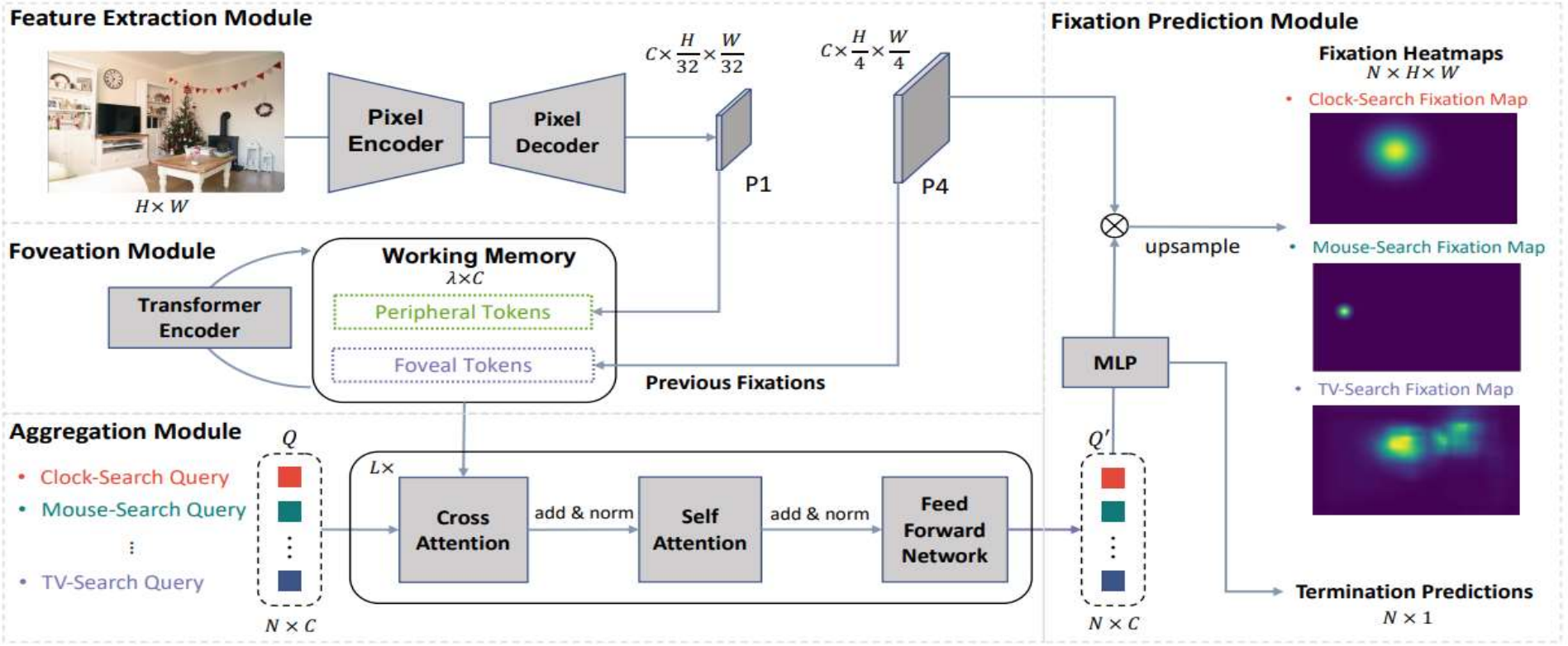
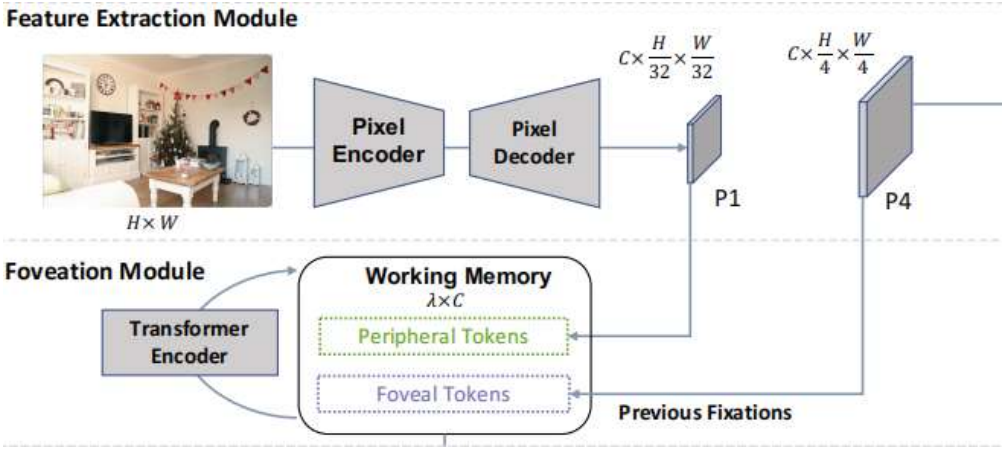


Figure 2. **HAT overview.** We use encoder-decoder CNNs to extract two sets of feature maps P_1 and P_4 of different spatial resolutions. A working memory with a capacity of λ tokens is constructed by combining all feature vectors from P_1 with the feature vectors of P_4 at previously fixated locations, representing information extracted from the periphery and central fovea. A transformer encoder is used to dynamically update the working memory at every new fixation. Then, HAT produces N per-task queries of dimension C (e.g., clock search and mouse search), with each learning to aggregates task-specific information from the shared working memory for predicting the fixations for its own task. Finally, the updated queries are convolved with P_4 to yield the fixation heatmaps after a MLP layer, and projected to the termination probabilities in parallel. Note, although this figure depicts visual search, the framework also applies for free viewing.

Pixel Encoder:(ResNet,Swin transformer)

Pixel Decoder:(FPN,MSDeformAttn)



The foveation module constructs a dynamic working memory using the feature maps P1 and P4 to represent the information a person acquires from the peripheral and foveal vision, respectively. Finally, we apply a Transformer encoder to dynamically update the working memory with the information acquired at a new fixation. Figure 3 illustrates the construction of the working memory.

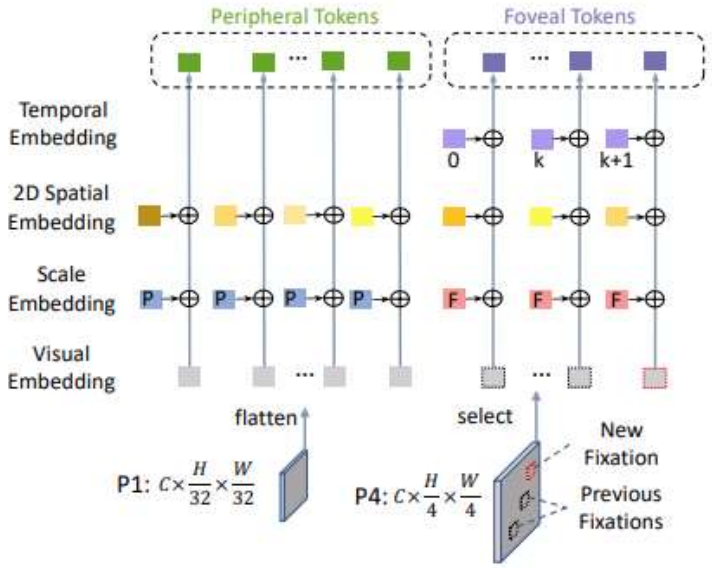
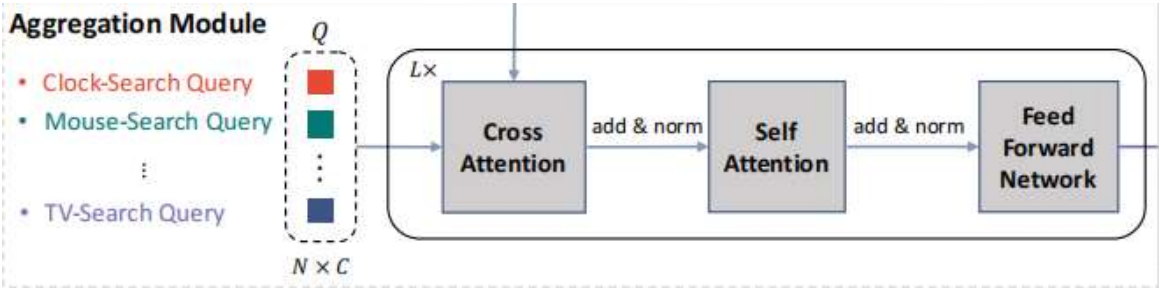
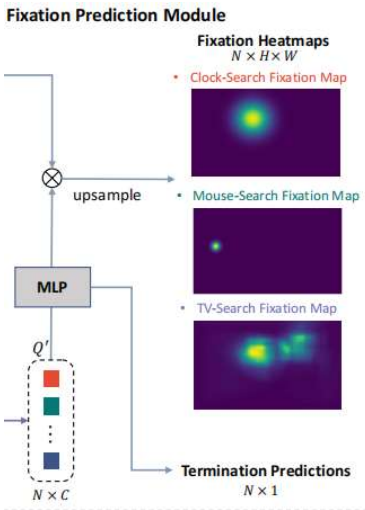


Figure 3. **Working memory construction.** We construct the working memory by starting with the visual embeddings ("what") flattened from P_1 over the spatial axes and selected from P_4 at previous fixation locations. A scale embedding is introduced to capture scale information. Spatial embeddings and temporal embeddings are further added to the tokens to enhance the "where" and "when" signals. At every new fixation (marked in red), we simply add a new foveal token while keeping other tokens unchanged.

The aggregation module is a transformer decoder that selectively aggregates information from the working memory using a set of learnable, task-specific queries $Q \in R^{N \times C}$, where N is the number of tasks (e.g., N = 18 for COCO-Search18 and N = 1 for free-viewing datasets).
Different from the standard transformer decoder, we switch the order of cross-attention and self-attention module.



The fixation prediction module yields the final prediction—a fixation heatmap $\hat{Y}_t$ and a termination probability $\hat{\tau}_t$ for each task t.



For the termination prediction, a linear layer followed by a sigmoid activation is applied on top of each updated query qt ∈ Q:

$$\hat{\tau}_t = \text{sigmoid}(W q_t^T + b),$$

For the fixation heatmap prediction, a Multi-Layer Perceptron (MLP) with two hidden layers first transforms qt into a task embedding, which is then convolved with the high resolution feature map P4 to get the fixation heatmap Y^t after a sigmoid layer:

$$\hat{Y}_t = \text{sigmoid}(P_4 \odot \text{MLP}(q_t))$$

Training Loss:

$$\mathcal{L} = \mathcal{L}_{\text{fix}}(\hat{Y}_t, Y) + \mathcal{L}_{\text{term}}(\hat{\tau}_t, \tau),$$

where $Y \in [0,1]^{H \times W}$ and $\tau \in \{0,1\}$ are the ground-truth fixation heatmap and termination label for task t, respectively. We compute Y by smoothing the ground-truth fixation map with a Gaussian kernel with the kernel size being one degree of visual angle. $\mathcal{L}_{\text{fix}}$ denotes the fixation loss and is computed using pixel-wise focal loss [34, 37]:

$$\mathcal{L}_{\text{fix}} = \frac{-1}{HW} \sum_{i,j} \begin{cases} (1 - \hat{Y}_{ij})^{\alpha} \log(\hat{Y}_{ij}) & \text{if } Y_{ij} = 1, \\ (1 - Y_{ij})^{\beta} (\hat{Y}_{ij})^{\alpha} & \\ \log(1 - \hat{Y}_{ij}) & \text{otherwise,} \end{cases} \quad (4)$$

where Y_{ij} represents the value of Y at location (i, j) and we set $\alpha = 2$ and $\beta = 4$ following [34, 63]. $\mathcal{L}_{\text{term}}$ is the termination loss and is computed by applying a binary cross entropy (negative log-likelihood) loss, i.e.,

$$\mathcal{L}_{\text{term}} = -\omega \cdot \tau \log(\hat{\tau}_t) - (1 - \tau) \log(1 - \hat{\tau}_t), \quad (5)$$

where ω is a weight to balance the loss of positive and negative training examples since there are many more negative labels than positive labels for training a termination prediction, especially for target-absent visual search and free-viewing tasks where scanpath are long. We set ω to be the ratio of the number of negative training instances to the number of positive ones.

Datasets:

Target_Present: CoCo-Search18(TP)
Target_Absent: CoCo-Search18(TA)
Free_View: CoCo-FreeView, MIT1003, OSIE

Evaluation metrics:

two aspects:
1) how similar the predicted scanpaths are to the human scanpaths;

Sequence Score(SS), Semantic Sequence Score(SemSS)

2) how accurate a model predicts the next fixation given all previous fixations.

conditional saliency metrics: cIG, cNSS and cAUC,

which measure how well a predicted fixation probability map of a model predicts the ground-truth (next) fixation when the model is provided with the fixation history of the scanpath in consideration, using the widely used saliency metrics, IG, NSS and AUC

Target-present search

	SemSS	SS	cIG	cNSS	cAUC
Human consistency	0.500	0.500	-	-	-
Detector	0.523	0.449	0.182	2.346	0.905
Fixation heuristic	0.506	0.437	1.107	2.186	0.917
IVSN [68]	0.368	0.326	-0.192	1.318	0.901
PathGAN [1]	0.280	0.239	-	-	-
IRL [62]	0.486	0.422	-9.709	1.977	0.913
Chen *et al*. [10]	0.518	0.445	-1.273	2.606	0.956
FFMs [63]	0.500	0.451	1.548	2.376	0.932
Gazeformer [42]	0.499	**0.489**	-	-	-
HAT (ours)	**0.543**	0.470	**2.399**	**5.086**	**0.977**

Table 1. **Target-present search scanpath prediction comparison** on the target-present test set of COCO-Search18. We highlight the best results in bold.

Target-absent search.

	SemSS	SS	cIG	cNSS	cAUC
Human consistency	0.372	0.381	-	-	-
Detector	0.332	0.321	-0.516	0.446	0.783
Fixation heuristic	0.309	0.298	-0.599	0.405	0.798
IVSN [68]	0.279	0.260	-0.219	0.884	0.867
PathGAN [1]	0.315	0.250	-	-	-
IRL [62]	0.329	0.319	0.032	1.202	0.893
Chen *et al*. [10]	0.340	0.331	-3.278	1.600	0.925
FFMs [63]	0.376	0.372	0.729	1.524	0.916
Gazeformer [42]	0.374	0.357	-	-	-
HAT (ours)	**0.382**	**0.402**	**1.686**	**3.103**	**0.961**

Table 2. **Target-absent search scanpath prediction comparison** on the target-absent test set of COCO-Search18. We highlight the best results in bold.

Free Viewing.

	SS	cIG	cNSS	cAUC
Human consistency	0.349	-	-	-
Fixation heuristic	0.329	0.319	1.621	0.930
PathGAN [1]	0.181	-	-	-
IRL [62]	0.300	-0.213	1.018	0.888
Chen *et al*. [10]	0.365	-1.263	1.655	0.922
DeepGaze III [32]	0.339	0.140	1.418	0.910
FFMs [63]	0.329	0.329	1.432	0.918
Gazeformer [42]	0.280	-	-	-
HAT	**0.369**	**1.485**	**3.382**	**0.965**

Table 3. **Comparing free-viewing scanpath prediction algorithms** (rows) using multiple metrics (columns) on the test set of COCO-FreeView. The best results are highlighted in bold.

HAT learns the entire scanpath distribution from multiple subjects whereas GazeFormer overfits to the "average person" and fails to predict scanpaths from different subjects.

Scanpath visualization

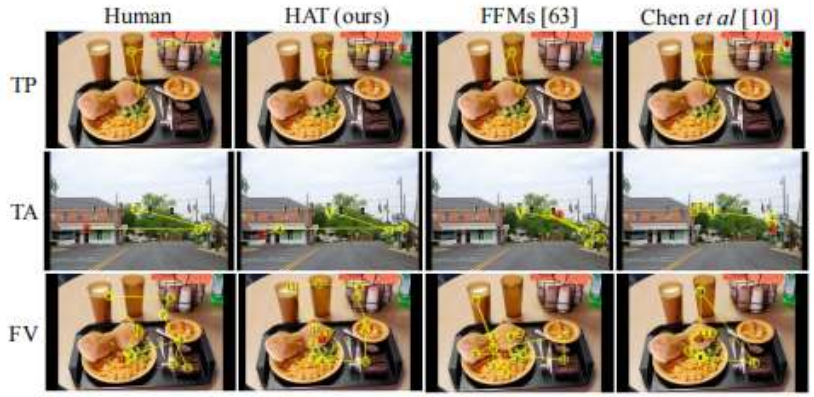


Figure 4. **Visualization of the ground-truth human scanpaths and predicted scanpaths of different methods (columns).** Three different settings (rows) including target-present bottle search, target-absent stop sign search and free viewing are shown from the top to bottom. The final fixation of each scanpath is highlighted in red circle. For methods without termination prediction, i.e., IRL, detector and fixation heuristic, we visualize the first 6 fixations for visual search and 15 for free viewing. The rightmost column shows the predicted scanpaths of the heuristic methods (detector 630 for visual search and fixation heuristic for free-viewing)

Ablation

	SemSS	SS	cIG	cNSS	cAUC
baseline (80×128)	**0.382**	**0.402**	**1.686**	**3.103**	**0.961**
– peripheral tokens	0.375	0.396	1.600	3.003	0.960
– foveal tokens	0.358	0.385	1.179	2.380	0.948
low-res (20×32)	0.374	0.389	1.534	2.760	0.955

Peripheral and foveal tokens

We verify the effectiveness of peripheral tokens and foveal tokens by ablating them one at a time. It is shown in Tab. 4 that ablating any one of them incur a performance drop over all metrics.

Output resolution

a reduced resolution incurs a noticeable performance drop in HAT, HAT still outperforms prior state-of-the-art FFMs with the same output resolution and Chen et al. using a higher output resolution